

SPRINGER
REFERENCE

Alain Marciano
Giovanni Battista Ramello
Editors

Encyclopedia of Law and Economics

Encyclopedia of Law and Economics

Alain Marciano
Giovanni Battista Ramello
Editors

Encyclopedia of Law and Economics

With 146 Figures and 50 Tables

 Springer

Editors

Alain Marciano
MRE and University of Montpellier
Montpellier, France

Faculté d'Economie
Université de Montpellier and
LAMETA-UMR CNRS
Montpellier, France

Giovanni Battista Ramello
DiGSPES
University of Eastern Piedmont
Alessandria, Italy

IEL
Torino, Italy

ISBN 978-1-4614-7752-5 ISBN 978-1-4614-7753-2 (eBook)

ISBN 978-1-4614-7754-9 (print and electronic bundle)

<https://doi.org/10.1007/978-1-4614-7753-2>

Library of Congress Control Number: 2019930845

© Springer Science+Business Media, LLC, part of Springer Nature 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

We dedicate this book to Jürgen Backhaus. The “encyclopedia” would never exist without him. Not only did he initiate the project but he also gave us the opportunity to carry it to publication. Jürgen Backhaus is a pioneer in modern European law and economics; he has served as a scientific bridge between the North American and the European law and economics community and he deserves to be recognized as a totemic figure in this field.

Preface

This book is the outcome of a dream Jürgen Backhaus had several years ago, to endow the law and economics community with a new and broad resource, making available to scholars an accessible and “living” reference. While other valuable resources already exist, providing a number of essays on given topics, this encyclopedia is original in that it aims at offering a very broad array of short references within law and economics or tangential to the discipline. Indeed, a large number of these entries are at the limits of law and economics and other fields that are close to law and economics – public choice, institutional economics, transaction cost economics, industrial economics, etc. This reveals that law and economics is much more difficult to delineate than what is usually assumed. And there is nothing to surprise us here: laws, legal rules, and institutions are everywhere. They are the bases of human societies and at the core of economic activities. It is indeed almost impossible to be an economist without taking law and institutions into account. This is what we tried to do in this encyclopedia. Finally, also inspired by Jürgen Backhaus, we chose to make room for historical and methodological entries about the founders of the field and about some of the important scholars who, directly or indirectly, contributed to make law and economics what it became. These three dimensions reflect the *parti pris* that was chosen to build this book. They reflect our convictions about law and economics.

The project was very ambitious, and we tried our best to keep it on the original tracks, nonetheless adding a new objective: making the encyclopedia the common ground, the cross-roads, of the international scholarly community, and hence inviting contributors from all around the world and what is more from different generations in order to favor the widest possible cross-fertilization. We were glad to see how enthusiastic and generous our colleagues were in answering us – the encyclopedia comprises now almost 400 entries. We are indeed grateful for the many ideas, topics, and innovations that they have supplied. At some point, we simply became the administrators of a commons, and this has been a wonderful experience.

The publishing agenda required us to “freeze” the process at a given date in order to make available the printed version, but this has not stopped the ongoing process, and many other contributors are still working toward the development of new and interesting entries that will be subsequently available online and then printed. Of course, despite the large number of entries, many topics are still missing, and this is somewhat unavoidable in the case of such a large project and its “living” nature, that makes it always “in progress.” We beg

your pardon if you don't find what may be of interest to you, but we are sure that you will find many stimulating perspectives, and we invite you, in case you really feel that a specific entry is needed, to think about becoming contributors and to send us a proposal.

Alain Marciano
Giovanni Battista Ramello

Acknowledgments

Every book is a collective endeavor. Even more when it is a book of this type. We could not have completed it without the help of a large number of people. We thank them all heartily. We thank Lorraine Klimowich and Michael Hermann, from Springer, who trusted us enough to give us the opportunity to edit it. Also, we thank Barbara Wolf, Veronika Mang, and Arianne Israel with whom we interacted regularly. Their help was invaluable. We also thank all the contributors who accepted the rules of the game.

About the Editors



Alain Marciano is Full Professor of Economics at the University of Montpellier. He is mainly interested in the emergence and existence of norms (formal and informal, juridical and non-juridical) and mechanisms of coordination/cooperation among individuals. A major area of his work consists in studying the history and methodology of recent economics – with a focus on law and economics and public choice. He has edited many books in law and economics and published a reader in law and economics. He has published articles in different journals (*International Review of Law and Economics*, *Public Choice*, *Constitutional Political Economy*, the *Journal of Economic Behavior and Organization*, *Law and Contemporary Problems*, the *Journal of Institutional Economics*, *History of Political Economy*, the *Journal of the History of Economic Thought*, the *European Journal of the History of Economic Thought*, the *Journal of Economic Methodology*). He serves as Co-Editor-in-Chief of the *European Journal of Law and Economics* and is in the boards of different journals. He is Series Editor of *Palgrave Studies in Institutions, Economics and Law* (Palgrave Macmillan).



Giovanni Battista Ramello is Full Professor of Industrial Economics at Università del Piemonte Orientale. In addition, he coordinates the IEL International Ph.D. Program in Comparative Analysis of Institutions, Economics and Law hosted at Collegio Carlo Alberto. He mainly teaches industrial economics and law and economics, while his research covers topics in industrial organization, antitrust and regulation, intellectual property rights and market structure, and economic analysis of litigation and of judicial systems. He published articles in several international leading journals and has worked as an advisor for ministries of countries all around the world and other international institutions. Furthermore, he serves as Co-Editor-in-Chief of the *European Journal of Law and Economics*, as Editor of the *European Journal of Comparative Economics*, and as an Associate Editor of the *Journal of Industrial and Business Economics*. Besides the editorship of the *Encyclopedia of Law and Economics*, he is Series Editor of *Palgrave Studies in Institutions, Economics and Law* (Palgrave Macmillan).

Contributors

Alden F. Abbott Heritage Foundation, Washington, DC, USA

Amanda Agan Princeton University, Princeton, NJ, USA

Varouj A. Aivazian Department of Economics, University of Toronto, Toronto, ON, Canada

Gianmaria Ajani Department of Law, University of Torino, Torino, Italy

Pinar Akman School of Law, University of Leeds, Leeds, UK

Hüseyin Can Aksoy Faculty of Law, Bilkent University, Bilkent, Ankara, Turkey

Maria Rosaria Alfano Economics Department, Seconda Università degli Studi di Napoli, Italy

Paul Dragos Aligica Mercatus Center, George Mason University, Arlington, VA, USA

Sofia Amaral-Garcia Center for Law and Economics, ETH Zurich, Zürich, Switzerland

Angela Ambrosino Department of Economics and Statistics Cognetti de Martiis, University of Turin, Torino, Italy

David A. Anderson Department of Economics, Centre College, Danville, KY, USA

Charles H. Anderton College of the Holy Cross, Worcester, MA, USA

Cristiano Antonelli Dipartimento di Economia e Statistica “Cognetti de Martiis”, Università di Torino, Collegio Carlo Alberto, Torino, Italy

Alessandro Antonietti Department of Psychology, Catholic University of the Sacred Heart, Milan, Italy

Benito Arruñada Pompeu Fabra University and BGSE, Barcelona, Spain

Mehmet Bac Sabanci University, Istanbul, Turkey

Matthew J. Baker Hunter College and the Graduate Center, CUNY, New York, NY, USA

Przemysław Banasik Faculty of Management and Economics, Gdansk University of Technology, Gdańsk, Poland

Anna Laura Baraldi Department of Economics, Second University of Naples, Capua, Italy

Guglielmo Barone Bank of Italy and RCEA, Bologna, Italy

Christian Barrère Faculty of Economics, Laboratoire R.E.G.A.R.D.S, University of Reims, Reims, France
ISMEA, Paris, France

Florian Baumann Center for Advanced Studies in Law and Economics (CASTLE), University of Bonn, Bonn, Germany

Cécile Bazart CEE-M – Univ Montpellier – CNRS – INRA – SupAgro, University of Montpellier, Montpellier, France
Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier 1, Montpellier, France

Flavio Bazzana Department of Economics and Management, University of Trento, Trento, TN, Italy

Sylvain Béal CRESE EA3190, Université Bourgogne Franche-Comté, Besançon, France

Paul Belleflamme CORE and LSM, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Tanja Benedict InnovationLab GmbH, Legal Services and IP Management, Heidelberg, Germany

Bruce L. Benson Department of Economics, Florida State University, Tallahassee, FL, USA

Pierre Bernhard Biocore Team, Université Côte d'Azur-INRIA, Sophia Antipolis Cedex, France

Enrico Bertacchini Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Torino, Italy

Elodie Bertrand CNRS and University Paris 1 Panthéon-Sorbonne (ISJPS, UMR 8103), Paris, France

Samantha Bielen Faculty of Applied Economics, Hasselt University, Diepenbeek, Belgium
Faculty of Law, University of Antwerp, Antwerp, Belgium

Magdalena Bielenia-Grajewska Intercultural Communication and Neurolinguistics Laboratory, Department of Translation Studies, Faculty of Languages, University of Gdansk, Gdansk, Poland

Sophie Bienenstock Economix, Université Paris Nanterre, Paris, France

Walter E. Block Joseph A. Butt, S.J. College of Business, Loyola University
New Orleans, New Orleans, LA, USA

Ulrich Blum Faculty of Law and Economics, Martin-Luther-University of
Halle Wittenberg, Halle, Germany
University of International Business and Economics, Beijing, China

Peter J. Boettke Department of Economics, George Mason University,
Fairfax, VA, USA

J. R. Borrell Secció de Polítiques Públiques, Dep. d'Econometria,
Estadística i Economia Aplicada, Grup de Governs i Mercats (GiM), Institut
d'Economia Aplicada (IREA), Universitat de Barcelona, Barcelona, Spain
IESE Business School, Public-Private Sector Research Center, University of
Navarra, Navarra, Spain

Patrice Bougette Department of Economics, Université Côte d'Azur, CNRS,
GREDEG, Nice, France

Afef Boughanmi University of Lorraine (IUP Finance), BETA-CNRS
Laboratory, Nancy, France

Jenny Bourne Carleton College, Northfield, MN, USA

Cécile Bourreau-Dubois Department of Economics, BETA UMR 7522,
Université de Lorraine, Nancy, France

Marcel Boyer Department of Economics, Université de Montréal, Montréal,
QC, Canada
Toulouse School of Economics, Toulouse, France

Jurgen Brauer Hull College of Business, Georgia Regents University,
Augusta, GA, USA

Karine Brisset Centre de Recherche sur les Stratégies Economiques,
University of Franche-Comté, Besançon, France

Maria T. Brouwer Amsterdam, The Netherlands

Christian Boris Brunner School of Agriculture, Policy and Development,
University of Reading, Reading, Berkshire, UK

Wolfgang Buchholz University of Regensburg, Regensburg, Germany

Paolo Buonanno Department of Management, Economics and Quantitative
Methods, University of Bergamo, Bergamo, Italy

Mauro Bussani University of Trieste, Trieste, Italy
University of Macau, S.A.R. of the P.R.C., Macau, China

Jeffrey L. Callen Rotman School of Management, University of Toronto,
Toronto, ON, Canada

Samuel Cameron Division of Economics, University of Bradford, Bradford, UK

Rosolino A. Candela Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

Laurent Carnis University Paris Est and IFSTTAR, Noisy-le-Grand, France

André Casajus HHL Leipzig Graduate School of Management, Leipzig, Germany

LSI Leipziger Spieltheoretisches Institut, Leipzig, Germany

Alessandra Cassar University of San Francisco, San Francisco, CA, USA

Danny Cassimon Institute of Development Policy and Management (IOB), University of Antwerp, Antwerp, Belgium

Roy Cerqueti Department of Economics and Law, University of Macerata, Macerata, Italy

Günther Chaloupek Austrian Chamber of Labour, Vienna, Austria

Nathalie Chappe University of Franche Comté CRESE, Besançon, France

Christophe Charlier Department of Economics, Université Côte d'Azur, CNRS, GREDEG, Nice, France

Alexandre Chirat Triangle, Université Lumière Lyon II, Lyon, France

Arianna Ciabattini Department of Management, University of Turin, Turin, Italy

Aitor Ciarreta University of the Basque Country, UPV/EHU, Bilbao, Spain

Valérie Clément MRE, EA 7491, Université de Montpellier, Montpellier, France

Ignacio N. Cofone NYU Information Law Institute, New York, NY, USA

Enrico Colombatto Università di Torino, Turin, Italy

IREF (Institut de Recherches Économiques et Fiscales), Lyon, France

Barbara Colombo Department of Psychology, Catholic University of the Sacred Heart, Brescia, Italy

Vittoria Colombo Università degli Studi di Torino, Turin, Italy

Raffaella Coppier Department of Economics and Law, University of Macerata, Macerata, Italy

Cristina Costantini Department of Law, University of Perugia, Perugia, Italy

Christopher J. Coyne Department of Economics, George Mason University, Fairfax, VA, USA

Samuel Cameron has retired.

Günther Chaloupek has retired.

John M. Crespi Iowa State University, Ames, IA, USA

Katalin J. Cseres Amsterdam Centre for European Law and Governance, Amsterdam Center for Law and Economics, University of Amsterdam, Amsterdam, The Netherlands

Péter Cserne University of Hull, Hull, UK

Elena D'Agostino Department of Economics, University of Messina, Messina, Italy

Lucia dalla Pellegrina DEMS, University of Milano-Bicocca and BAFFI CAREFIN Centre, Università Bocconi, Milan, Italy

Maria De Benedetto Department of Political Science, Roma Tre University, Rome, Italy

Gerrit De Geest Washington University in St. Louis, St. Louis, MO, USA

Jef De Mot Postdoctoral Researcher FWO, Faculty of Law, University of Ghent, Belgium

Erwin Dekker Erasmus School of History, Culture and Communication, Erasmus University Rotterdam, Rotterdam, The Netherlands

Ben Depoorter Hastings College of the Law, University of California, San Francisco, CA, USA

Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium

Marc Deschamps CRESE EA3190, Université Bourgogne Franche-Comté, Besançon, France

Claudio Detotto DiSea – CRENoS, University of Sassari, Sassari, Sardinia, Italy

Fatih Deyneli Department of Public Finance, Pamukkale University, Denizli, Turkey

Viviana Di Giovinazzo Department of Sociology and Social Research, University of Milano Bicocca, Milan, Italy

Andrea Di Liddo Department of Economics, University of Foggia, Foggia, Italy

Giuseppe Di Vita Department of Economics and Business, University of Catania, Catania, Italy

Mark A. Dijkstra Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

Robert J. Dijkstra Tilburg Institute for Private Law, Tilburg University, Tilburg, The Netherlands

Alexander Dilger Institute for Organisational Economics, University of Münster, Münster, Germany

Antony W. Dnes Northcentral University, Arizona, AZ, USA

David Dolejší Faculty of Social and Economic Studies, Jan Evangelista Purkyně University, Ústí nad Labem, Czech Republic

Goran Dominioni Rotterdam Institute of Law and Economics, Erasmus School of Law, Rotterdam, The Netherlands

John J. Donohue III Stanford Law School and National Bureau of Economic Research, Stanford, CA, USA

Myriam Doriat-Duban Department of Economics, BETA UMR 7522, Université de Lorraine, Nancy, France

Thomas Eger Faculty of Law, Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Michael Eichenseer University of Regensburg, Regensburg, Germany

Nora El-Bialy Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Javier Elizalde Department of Economics, University of Navarra, Pamplona, Navarra, Spain

Alfred Endres FernUniversität Hagen, Hagen, Germany

Peter-Jan Engelen Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

Martha M. Ertman Francis King Carey School of Law, University of Maryland, Baltimore, MD, USA

María Paz Espinosa University of the Basque Country, UPV/EHU, Bilbao, Spain

Giuseppe Eusepi Department of Law and Economics of Productive Activities, Sapienza University of Rome, Faculty of Economics, Rome, Italy

Francois Facchini Faculté Jean Monnet, University of Paris 11 (RITM) and Associate Economist, Centre d'Economie de la Sorbonne, University of Paris 1 (France), Sceaux, France

Valeria Faralla Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università del Piemonte Orientale, Alessandria, Italy

Joëlle Farchy Faculty of Economics, Pantheon-Sorbonne University Paris 1, Paris, France

Anne Catherine Faye Analysis Group, Montréal, Canada

Silvia Fedeli Department of Economics and Law, Sapienza - University of Rome, Rome, Italy

Claudio Feijóo Sino-Spanish Campus, Tongji University - Technical University of Madrid, Shanghai, China

Samuel Ferey CNRS, BETA, University of Lorraine, Nancy, France

Cyrille Ferraton Université Montpellier 3, Montpellier, France

Sergio Pinheiro Firpo Insper Institute of Education and Research, São Paulo, Brazil

Jean-Baptiste Fleury THEMA, University of Cergy-Pontoise, Cergy-Pontoise, France

Mataka P. Flynn Faculty of Law, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Francesco Forte Department of Economics and Law, Sapienza - University of Rome, Rome, Italy

Panagiotis N. Fotis Hellenic Competition Commission, Athens, Greece

Luigi Alberto Franzoni University of Bologna, Bologna, Italy

Tim Friehe Public Economics Group, Philipps-University Marburg, Marburg, Germany

Ludovic Frobert Centre National de la Recherche Scientifique, Paris, France

Yannick Gabuthy BETA, CNRS, University of Strasbourg, Strasbourg, France

University of Lorraine, Nancy, France

Alicia García-Herrera Ilustre Colegio de Abogados de Valencia, Valencia, Spain

Isabel-María García-Sánchez Universidad de Salamanca, Salamanca, Spain

Bianca Gardella Tedeschi DiSEI, Dipartimento per lo Studio dell'Economia e dell'Impresa, Università del Piemonte Orientale, Novara, Italy

Dustin Garlitz Department of Philosophy, University of South Florida, Tampa, FL, USA

Martin Gelter Fordham University School of Law, New York, NY, USA

Nikolaos Georgantzis Economic and Social Sciences, School of Agriculture Policy and Development, University of Reading, Reading, Berkshire, UK
LEE and Economics Department, Universitat Jaume I, Castellon, Spain

Federica Gerbaldo IEL- Institutions, Economics and Law, Collegio Carlo Alberto, University of Torino, Turin, Italy

Daniel Gervais Vanderbilt Intellectual Property Program, Vanderbilt University Law School, South Nashville, TN, USA

Benny Geys Department of Economics, Norwegian Business School BI, Oslo, Norway

Vrije Universiteit Brussel, Ixelles, Belgium

Giuseppina Gianfreda University of Tuscia, Viterbo, Italy

Katrin Gödker School of Business, Economics and Social Sciences, University of Hamburg, Hamburg, Germany

Laszlo Goerke IAAEU (Institute for Labour Law and Industrial Relations in the European Union), University Trier, Trier, Germany

IZA, Bonn, Germany

CESifo, Munich, Germany

José Luis Gómez-Barroso Dpto. Economía Aplicada e Historia Económica, UNED (Universidad Nacional de Educación a Distancia), Madrid, Spain

Brady P. Gordon Trinity College, University of Dublin, Dublin, Ireland

Kateryna Grabovets Rotterdam Business School, Rotterdam University of Applied Sciences, Rotterdam, The Netherlands

Gerhard Graf Johamer Qutenberg-Universitat Maine, Nieder-Olm, Germany

Kristoffel Grechenig Max Planck Institute for Research on Collective Goods, Bonn, Germany

Luciano Greco Department of Economics and Management and CRIEP, University of Padua, Padua, Italy

Veronica Grembi Copenhagen Business School, Copenhagen, Denmark

Gilles Grolleau Supagro, UMR 1135 LAMETA, Montpellier, France

Burgundy School of Business – LESSAC, Dijon, France

Pauline Grosjean School of Economics, University of New South Wales, Sydney, NSW, Australia

Sven Grüner Agribusiness Management, Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Brishti Guha Centre of International Trade and Development, School of International Studies, Jawaharlal Nehru University, New Delhi, India

Sonja Elisabeth Haberl University of Ferrara, Ferrara, Italy

Paul Hallwood Department of Economics, University of Connecticut, Storrs, CT, USA

Philip C. Hanke Department of Public Law, University of Bern, Bern, Switzerland

Michael Hantke-Domas Centre for Water Law, Policy and Science, University of Dundee, Scotland, United Kingdom

Third Environment Court, Valdivia, Chile

Sophie Harnay EconomiX CNRS UMR 7235, University Paris Ouest Nanterre La Defense, Nanterre, France

Ozan Hatipoglu Department of Economics, Bogazici University, Istanbul, Turkey

Aristides Hatzis Department of History and Philosophy of Science, National and Kapodistrian University of Athens, Athens, Greece

Klaus Heine Rotterdam Institute of Law and Economics, Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Jenny Helstroffer BETA, Université de Lorraine, Nancy, France

Oystein Hernæs Department of Economics, European University Institute, Florence, Italy
Norwegian Social Research, Oslo, Norway

Anthony Heyes University of Ottawa, Ottawa, Canada

Norbert Hirschauer Agribusiness Management, Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Geoffrey M. Hodgson Hertfordshire Business School, University of Hertfordshire, Hatfield, Hertfordshire, UK

Sven Hoepfner Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium
Guest Researcher, Max Planck Institute for Research on Collective Goods, Bonn, Germany

Manfred J. Holler Center of Conflict Resolution (CCR), University of Hamburg, Institute of SocioEconomics (ISE), Hamburg, Germany

Lars Hornuf Department of Economics, Trier University, Trier, Germany
CESifo, Max Planck Institut for Innovation and Competition, Munich, Germany

Herbert Hovenkamp The College of Law, The University of Iowa, Iowa City, IA, USA

Laura Huici Sancho Department of International Law and Economics, University of Barcelona, Barcelona, Spain

Keith N. Hylton Boston University School of Law, Boston, MA, USA

Ari Hyytinen Jyväskylä School of Business and Economics, University of Jyväskylä, Jyväskylä, Finland

Lisette Ibanez LAMETA, CNRS, INRA, Montpellier SupAgro, Université de Montpellier, Montpellier, France

Eka Ikpe King's College London, London, UK

Giovanni Immordino Department of Economics and Statistics, University of Naples Federico II and CSEF, Naples, Italy

Marta Infantino University of Trieste, Trieste, Italy

Roberto Ippoliti Department of Management, University of Turin, Turin, Italy

Scientific Promotion, Hospital of Alessandria – “SS. Antonio e Biagio e Cesare Arrigo”, Alessandria, Italy

Karl-Friedrich Israel Faculty of Economics and Business Administration, Humboldt-Universität zu Berlin, Berlin, Germany

Department of Law, Economics, and Business, University of Angers, Angers, France

Hadar Yoana Jabotinsky Faculty of Law, Hebrew University of Jerusalem, Jerusalem, Israel

Niklas Jakobsson Oslo and Akershus University College of Applied Sciences, Norwegian Social Research and Karlstad University, Oslo, Norway

Bruno Jeandidier Bureau d'Economie Théorique et Appliquée (BETA), CNRS and University of Lorraine, Nancy, France

J. L. Jiménez Facultad de Economía, Empresa y Turismo, Departamento de Análisis Económico Aplicado. Despacho D.2-12, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Jean-Michel Josselin University of Rennes 1 and CREM UMR-CNRS, Rennes, France

Juha-Pekka Kallunki Oulu Business School, University of Oulu, Oulu, Finland

Sandeep Kapur Birkbeck College, University of London, London, UK

Vaia Karapanou Athens, Greece

Dennis W. K. Khong Centre for Law and Technology, Faculty of Law, Multimedia University, Melaka, Malaysia

Iljoong Kim Department of Economics, SungKyunKwan University (SKKU), Jongno-Gu, Seoul, South Korea

Jeong-Yoo Kim Economics, Kyung Hee University, Seoul, Republic of Korea

Christopher Kingston Amherst College, Amherst, MA, USA

Haksoo Ko School of Law, Seoul National University, Seoul, Republic of Korea

Stefan Kolev Wilhelm Röpke Institute, Erfurt and University of Applied Sciences Zwickau, Zwickau, Germany

Georgia Kontogeorga International Board of Auditors for NATO (IBAN), Brussels, Belgium

University of Paris 1- Pantheon Sorbonne, Paris, France

Andreas Kotsadam Department of Economics, University of Oslo, Oslo, Norway

Mitja Kovac Department of Economic Theory and Policy, Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

Pavel Kuchař Department of Economics and Finance, University of Guanajuato, DCEA-Sede Marfil, Guanajuato, GTO, Mexico

Facultad de Economía-División de Estudios de Posgrado, Universidad Nacional Autónoma de México, Ciudad de México, México

Joanna Kuczevska Faculty of Economics, University of Gdansk, Sopot, Poland

Andreas P. Kyriacou Universitat de Girona, Girona, Spain

Helfried Labrenz Institut für Unternehmensrechnung, Finanzierung und Besteuerung, Wirtschaftswissenschaftliche Fakultät, Universität Leipzig, Leipzig, Germany

LSI Leipziger Spieltheoretisches Institut, Leipzig, Germany

Eve-Angéline Lambert BETA, CNRS, University of Strasbourg, Strasbourg, France

BETA UMR 7522, Université de Lorraine, Nancy, France

Fabio Landini LUISS University, Rome, Italy

Pierre-Carl Langlais Université Paris IV – Sorbonne, Paris, France

Régis Lanneau Law School, CRDP, Université de Paris Nanterre, Nanterre, France

David Allen Larson Hamline University School of Law, Saint Paul, MN, USA

Shay Lavie The Buchmann Faculty of Law, Tel Aviv University, Tel Aviv, Israel

Kangoh Lee Department of Economics, San Diego State University, San Diego, CA, USA

Yong-Shik Lee The Law and Development Institute, Decatur, GA, USA

Dennis Leech University of Warwick, Coventry, UK

Peter T. Leeson Department of Economics, George Mason University, Fairfax, VA, USA

Martin A. Leroch Politics and Economy, Johannes Gutenberg Universität Mainz, Mainz, Germany

Jacek Lewkowicz Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland

Ines Lindner Department of Econometrics and OR, VU University Amsterdam, Amsterdam, The Netherlands

Jason M. Lindo Texas A&M University, College Station, TX, USA
NBER, Cambridge, MA, USA
IZA, Bonn, Germany

Maurizio Lisciandra Department of Economics, University of Messina, Messina, Italy

Michael Litschka Department of Media Economics, University of Applied Sciences St. Pölten, St. Pölten, Austria

Rafael Llorca-Vivero University of Valencia, Valencia, Spain

Emanuele Lo Gerfo Economics, Management and Statistics Department, University of Milano Bicocca, Milan, Italy

C. Locatelli GAEL, CNRS, Grenoble, INP, INRA, University of Grenoble-Alpes, Grenoble, France

Gianna Lotito DiGSPES, University of Eastern Piedmont, Alessandria, Italy

Adrien Lutz GATE L-SE, Université Jean Monnet, Saint-Étienne, France

Ronny Müller Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

Magdalena Malecka TINT – Centre for Philosophy of Social Science, University of Helsinki, Helsinki, Finland

Rafael Emmanuel Macatangay Centre for Energy, Petroleum and Mineral Law and Policy School of Social Sciences, University of Dundee, Dundee, Scotland, UK

Anna Maffioletti ESOMAS, University of Turin, Turin, Italy

Marie-Helen Maras John Jay College of Criminal Justice, City University of New York, New York, NY, USA

Miriam Marcén Universidad de Zaragoza, Zaragoza, Spain

Carla Marchese Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, POLIS Institute, Università degli Studi del Piemonte Orientale ‘Amedeo Avogadro’, Alessandria, Italy

Alain Marciano MRE and University of Montpellier, Montpellier, France
Faculté d’Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France

Stephan Marette UMR Economie Publique INRA, Paris, France

Wim Marneffe Faculty of Applied Economics, Hasselt University, Diepenbeek, Belgium

Frédéric Marty CNRS - UMR 7321 GREDEG, Université Côte d’Azur, Université Nice Sophia Antipolis, Valbonne, France

Donato Masciandaro Department of Economics, Bocconi University, Milan, Italy

Nicola Matteucci Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy

Karsten Mause Department of Political Science (IfPol), University of Münster, Münster, Germany

Bryan McCannon Saint Bonaventure University, St Bonaventure, NY, USA

Steven G. Medema Department of Economics, University of Colorado Denver, Denver, CO, USA

Daniel Meierrieks Department of Economics, University of Freiburg, Freiburg, Germany

Alessandro Melcarne EconomiX, CNRS and Université Paris Nanterre, Nanterre, France

Katarzyna Metelska-Szaniawska Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland

Mehdi Mezaguer CNRS – GREDEG, Université Côte d’Azur, Université Nice Sophia Antipolis, Valbonne, France

Thomas J. Miceli Department of Economics, University of Connecticut, Storrs, CT, USA

Matteo Migheli Università di Torino, Torino, Italy
CeRP-Collegio Carlo Alberto, Torino, Italy
IEL, Torino, Italy

Michelle D. Miranda Farmingdale State College State, University of New York, Farmingdale, NY, USA

Michael Mitsopoulos Brookings Institution, Athens, Greece

Sergio Mittlaender Max Planck Institute for Social Law and Social Policy, Munich, Germany

Franklin G. Mixon Jr. Center for Economic Education, Columbus State University, Columbus, GA, USA

Anne-Marie Mohammed Department of Economics, Faculty of Social Sciences, The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

Pier Giuseppe Monateri SciencesPo, Ecole de droit, University of Torino, Torino, Italy

Josef Montag International School of Economics, Kazakh-British Technical University, Almaty, Kazakhstan
Faculty of Law, Charles University, Prague, Czech Republic

Sylvia Morawska Warsaw School of Economics, Collegium of Business Administration, Warszawa, Poland

David Moroz Champagne School of Management, Groupe ESC Troyes, France

P. Sean Morris Faculty of Law, University of Helsinki, Helsinki, Finland

Nathalie Moureau Université Paul-Valéry MONTPELLIER 3, Montpellier, France

ARTdev, UMR 5281, Université Paul Valéry, Montpellier, France

Naoufel Mzoughi INRA, UR 767 Ecodéveloppement, Avignon, France

Giulio Napolitano Law Department, Roma Tre University, Rome, Italy

Gaia Narciso Trinity College Dublin, Dublin, Ireland

Jannik A. Nauerth Faculty of Business and Economics, University of Technology Dresden, Dresden, Germany

Yew-Kwang Ng Department of Economics, Nanyang Technological University, Singapore, Singapore

School of Economics, Fudan University, Shanghai, China

Daniel Nientiedt Walter Eucken Institute, Freiburg, Germany

Nirjhar Nigam ICN Business School, CEREFIGE and LARGE Laboratory, Metz, France

Marco Novarese Department of Law and Economics, University of Eastern Piedmont, Centre for Cognitive Economics, Alessandria, Italy

Jacob Nussim Faculty of Law, Bar-Ilan University, Ramat-Gan, Israel

Marie Obidzinski CRED, Université Panthéon-Assas Paris II, Paris, France

Cornelia Ohl HA Hessen Agentur GmbH, Transferstelle für Emissionshandel und Klimaschutz, Wiesbaden, Germany

J. M. Ordóñez-de-Haro Departamento de Teoría e Historia Económica y Cátedra de Política de Competencia, Universidad de Málaga, Málaga, Spain

Guido Ortona Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università del Piemonte Orientale, Alessandria, Italy

Elisabetta Ottoz Department of Economics and Statistics Cognetti De Martiis, University of Turin, Turin, Italy

Alessio M. Paccès Erasmus School of Law, Rotterdam Institute of Law and Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

Ugo Pagano University of Siena, Siena, Italy
Central European University, Budapest, Hungary

Mika Pajarinen The Research Institute of the Finnish Economy (ETLA), Helsinki, Finland

Martin Peitz Department of Economics, University of Mannheim, Mannheim, Germany

Theodore Pelagidis Brookings Institution, University of Piraeus, Piraeus, Attica, Greece

J. Perdiguero Departament d'Economia Aplicada, Research Group of "Economia Aplicada" (GEAP), Universitat Autònoma de Barcelona, Bellaterra, Spain

Ennio Piano Department of Economics, George Mason University, Fairfax, VA, USA

Barbara Piattoli Girard Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università degli Studi del Piemonte Orientale "Amedeo Avogadro", Alessandria, Italy

Joshua Pierson Department of Economics, George Mason University, Fairfax, VA, USA

Pasquale Pistone University of Salerno, Salerno, Italy

IBFD, Amsterdam, The Netherlands

WU Vienna University of Economics, Vienna, Austria

Marta Podemska-Mikluch Faculty of Economics, Gustavus Adolphus College, Saint Peter, MN, USA

Paolo Polidori Department of Law, University of Urbino, Urbino, Italy

Ferruccio Ponzano Department of Law and Political, Economic and Social Sciences, University of Piemonte Orientale, Alessandria, Italy

Donatella Porrini Department of Management, Economics, Mathematics and Statistics, University of Salento, Lecce, Italy

Aurelien Portuese University of Westminster, London, UK

J. J. Prescott Law School, University of Michigan, Ann Arbor, MI, USA

Derek Pyne Thompson Rivers University, Kamloops, BC, Canada

Elena Quercioli Department of Economics, College of Business Administration, Central Michigan University, Mt. Pleasant, MI, USA

Laurie Rachet Jacquet Hospinnomics, Paris School of Economics, Paris, France

Véronique Raimond Haute Autorité de santé, Saint-Denis, France

Shruti Rajagopalan Department of Economics, State University of New York, Purchase College, Anderson, NY, USA

Juan Ramón Rallo OMMA Center of Studies, Madrid, Spain

Giovanni Battista Ramello DiGSPES, University of Eastern Piedmont, Alessandria, Italy

IEL, Torino, Italy

Ringa Raudla Faculty of Social Sciences, Ragnar Nurkse School of Innovation and Governance, Tallinn University of Technology, Tallinn, Estonia

Tim Reuter RBB Economics, Brussels, Belgium

Marco Ricolfi Department of Legal Studies, Department of Law, Turin University, Turin, Italy

Mario J. Rizzo Department of Economics, New York University, New York, NY, USA

Matteo Rizzolli LUMSA University, Rome, Italy

Lise Rochaix Hospinnomics, Paris School of Economics, Paris, France

Fabrice Rochelandet IRCAV, Université Sorbonne Nouvelle Paris 3, Paris Cedex 05, France

Antonio Rodriguez Andres Departamento de ADE y Economía, Universidad Camilo José Cela, Villanueva de la Cañada, Madrid, Spain

Claudia Rodella Department of Psychology, Catholic University of the Sacred Heart, Brescia, Italy

Rustam Romaniuc Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier; International programme in institutions, economics and law (IEL), University of Turin, Montpellier, France

Alessandro Romano China-EU School of Law, China University of Political Science and Law, Beijing, China

Caspar Rose Copenhagen Business School, Copenhagen, Denmark

Maria Alessandra Rossi Department of Economics and Statistics, University of Siena, Siena, Italy

Sébastien Roussel Department of Economics, CEE-M, Université Montpellier, CNRS, INRA, SupAgro, Université Paul Valéry Montpellier 3, Montpellier, France

Silvia Ručinská Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

Luca Rubini University of Birmingham, Birmingham, UK

Francesco F. Russo Department of Economics and Statistics, University of Naples Federico II and CSEF, Naples, Italy

Matt E. Ryan Department of Economics, Duquesne University, Pittsburgh, PA, USA

Amit M. Sachdeva Ernst & Young LLP, Houston, TX, USA

New York University School of Law (NYU), New York, USA

London School of Economics (LSE), London, UK

The Hague Academy of International Law, The Hague, The Netherlands

Pablo Salvador Coderch Universitat Pompeu Fabra, Barcelona, Spain

Margherita Saraceno DEMS, University of Milano-Bicocca and BAFFI CAREFIN Centre, Università Bocconi, Milan, Italy

George Saridakis Small Business Research Centre, Business School, Kingston University, London, UK

Paola Savini Autorità per le garanzie nelle comunicazioni – AGCOM, Rome, Italy

Pasquale L. Scandizzo CEIS (Centre for International Economic Studies), University of Rome “Tor Vergata”, Rome, Italy

Hans-Bernd Schäfer Bucerius Law School, Hamburg, Germany
Faculty of Law, St. Gallen University, St. Gallen, Switzerland

Jessamyn Schaller The University of Arizona, Tucson, AZ, USA

Sebastian Scheerer Institute for Criminological Social Research, Faculty of Economics and Social Sciences, University of Hamburg, Hamburg, Germany

Friedrich Schneider Department of Economics, Johannes Kepler University of Linz, Linz-Auhof, Austria

Martin Schneider Faculty for Economics and Business Administration, University of Paderborn, Paderborn, Germany

Jan Schnellenbach Brandenburgische Technische Universität Cottbus-Senftenberg Chair for Microeconomics, Cottbus, Germany

Philipp Schumacher Wildau Institute of Technology, Technical University of Applied Science, Wildau, Brandenburg, Germany

Otmar Schuster Haus der Geoinformation, Mülheim an der Ruhr, Germany

Kathleen Segerson Department of Economics, University of Connecticut, Storrs, CT, USA

Régis Servant PHARE, University Paris 1 Panthéon-Sorbonne, Paris, France

Ram Singh Department of Economics, Delhi School of Economics, University of Delhi, Delhi, India

Emiliano Sironi Department of Statistical Sciences, Catholic University of Milan, Milan, Italy

Lones Smith Department of Economics, University of Wisconsin, Madison, WI, USA

Nicholas A. Snow Department of Economics, Indiana University, Bloomington, IN, USA

Giuseppe Sobbrío Department of Economics, University of Messina, Messina, Italy

Tomáš Sobek Faculty of Law, Masaryk University, Brno, Czech Republic

Philippe Solal CNRS UMR 5824 GATE, Université de Saint-Etienne, Saint-Etienne, France

Sandra Sookram Sir Arthur Lewis Institute of Social and Economic Studies, The University of The West Indies, St. Augustine, Trinidad and Tobago, West Indies

Alessandro Sterlacchini Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy

Anna Sting International and European Union Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Edward Peter Stringham Economic Organizations and Innovation, Trinity College, Hartford, CT, USA

American Institute for Economic Research, Great Barrington, MA, USA

S. I. Strong University of Missouri School of Law, Columbia, MO, USA

Marianna Succurro Department of Economics, Statistics and Finance, University of Calabria, Rende (CS), Italy

Silvio H. T. Tai PPGE, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

RITM, Université Paris-Sud, Paris, France

Léa Tardieu BETA, Université Lorraine, INRA, AgroParisTech, Nancy, France

Valerio Tavormina Università Cattolica del Sacro Cuore, Milan, Italy

Désirée Teobaldelli Department of Law, University of Urbino, Urbino, Italy

Henk J. ter Bogt University of Groningen, Groningen, The Netherlands

Antoni Terra Ibáñez Stanford Law School, Stanford, CA, USA

Andrew Torre School of Accounting, Economics and Finance, Deakin University, Melbourne, VIC, Australia

Jacopo Torriti University of Reading, Reading Berkshire, UK

Michael Trebilcock University of Toronto, Toronto, ON, Canada

George Tridimas Department of Accounting, Finance and Economics, Ulster University Business School, Belfast, Northern Ireland

Aspasia Tsaoussi School of Law, Aristotle University of Thessaloniki, Thessaloniki, Greece

Fav Tsoin Lai National Chi Nan University, Puli, Taiwan

George Tsourvakas School of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, Greece

Jaime Vallés-Giménez University of Zaragoza, Zaragoza, Spain

Remus Valsan Edinburgh Law School, University of Edinburgh, Edinburgh, Scotland, UK

Hanne Van Cappellen Institute of Development Policy (IOB), University of Antwerp, Antwerp, Belgium

Joël J. van der Weele Department of Economics, University of Amsterdam, Amsterdam, The Netherlands

Rogier J. van der Wolk Faculty of Law, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Faculty of Law, Universiteit van Amsterdam, Amsterdam, The Netherlands

Robert Van Horn Economics Department, University of Rhode Island, Kingston, RI, USA

Luc Van Liedekerke Faculty of Applied Economics, University of Antwerp, Antwerpen, Belgium

Bruno van Pottelsberghe de la Potterie Solvay Brussels School of Economics and Management, Université libre de Bruxelles, Brussels, Belgium

Ben C. J. van Velthoven Leiden Law School, Leiden University, Leiden, The Netherlands

Peter W. van Wijck Leiden Law School, Leiden University, Leiden, The Netherlands

Ann-Sophie Vandenberghe Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Marco Vannini Department of Economics and CRENoS, University of Sassari, Sassari, Italy

Juan F. Vargas Department of Economics, Universidad del Rosario, Bogota, Colombia

Massimiliano Vatterio “Brenno Galli” Chair of Law and Economics, Institute of Law (IDUSI), Università della Svizzera italiana (USI), Lugano, Switzerland

Roland Vaubel Universität Mannheim, Mannheim, Germany

Patrick Velte Leuphana University, Lüneburg, Germany

Lode Vereeck Faculty of Applied Economics, Hasselt University, Diepenbeek, Belgium

Louis Visscher Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Stefan Voigt Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Uwe Vollmer University of Leipzig, Leipzig, Germany

Richard E. Wagner Department of Economics, George Mason University, Fairfax, VA, USA

Andre Wakefield Pitzer College, Claremont Colleges, Claremont, CA, USA

Bruce Wardhaugh School of Law, Queen's University Belfast, Belfast, UK

Franziska Weber Faculty of Law, Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Wolfgang Weigel Faculty of Business, Economics and Statistics, Department of Economics, Joseph von Sonnenfels Center for the Study of Public Law and Economics, Vienna, Austria

Paul-Ludwig Weinacht Güntersleben, Germany
Würzburg University, Würzburg, Germany

Mark D. White Department of Philosophy, College of Staten Island/CUNY, Staten Island, NY, USA

Michelle J. White University of California, San Diego, La Jolla, California, USA

Sam Whitt High Point University, High Point, NC, USA

Edwin Woerdman Faculty of Law, Department of Law and Economics, University of Groningen, Groningen, The Netherlands

Garrett Wood Department of Economics, George Mason University, Fairfax, VA, USA

Alexander J. Wulf SRH Hochschule Berlin, Berlin, Germany

Jen-Te Yao Fu-Jen Catholic University, New Taipei City, Taiwan

Oleg Yerokhin University of Wollongong, Wollongong, Australia

Luciana Yeung Insper Institute of Education and Research, São Paulo, Brazil

Qianwei Ying Business School, Sichuan University, Chengdu, Sichuan, China

Martin Zagler UPO University of Eastern Piedmont, Novara, Italy
WU Vienna University of Economics, Vienna, Austria

Katarina Zajc Laws School, University of Ljubljana, Ljubljana, Slovenia

Anabel Zárate-Marco University of Zaragoza, Zaragoza, Spain

Ainhoa Zarraga University of the Basque Country, UPV/EHU, Bilbao, Spain

Nikolaos E. Zevgolís Hellenic Competition Commission, Athens, Greece

Michel S. Zouboulakis Department of Economics, University of Thessaly, Volos, Greece

Todd J. Zywicki Antonin Scalia Law School, George Mason University, Arlington, VA, USA

Absolute Priority

Alexander Dilger
Institute for Organisational Economics,
University of Münster, Münster, Germany

Synonyms

[Liquidation Preference](#)

Definition

Absolute priority is a rule by which different classes of creditors are paid in full one after another in a corporate insolvency and the owners are paid last if anything remains.

Explanation

Absolute priority means that in a corporate insolvency, different classes of creditors are paid one after another, while the (former) owners of an insolvent firm get the rest if anything remains after all creditors have been paid in full. Normally this means that the owners get nothing because otherwise the company would not be insolvent in the first place. Accordingly, absolute priority is not relevant for solvent companies because they pay everything they owe their creditors and the owners own the rest. If a company cannot pay

everything it owes because its liabilities are higher than the value of all its assets, then it is insolvent and all its debts are due immediately. Following absolute priority the most senior creditors are paid first. If the value of the insolvent firm is high enough, the most senior creditors are paid in full and the second highest class of creditors is next. Otherwise, if the value of the insolvent firm is lower than all the claims of the most senior creditors, these claims are satisfied proportionally and all other creditors with lower priority as well as the owners get nothing. In general, one class of creditors is partially satisfied, while all more senior creditors are satisfied completely and all more junior creditors get nothing. Conversely, it is a violation of absolute priority if any junior creditor or an owner gets anything before the more senior claims have been paid in full.

It is quite easy to comply with absolute priority as long as the value of the insolvent company is known. This is especially the case after the liquidation of a firm when all its assets have been sold. However, in many cases, the value of a firm is higher as a going concern. When it is sold as a whole, the total revenue is known and can be distributed according to absolute priority. Yet the market for insolvent companies may be thin such that only a fire sale with a depressed price is possible (cf. Gale and Gottardi 2011) and it is better for all creditors together to continue the business by themselves at least for a while. In this case absolute priority requires an estimation of the company's value. In estimating this value, there is a conflict

of interest because the most senior creditors get a higher share of this company when the valuation is low, whereas the more junior creditors prefer a higher valuation to get more (than nothing). Bebchuk (1988) proposes the use of options as a fair solution to this problem. The most senior creditors get all new shares of the then debt-free company, while more junior creditors or even owners can buy shares by paying the corresponding ratio of more senior claims. Dilger (2006) shows that one auction of all shares is both simpler and fairer than Bebchuk's approach with options.

Besides the question on how to observe absolute priority, one may ask why it is important. That creditors should be more senior than owners in a bankruptcy is part of any meaningful debt contract. The differences between creditors (or also different categories of owners) such that some are more senior than others can also be arranged by contract or regulated by law. Violations of absolute priority are then a breach of these contracts or laws. Thus, they limit the freedom of contract and are inefficient *ex ante* (cf. Longhofer 1997; Bebchuk 2002). However, strictly maintaining absolute priority can be inefficient *ex post* if there are costly conflicts while determining the value of a company (cf. Baird and Bernstein 2006) or if nobody has sufficient incentives to begin or end a bankruptcy procedure (cf. Baird 1991). Empirically one can observe many deviations from absolute priority (cf. Weiss 1990). They seem to be fair and efficient as long as the more senior creditors agree. Thus, a veto power of them, individually or even by majority rule by creditor classes, against violations of absolute priority is better than a strict adherence to it in any case (cf. Warren 1991). Although there is much variation, the bankruptcy laws in many countries are in accordance with this (cf. Campbell 1992; Claessens and Klapper 2005).

References

- Baird DG (1991) The initiation problem in bankruptcy. *Int Rev Law Econ* 11:223–232
- Baird DG, Bernstein DS (2006) Absolute priority, valuation uncertainty, and the reorganization bargain. *Yale Law J* 115:1930–1970

- Bebchuk LA (1988) A new approach to corporate reorganizations. *Harv Law Rev* 101:775–804
- Bebchuk LA (2002) *Ex ante* costs of violating absolute priority in bankruptcy. *J Financ* 57:445–460
- Campbell D (ed) (1992) *International corporate insolvency law*. Butterworths Law, London
- Claessens S, Klapper LF (2005) Bankruptcy around the world: explanations of its relative use. *Am Law Econ Rev* 7:253–283
- Dilger A (2006) Forced to make mistakes: reasons for complaining about Bebchuk's scheme and other market-oriented insolvency procedures. *Eur J Law Econ* 21:79–94
- Gale D, Gottardi P (2011) Bankruptcy, finance constraints, and the value of the firm. *Am Econ J Microecon* 3:1–37
- Longhofer SD (1997) Absolute priority rule violations, credit rationing, and efficiency. *J Financ Intermed* 6:249–267
- Warren E (1991) A theory of absolute priority. *NY Univers Ann Surv Am Law* 9:9–48
- Weiss LA (1990) Bankruptcy resolution: direct costs and violation of priority of claims. *J Financ Econ* 27:285–314

Abuse of Dominance

Pinar Akman

School of Law, University of Leeds, Leeds, UK

Abstract

Article 102 of the Treaty on the Functioning of the European Union (TFEU) prohibits the abuse of a dominant position within the internal market or in a substantial part of it, as incompatible with the internal market, insofar as it may affect trade between Member States. This essay examines the constituent elements of the prohibition found in Article 102 TFEU. These elements are: (i) one or more undertakings of a dominant position within the internal market or in a substantial part of it; (ii) effect on trade between Member States; (iii) abuse.

Definition

Article 102 of the Treaty on the Functioning of the European Union (TFEU) prohibits any abuse by one or more undertakings of a dominant position

within the internal market or in a substantial part of it, as incompatible with the internal market, insofar as it may affect trade between Member States. Article 102 TFEU also provides examples of abusive practices. This entry examines the constituent elements of the prohibition found in Article 102 TFEU.

Introduction

Article 102 of the Treaty on the Functioning of the European Union (TFEU) prohibits the abuse of a dominant position within the internal market or in a substantial part of it, as incompatible with the internal market, insofar as it may affect trade between Member States. This entry examines the constituent elements of the prohibition found in Article 102 TFEU. These elements are (i) one or more undertakings of a dominant position within the internal market or in a substantial part of it, (ii) effect on trade between Member States, and (iii) abuse.

One or More Undertakings of a Dominant Position Within the Internal Market or in a Substantial Part of It

The Concept of an “Undertaking”

First and foremost, it is clear from Article 102 TFEU that a dominant position can only be held by “one or more undertakings” for the purposes of that provision. An “undertaking” for the purposes of EU competition law is “every entity engaged in an economic activity regardless of the legal status of the entity and the way in which it is financed” (Case C-41/90 *Hofner and Elser v Macrotron GmbH* [1991] ECR I-1979 [21]). In turn, “economic activity” is defined as “any activity consisting in offering goods or services on a given market” (Cases C-180/98 etc *Pavel Pavlov and others v Stichting Pensioenfonds Medische Specialisten* [2000] ECR I-6451 [75]). A “functional approach” is adopted when determining whether an entity is an “undertaking” for the purposes of competition law, meaning that the same legal entity may be acting as an undertaking

when carrying on one activity but not when carrying on another activity (Whish and Bailey 2012, 84–85). Similarly, the fact that the entity in question does not have a profit motive or economic purpose is irrelevant in establishing whether it is engaged in “economic activity” (Case C-67/96 etc *Albany International BV v SBT* [1999] ECR I-5751 [85]; Case 155/73 *Italy v Sacchi* [1974] ECR 409 [13–14]). Activities that are *not* economic are those provided on the basis of “solidarity” and those that consist in the exercise of public power and procurement pursuant to a noneconomic activity (Whish and Bailey 2012, 87 et seq). “Solidarity” in turn is defined as “the inherently uncommercial act of involuntary subsidisation of one social group by another” (Opinion of Advocate General Fennelly in Case C-70/95 *Sodemare v Regione Lombardia* [1997] ECR I-3395 [29]). The case law on this has mostly been concerned with national health services, compulsory insurance schemes, pension schemes, etc., as well as provision of services by public authorities which fulfill essential functions of the State.

Dominant Position

Dominant position is not defined in the Treaty. As is clear from Article 102 TFEU, it can be held by one or more undertakings. Where the dominant position is held by more than one undertaking, they would have a position of “collective dominance.” The first step in establishing whether an undertaking enjoys a dominant position for the purposes of Article 102 TFEU is to define the “relevant market.” This is because a dominant position cannot be held in the abstract but can only be held over a relevant market.

Market Definition

Defining the relevant market delineates the products or services that are in competition and enables one to gauge how much power an undertaking has over its competitors and consumers (Rodger and MacCulloch 2015, 95). For the purposes of competition law, the relevant market has to be defined with regard to product, geography, and occasionally time. According to the Court of Justice of the European Union (CJEU), the

definition of the market is one of establishing interchangeability: if products or services are seen as interchangeable, then they are part of the same market (e.g., Case 6/72 *Europemballage Corporation and Continental Can Company Inc v EC Commission* [1973] ECR 215 [32]). The European Commission has published a Notice on the definition of the relevant market ([1997] OJ C 372/5). According to the Notice, the main purpose of market definition is to identify in a systematic way the competitive constraints that the relevant undertakings face (Notice [2]). The objective of defining a market in both its product and geographic dimension is to identify those actual competitors of the undertaking involved that are capable of constraining that undertaking's behavior and of preventing it from behaving independently of effective competitive pressure (Notice [2]). Consequently, market definition makes it possible inter alia to calculate market shares that would convey meaningful information regarding market power for the purposes of assessing dominance (Notice [2]).

There are three main sources of competitive constraints on an undertaking's conduct: demand substitutability, supply substitutability, and potential competition (Notice [13]). In the Notice, regarding demand substitutability, the Commission adopts the so-called "hypothetical monopolist" test. This test asks whether a hypothetical small but significant non-transitory increase in price (SSNIP) of the product produced by the undertaking under investigation would lead to customers switching to other products (Notice [17]). The range of such an increase is usually 5–10%. If customers would switch to other products following such an increase, then the product of the undertaking under investigation and those other products to which the customers would switch are considered to be in the same product market (Notice [18]). Supply substitutability seeks to establish if other suppliers would switch their production to the products of the undertaking under investigation and market them in the short term without incurring significant additional costs, if there was a small and permanent increase in the price of the product under investigation (Notice [20]). However, the Notice stipulates

that the Commission would take into account supply substitutability only where its effects are equivalent to those of demand substitution in terms of effectiveness and immediacy (Notice [20]). Other than the relevant product market, the relevant geographical market has to be also defined. According to the CJEU, the geographical market is limited to an area where the objective conditions of competition applying to the product in question are sufficiently homogenous for all traders (Case 27/76 *United Brands Continentaal BV v EC Commission* [1978] ECR 207 [11]). In certain cases, it may be necessary to define the temporal market as well since competitive conditions may change depending on season, weather, time of the year, etc. It should be noted that the narrower that the market is defined, the more likely that the undertaking under investigation will be found to be dominant.

Establishing Dominance

Once the relevant market is defined, then it can be established whether a given undertaking is dominant. The CJEU has defined dominant position as "a position of economic strength enjoyed by an undertaking which enables it to prevent effective competition being maintained on the relevant market by affording it the power to behave to an appreciable extent independently of its competitors, customers and ultimately of its consumers" (*United Brands* [65]). Such a position does not exclude some competition – which it does where there is a monopoly or quasi-monopoly – but enables the undertaking which profits by the position, if not to determine, at least to have an appreciable influence on the conditions under which that competition will develop and in any case to act largely in disregard of it so long as such conduct does not operate to its detriment (Case 85/76 *Hoffmann-La Roche & Co AG v EC Commission* [1979] ECR 461 [39]). Although the definitions from case law appear to relate to only undertakings on the supply side, clearly, an undertaking on the buying side can also be a dominant undertaking for the purposes of competition law, as was the case in, for example, *British Airways* (see Case C-95/09 P *British Airways v EC Commission* [2007] ECR I-2331).

According to the European Commission, dominance entails that competitive constraints are ineffective, and hence, the undertaking in question enjoys substantial market power over a period of time (Guidance on the Commission's enforcement priorities in applying Article 82 of the EC Treaty to abusive exclusionary conduct by dominant undertakings [2009] OJ C45/7 [20]). An undertaking which is capable of profitably increasing prices above the competitive level for a significant period of time does not face sufficiently effective competitive constraints and can thus generally be regarded as dominant (Guidance [11]). The assessment of dominance will take into account the competitive structure of the market and in particular factors, such as (i) constraints imposed by existing supplies from, and the position on the market of, actual competitors, (ii) constraints imposed by the credible threat of future expansion by actual competitors or entry by potential competitors, and (iii) constraints imposed by the bargaining strength of the undertaking's customers (countervailing buyer power) (Guidance [12–18]). Regarding the first of these factors, market shares provide a useful first indication of the market structure and of the relative importance of various undertakings on the market (Guidance [13]). However, market shares will be interpreted in light of the relevant market conditions and, in particular, of the dynamics of the market and of the extent to which products are differentiated (Guidance [13]). According to the CJEU, “although the importance of the market share vary from one market to another the view may legitimately be taken that very large shares are in themselves, and save in exceptional circumstances, evidence of the existence of a dominant position” (Case 85/76 *Hoffmann-La Roche & Co v EC Commission* [1979] ECR 461 [41]). Interestingly, in *AKZO* the CJEU held that a market share of 50% could be considered to be very large, and in the absence of exceptional circumstances, an undertaking with such a market share would indeed be presumed to be dominant (Case C-62/85 *AKZO Chemie BV v Commission* [1991] ECR I-3359 [60]). Consequently, an undertaking with 50% of market share would have to rebut this presumption to prove that it is *not* dominant.

Interestingly, in its Guidance, instead of referring to a presumption of dominance, the Commission refers to the fact that low market shares are generally a good proxy for the absence of substantial market power: dominance is not likely if the undertaking's market share is below 40% in the relevant market (Guidance [14]). It should be noted that the Guidance is not intended to constitute a statement of the law and is without prejudice to the interpretation of Article 102 TFEU by the CJEU or the General Court; the Guidance sets the enforcement priorities of the European Commission in applying Article 102 TFEU to certain types of abusive practices (see Akman 2010 on the Guidance and its legal position as a soft law instrument). It is noteworthy that the lowest market share that an undertaking has been held to be dominant by the Commission and confirmed on appeal by the General Court is 39.7% in *British Airways*. In any case, the Commission also acknowledges in the Guidance that there may be specific cases below the threshold of 40% market share where competitors are not in a position to constrain effectively the conduct of a dominant undertaking, for example, where they face serious capacity limitations, and such cases may also deserve the attention of the Commission (Guidance [14]).

Other than market shares, potential competition is also important for establishing dominance. Assessing potential competition entails identifying the potential threat of expansion by actual competitors, as well as potential entry by other undertakings into the relevant market. These are relevant factors since competition is a dynamic process and an assessment of the competitive constraints on an undertaking cannot be based merely on the existing market situation (Guidance [16]). An undertaking can be deterred from increasing prices if expansion or entry is likely, timely, and sufficient (Guidance [16]). According to the Commission, in this context, “likely” refers to expansion or entry being sufficiently profitable for the competitor or entrant, taking into account factors such as barriers to expansion or entry, the likely reactions of the allegedly dominant undertaking and other competitors, and the risks and costs of failure

(Guidance [16]). For expansion or entry to be considered “timely,” it must be sufficiently swift to deter or defeat the exercise of substantial market power (Guidance [16]). Finally, to be “sufficient,” expansion or entry has to be more than small-scale entry and be of such a magnitude to be able to deter any attempt to increase prices by the allegedly dominant undertaking (Guidance [16]).

The issue of potential competition brings to the fore the discussion of “barriers to entry or expansion.” There is considerable debate over what should be included within the term “barriers to entry” (Rodger and MacCulloch 2015, 101): one school of thought (Chicago) would only accept as a barrier to entry a cost to new entrants which was not applicable to the existing operators when they entered the market, whereas another school of thought (including the European Commission) views barriers to entry to be much wider, including any factor that would tend to discourage new entrants from entering the market. According to the Commission, barriers to expansion or entry can take various forms, such as legal barriers, tariffs or quotas, and advantages enjoyed by the allegedly dominant undertaking (such as economies of scale and scope, privileged access to essential input or natural resources, important technologies, or an established distribution and sales network) (Guidance [17]). Barriers to expansion or entry can also include costs and other impediments faced by customers in switching to a different supplier (Guidance [17]). Furthermore, the allegedly dominant undertaking’s own conduct may also create barriers to entry, for example, where it has made significant investments which entrants or competitors would have to match or where it has concluded long-term contracts with its customers that have appreciable foreclosing effects (Guidance [17]). Considering the conduct of the undertaking to be an entry barrier is clearly controversial since conduct is normally taken into account when assessing the “abuse” element of the provision rather than the element of dominance. Such an approach is circular in that conduct will not normally be considered abusive until dominance is established, but if conduct can also indicate dominance, then the likelihood of a

finding of abuse clearly increases (Rodger and MacCulloch 2015, 102).

The final factor to be taken into account in establishing dominance is countervailing power held by the trading partners of the allegedly dominant undertaking. Competitive constraints may be exerted not only by actual or potential competitors but also by customers (or suppliers, if the dominant undertaking is on the buying side) of the allegedly dominant undertaking. According to the Commission, even an undertaking with a high market share may not be able to act to an appreciable extent independently of customers (or suppliers, if the dominant undertaking is on the buying side) with sufficient bargaining strength (Guidance [18]). Such bargaining power may result from the trading partner’s size, commercial significance, ability to switch, ability to vertically integrate, etc. (Guidance [18]).

Dominant Position “Within the Internal Market or in a Substantial Part of It”

According to Article 102 TFEU, to be subject to the prohibition therein, the relevant undertaking has to have a dominant position “within the internal market or in a substantial part” of the internal market. This has been noted to be the equivalent of the *de minimis* doctrine under Article 101 TFEU, according to which agreements of minor importance are not caught by the prohibition found in Article 101 TFEU (Whish and Bailey 2012, 189). Clearly, where dominance is established to exist throughout the EU, there is no difficulty in deciding that the dominant position is held within the internal market or in a substantial part of the internal market. Where dominance is more localized than this, then it will have to be decided what “substantial” refers to. Substantiality does not simply refer to the physical size of the geographic market within the EU (Whish and Bailey 2012, 190). Rather, what matters is the relevance of the market in terms of volume and economic opportunities of sellers and buyers (Cases 40/73 etc *Suiker Unie and others v EC Commission* [1975] ECR 1663 [371]). Each Member State is likely to be considered to be a substantial part of the internal market, as well as parts of a Member State (Whish and Bailey 2012, 190).

Effect on Trade Between Member States

The prohibition in Article 102 TFEU is only applicable to the extent that the conduct of the dominant undertaking “may affect trade between Member States.” This is a *jurisdictional* criterion that establishes whether EU competition law is applicable, as well as demarcating the application of EU competition law from national competition laws of the Member States. The Commission has published Guidelines on the effect on trade concept contained in Articles 101 and 102 TFEU ([2004] OJ C 101/81). First and foremost, Articles 101 and 102 TFEU are only applicable where the effect on trade between Member States is *appreciable* (Guidelines [13]). Second, the concept of “trade” is not limited to exchange of goods/services but covers all cross-border economic activity, including the establishment of agencies, branches, subsidiaries, etc. in Member States (Guidelines [19, 30]). The concept of “trade” also includes cases where conduct results in a change in the *structure* of competition on the internal market (Cases 6 and 7/73 *Commercial Solvents v EC Commission* [1974] ECR 223 [33]; Guidelines [20]).

Regarding the notion of “may affect” trade between Member States, the CJEU has noted that this means that it must be possible to foresee, with a sufficient degree of probability on the basis of a set of objective factors of law or of fact, that the conduct may have an influence, direct or indirect, actual or potential, on pattern of trade between Member States (Guidelines [23]). It should be noted that this is a neutral test; it is not a condition that trade be restricted or reduced (O’Donoghue and Padilla 2013, 864).

Abuse

According to the CJEU, a finding of a dominant position is not in itself a recrimination, but means that irrespective of the reasons for which it has such a position, the dominant undertaking has a “special responsibility” not to allow its conduct to impair genuine undistorted competition on the internal market (Case 322/81 *Michelin v EC*

Commission [1983] ECR 3461 [57]). Although the Court has created this concept of “special responsibility” for dominant undertakings, what exactly it entails – and if it entails anything above the parameters of the prohibition in Article 102 TFEU itself – is debatable. It has been suggested in the literature that dominant undertakings do not have any responsibility over and above complying with Article 102 TFEU itself (Akman 2012, 95).

Article 102 TFEU prohibits the abuse of a dominant position and lists examples of abuse in a non-exhaustive manner (*Continental Can* [26]). In general, it is considered that abusive conduct can be categorized as: (i) exploitative and (ii) exclusionary. Exploitative abuse relates to the dominant undertaking directly harming its customers (including consumers) as a result of, for example, the reduction in output and the increase in prices that the dominant undertaking can effect due to its market power. It has been defined as the dominant undertaking receiving advantages to the disadvantage of its customers that would not be possible *but for* its dominance (Akman 2012, 95, 303). In contrast, exclusionary abuse concerns the dominant undertaking’s conduct that harms the competitive position of its competitors, mainly by foreclosing the market. Most of the decisional practice under Article 102 TFEU has concerned exclusionary conduct, despite the fact that the examples listed in Article 102 TFEU are mostly – if not only – concerned with exploitative abuse. Indeed, it has been argued in the literature that Article 102 TFEU itself is merely concerned with exploitation and not exclusion (see Joliet 1970; Akman 2009, 2012). In line with the decisional practice, the Commission’s Guidance on enforcement priorities – the only official document on the application of Article 102 TFEU adopted by the Commission – is limited to exclusionary conduct. It must be noted that the Guidance was published at the end of a long period of debate on the role and application of Article 102 TFEU as the Commission’s enforcement of this provision, as well as the European Courts’ jurisprudence thereon had been criticized by many commentators for not being based on economic effects but on the form of conduct, for

failing to fall in line with the Commission's more modern approach to Article 101 TFEU and merger control, and "for protecting competitors, not competition" (see, e.g., O'Donoghue and Padilla 2013, 67 et seq. for an overview of the reform).

Exploitative Abuses

Unfair Pricing and Unfair Trading Conditions

Article 102(a) TFEU prohibits the imposition of unfair prices or unfair trading conditions. Although the prohibition is one of "unfair" pricing, it has been mostly interpreted as one of excessive pricing. In any case, there have been very few cases prohibiting prices as "excessive" or "unfair" since the prohibition poses many problems, such as the difficulty of defining what an "excessive" or "unfair" price is, the potential adverse effects on innovation and investment the prohibition could lead to, the lack of legal certainty resulting from a lack of a test for "excessive" or "unfair" prices, the inappropriateness of price regulation by competition authorities and courts, etc. As for the interpretation of "unfair pricing" by the CJEU, there is a two-staged test established in *United Brands*: first, it should be determined whether the price-cost margin is excessive, and if so, it should be determined whether a price has been imposed that is either "unfair in itself" or "when compared to competing products."

Regarding the abuse of imposing unfair trading conditions, there is similarly limited case law. Examples of unfair trading conditions include imposing obligations on trading partners which are not absolutely necessary and which encroach on the partners' freedom to exercise its rights, imposing commercial terms that fail to comply with the principle of proportionality, unilateral fixing of contractual terms by the dominant undertaking, etc. (Case 127/73 *Belgische Radio on Televisie v SV SABAM* [1974] ECR 313; *DSD* (Case COMP D3/34493) [2001] OJ L166/1; Case 247/86 *Alsatel v SA Novasam* [1988] ECR 5987 respectively).

Exclusionary Abuses

According to the Guidance paper, the Commission's enforcement activity in relation to exclusionary conduct aims to ensure that dominant

undertakings do not impair effective competition by foreclosing their competitors in an anticompetitive way, thus having an adverse impact on consumer welfare, whether in the form of higher prices than would have otherwise prevailed or in some other form such as limiting quality or reducing consumer choice (Guidance [19]). In turn, "anticompetitive foreclosure" refers to a situation where effective access of actual or potential competitors to supplies or markets is hampered or eliminated as a result of the conduct of the dominant undertaking whereby the dominant undertaking is likely to be in a position to profitably increase prices (or to influence other parameters of competition, such as output, innovation, quality, variety, etc.) to the detriment of consumers (Guidance [19]). It must be noted that the Guidance has not received formal approval from the EU Courts as they have not yet had to deal with a case in which the Commission applied the principles of the Guidance and the EU Courts have not necessarily been keen to revise their case law in order to adopt a more economic effects-based approach. The rest of this section will consider some common types of exclusionary conduct that have been identified as priorities in the Commission's Guidance paper. It should be noted that the Guidance paper distinguishes between price-based and non-price-based exclusionary conduct. For price-based exclusionary conduct, the test promoted in the Guidance is that of the "as efficient competitor" test, according to which, the Commission will only intervene where the conduct concerned has already been or is capable of hampering competition from competitors which are considered to be as efficient as the dominant undertaking (Guidance [23]). This test was used in some earlier case law already, but it is noteworthy that in the recent appeal of *Intel*, the General Court has found the test to be a neither necessary nor sufficient test of abuse, at least for the particular conduct in question, namely, rebates (Case T-286/09 *Intel Corp v European Commission* (not yet published) [143–146]).

Exclusive Dealing

A dominant undertaking may foreclose its competitors by hindering them from selling to

customers through use of exclusive purchasing obligations or rebates, which the Commission together refers to as “exclusive dealing” (Guidance [32] et seq.). “Exclusive purchasing” refers to an obligation imposed by the dominant undertaking on a customer to purchase exclusively or to a large extent only from the dominant undertaking (Guidance [33]). According to the CJEU, it is irrelevant whether the request to deal exclusively comes from the customer, and it is also irrelevant whether the exclusivity obligation is stipulated without further qualification or undertaken in return for a rebate (*Hoffmann-La Roche* [89]).

Regarding rebates granted to customers to reward them for a particular form of purchasing behavior, particularly the European Courts have adopted a very formalistic approach. Specifically, rebates that create “loyalty” to the dominant undertaking are condemned to a degree which some might argue to be a per se prohibition. For example, recently in *Intel* the General Court held that fidelity/exclusivity rebates are abusive if there is no justification for granting them, and the Commission does not have to analyze the circumstances of the case to establish a potential foreclosure effect (*Intel* [80–81]). Quantity rebates, namely, rebates which are linked solely to the volume of purchases from the dominant undertaking which reflect the cost savings of supplying at higher levels, are generally deemed not to have foreclosure effects (*Intel* [75]).

Tying and Bundling

Tying refers to situations where customers that purchase one product (the “tying product”) are required to also purchase another product from the dominant undertaking (the “tied product”) and can be contractual or technical (Guidance [48]). Bundling refers to the ways the dominant undertaking offers and prices its products: pure bundling occurs where products are *only* sold together in fixed proportions, whereas mixed bundling occurs where products are available separately, but the price of the bundle is lower than the total price when they are sold separately (Guidance [48]). As well as having potential efficiency benefits for customers, tying and bundling

can also be used by the dominant undertaking to foreclose the market for the tied product by leveraging its market power in the tying product market to the tied product market. For example, in the case of *Microsoft*, the tying of Microsoft Media Player to the Windows Operating System was found to be an abuse of Microsoft’s dominant position (Case T-201/04 *Microsoft Corp v EC Commission* [2007] ECR II-3601).

Predatory Pricing

Predation involves the dominant undertaking selling its products at a price below cost. As such, the dominant undertaking deliberately incurs losses or foregoes profits in the short term, so as to foreclose or be likely to foreclose one or more of its actual or potential competitors, with a view to strengthening or maintaining its market power (Guidance [63]). The obvious difficulty with sanctioning predation is that an erroneous condemnation would imply the prohibition of low prices, and price competition leading to low prices is clearly part of legitimate competition. In *AKZO* the CJEU decided that prices below average variable costs (costs that vary according to the quantity produced) by means of which a dominant undertaking seeks to eliminate a competitor must be regarded as abusive since a dominant undertaking has no interest in applying such prices except that of eliminating competitors so as to enable it subsequently to raise its prices by taking advantage of its monopolistic position, since each sale generates a loss (*AKZO* [71]). As for prices between average variable costs and average total costs (variable costs plus fixed costs), the Court held that such prices will be held abusive if they are part of a plan for eliminating a competitor (*AKZO* [72]). In its Guidance, the Commission uses average avoidable cost instead of average variable cost ([64] et seq.). It is not necessary to prove that the dominant undertaking can possibly recoup its losses for predation to be abusive (see, e.g., Case C-202/07 P *France Télékom v Commission* [2009] ECR I-2369 [110]).

Refusal to Supply and Margin Squeeze

Despite the fact that generally any undertaking, dominant or not, should have the right to choose

its trading partners and to dispose freely of its property, there are occasions on which a dominant undertaking can abuse its position by refusing to deal with a certain trading partner (Guidance [75]). Such a finding entails the imposition of an obligation to supply on the dominant undertaking, and the Commission acknowledges that such obligations may undermine undertakings' incentives to invest and innovate and, thereby, possibly harm consumers (Guidance [75]). According to the Commission, typically competition problems arise when the dominant undertaking competes on the downstream market with the buyer whom it refuses to supply (Guidance [76]). The downstream market refers to the market for which the refused input is needed in order to manufacture a product or provide a service (Guidance [76]). It is irrelevant whether the customer to whom supply is refused is an existing customer or a new customer, but it is more likely that the termination of an existing relationship will be found to be abusive than a *de novo* refusal to supply (Guidance [79], [84]). Instead of refusing to supply, a dominant undertaking may engage in "margin squeeze": it may charge a price for the product on the upstream market which, compared to the price it charges on the downstream market, does not allow an equally efficient competitor to trade profitably in the downstream market on a lasting basis (Guidance [80]).

For refusal to supply and margin squeeze to constitute abuse, it must be the case that the refusal relates to a product or service that is objectively necessary ("indispensable") to be able to compete effectively on the downstream market, is likely to lead to elimination of effective competition on the downstream market, and is not objectively justified (Whish and Bailey 2012, 699). A particularly controversial application of this doctrine is the area of intellectual rights where a finding of abuse implies that these rights may be made subject to compulsory licensing. For example, in *Microsoft*, abuse was found in Microsoft's refusal to provide interoperability information to its competitors which would enable them to develop and distribute products that would compete with Microsoft on the market for servers. According to the General Court, Microsoft's

conduct limited technical development under Article 102(b) TFEU (*Microsoft* [647]).

Objective Justification

Article 102 TFEU does not contain an exemption or exception clause like that found in Article 101(3) TFEU, which would "save" otherwise abusive conduct from breaching Article 102 TFEU due to any procompetitive gains the conduct might produce. However, the decisional practice and the case law have developed the concept of "objective justification" as a defense on the part of the dominant undertaking. The dominant undertaking may demonstrate that its conduct is objectively necessary or that the anti-competitive effect produced by the conduct is counterbalanced or outweighed by advantages in terms of efficiencies that also benefit consumers (Case C-209/10 *Post Danmark A/S v Konkurrenceradet*, not yet reported [41]; Guidance [28–31]). According to some commentators, the defense of objective justification is somewhat a tautology (O'Donoghue and Padilla 2013, 283). This is because, as noted by Advocate General Jacobs in *Syfait*, the very fact that conduct is characterized as abusive suggests that a negative conclusion has already been reached, and therefore, a more accurate conception would be to accept that certain types of conduct do not fall within the category of abuse at all (Case C-53/03 *Syfait and others v GlaxoSmithKline plc and another* [2005] ECR I-4609 [53]).

References

- Akman P (2009) Searching for the long-lost soul of Article 82EC. *Oxf J Leg Stud* 29(2):267–303
- Akman P (2010) The European commission's guidance on Article 102TFEU: from *Inferno* to *Paradiso*? *Mod Law Rev* 73(4):605–630
- Akman P (2012) The concept of abuse in EU competition law: law and economic approaches. Hart Publishing, Oxford
- Joliet R (1970) Monopolization and abuse of dominant position. Martinus Nijhoff, La Haye
- O'Donoghue R, Padilla J (2013) The law and economics of Article 102 TFEU, 2nd edn. Hart Publishing, Oxford
- Rodger BJ, MacCulloch A (2015) Competition law and policy in the EU and UK, 5th edn. Routledge, Oxford
- Whish R, Bailey D (2012) Competition law, 7th edn. OUP, Oxford

Access to Justice

David Allen Larson

Hamline University School of Law, Saint Paul,
MN, USA

Abstract

Access to justice has both procedural and substantive components. Context matters and a society that views its members as part of a collective, for example, may perceive access to justice differently than a more individualistic society. International law documents ensuring access to justice generally take either a general human rights approach or provide specific protections for disadvantaged populations. Although substantive access to justice appears to have improved over time, procedural access to justice may not have kept pace. Money and time are very real limitations. Physical barriers have a severe impact on persons with disabilities and individuals living in poverty. Institutional barriers also limit access to justice for reasons that include ponderous or bias court systems, discriminatory police conduct, expense, and political interference. Additionally, limited education and social custom impair access to justice. When public trust is lacking, individuals may not rely on justice institutions to settle disputes and resolve problems. Challenges remain concerning which substantive rights we need to protect and what efficient and effective procedures are available.

Definition

Access to justice, frequently abbreviated ATJ or A2J, refers to two different, but closely related, concerns. On the one hand, procedural access to justice focuses on both the processes that are available to help people enforce their rights and privileges under the law and the effectiveness of those processes. When one explores procedural access to justice, for example, one might examine

access to the physical locations of justice administration (such as courthouses and police stations), individuals' ability to understand and participate in proceedings (such as court hearings, police interviews, and conversations with one's attorney), and due process of law. On the other hand, the notion of substantive access to justice focuses on the nature and extent of the rights to be protected. Questions about substantive access to justice can focus on specific rights or address broad questions such as whether a society ensures equitably equal access to opportunities and benefits and whether all individuals have the ability to live a just life. It is difficult to imagine a just life that does not require the existence and enforceability of substantive rights, such as the right to travel freely or the right to freedom of conscience. But any discussion of access to justice must acknowledge that context matters. A society or culture that views its members as part of a collective may have a different understanding of substantive access to justice than a society that is more individualistic.

Access to Justice

Procedural access to justice is central to the enforcement of legal rights and privileges (Ortoleva 2011). When individuals and groups are excluded from society, talents and value are lost. Improving procedural access to justice may be particularly helpful for groups that traditionally and historically have been subject to discrimination (Ortoleva 2011). When a law or practice prevents or discourages individuals from participating in the mechanisms for enforcing laws (which include policing, civil and criminal court claims, alternative dispute resolution processes, appeals, and final judgments), it denies those individuals procedural access to justice (Ortoleva 2011).

The phrase access to justice also may be used when one focuses on the question of whether individuals truly are experiencing a just existence. Substantive access to justice is generally the goal, if not always the result, of procedural access to justice. Procedural access to justice provides a route to substantive access to justice. But even

the most carefully designed procedures may provide little benefit in the absence of recognized or cognizable substantive rights.

Many international law documents, as well as the constitutions and laws of most nations, ensure the right of people to access the courts. International law documents ensuring access to justice can be divided into two types: (1) those that take a general human rights approach and (2) those that make specific provisions for populations that historically have lacked access to justice for one reason or another. As promising as they sound, however, it may be difficult to enforce the rights described in these documents.

Documents belonging to the first category include the United Nations (UN) Universal Declaration of Human Rights and the Hague Convention on International Access to Justice, along with various regional declarations and conventions, such as the American Declaration of the Rights and Duties of Man and the European Convention for the Protection of Human Rights and Fundamental Freedoms. These documents are primarily aspirational, though some (such as the Hague Conventions and the International Covenant on Civil and Political Rights) potentially are binding on states. Aspirational documents tend to contain more substantive provisions than procedural ones, but they do address both types of access to justice.

Many general human rights documents express the same rights and freedoms as appear in the Universal Declaration of Human Rights. Rights believed important for access to justice include the right to recognition as a person before the law; the right to equal protection of the laws; the right to an effective, enforceable remedy for a violation of one's rights; and the right to a full, fair, and public hearing by an independent and neutral tribunal. Although the line is not always bright, these primarily procedural rights are central to the enforcement of substantive rights, such as the rights of freedom of thought, conscience, and religion; freedom to travel; and freedom from arbitrary arrest, detention, and exile.

The second category, consisting of population-focused documents, includes many UN conventions, such as the Convention to Eliminate All Forms of Discrimination Against Women, the

UN Convention for the Elimination of Racial Discrimination, and the UN Convention on the Rights of Persons with Disabilities. These documents reaffirm the principles of general human rights documents but also are sensitive to the special needs and circumstances of the populations they protect. For example, the Convention on the Rights of Persons with Disabilities contains many of the same rights as general human rights documents. But it also provides, for example, that states should ensure that persons with disabilities have equal access to information and that those working in fields of justice administration receive appropriate training to advance and protect the rights of persons with disabilities.

International bodies within the United Nations, such as the Human Rights Committee, also bear some of the responsibility for monitoring the impact of the international community's efforts to improve access to justice. The United Nations Development Programme, the primary UN body responsible for the UN's Millennium Development Goals, provides support to legal aid providers in several countries, with particular emphasis on providing services and education to the poor and other marginalized groups.

Many national constitutions secure rights that are necessary to access to justice. Such rights include the right to counsel, the right to equal treatment by the government, and the right to seek redress for wrongs through the courts. Although these rights may mirror those contained in general human rights documents, they can be more easily enforced because they are the laws of the state rather than aspirations of the international community.

In addition to including provisions in their constitutions, nations also promote access to justice through legislation. Countries may enact legislation to fulfill obligations under international conventions but also may be responding to the need in their own country for heightened protection for identifiable populations. For example, one can find legislation in many countries that promises protection for individuals based on their race, religion, gender, health, or country of origin. Anti-discrimination legislation at the national level, public educational initiatives, and programs

directed toward empowering and informing marginalized populations of their rights can improve access to justice by eliminating gaps in political, economic, and social power. Additionally, some nations have chosen not only to prohibit discrimination against marginalized populations but also encourage policy makers to continue to review court decisions and other governmental actions that may have a disproportionate effect on those identifiable populations. These efforts are most effective when they are a result of community participation and consensus.

When a person or group is denied access to justice, that denial may be either direct or indirect. Access to justice is denied in a direct manner when a specific person or group is explicitly prevented from attaining procedural or substantive justice. Historical examples include restrictions on personhood based on race and the shedding of a woman's legal personality upon marriage (coverture). Indirect denial of access to justice, in contrast, consists of restrictions that appear neutral but have a disparate impact on a specific group. This type of discrimination includes filing fees that may be affordable for most people but prevent the poor from accessing the courts; height, weight, or strength requirements that exclude qualified women from a job even though those qualities are not necessary to perform the actual duties of the job; or diploma, certificate, or skill requirements that again may not actually be needed to perform a well-paying job. Employment can improve access to justice by providing the resources (time, money, knowledge) necessary for a person to access the justice system.

Access to Justice in Historical Context

Before the modern era of human rights, the common view was that only states could invoke the protections of international law (Francioni 2007). Individuals had limited access to justice. They could seek redress only in the state where the wrong occurred, which meant that their rights and privileges were defined by that state's own laws (Francioni 2007). When individuals sought redress in foreign states that may have had more

favorable laws, they often faced local prejudices, language barriers, wide variances in substantive law, and other challenges that hindered access to justice (Francioni 2007).

In order for a person to enforce his or her rights in the international realm, a state would need to intercede on the person's behalf (Francioni 2007). Claims made by a state on behalf of one of its citizens were not considered individual claims for remedy, but rather diplomatic issues to be resolved between the states themselves (Francioni 2007). These claims often were addressed in ways that did not involve judicial mechanisms and which were not transparent (Francioni 2007). Even today, some forums are only available to states, such as the International Court of Justice.

Eventually, multiple international forums for enforcing individual rights came into existence (Francioni 2007). Regional and nongovernmental organizations have created forums for the vindication of individual rights, such as the European Court of Justice and the Inter-American Court of Human Rights. In addition to these international bodies, individuals today retain their historical access to domestic courts (Francioni 2007). Agreements between states which ensure that individuals have access to justice often include the individual's right to access the courts of a state regardless of nationality. While procedural and substantive barriers may still arise, discrimination based solely on nationality is generally forbidden, along with discrimination based on characteristics such as race, gender, and religion.

Access to justice has improved in some respects while lagging in others. In most societies, substantive access to justice appears to have been strengthened. When one examines modern constitutional and legislative language, it is not uncommon to find language guaranteeing access to justice for all persons. Bold and impressive language often promises access to justice for all persons regardless of factors such as age, race, and sex. Populations with other shared characteristics, such as persons with disabilities and those with HIV/AIDS, also have better access to justice as countries work to address histories of stigmatization. But even though substantive access to

justice may have improved, procedural access to justice may not have kept pace. Court systems and administrative agencies often are terribly backlogged, which results in significant delays and in some cases a denial of access to justice. Money and time can be very real limitations affecting access to justice in modern societies. The availability of alternative forums for justice, such as restorative justice processes, mediation, and arbitration, may increase access to justice, but the potential for unfair outcomes still looms when these processes are not held to the same standards of fairness, equality, and transparency as court systems.

Although research generally supports the commonly held assumption that early referral to mediation increases the chances that parties will reduce their expenses, it does not support the claim that mediation by itself reduces party costs (McEwen 2014). Analyses of civil mediation programs do not find any consistent differences in attorneys' fees, hours, or other costs between mediated and other cases (McEwen 2014). Litigation costs and litigation activity (depositions, interrogatories, and motions) do not automatically decline whenever parties choose mediation (McEwen 2014). Research focusing on business-to-business disputes does suggest, however, that if parties engage in mediation early in the litigation process, then it can significantly reduce costs (McEwen 2014). Cost savings appear most likely when mediation (and case management) alters normal litigation practices, particularly by reducing the amount of discovery and motions and by initiating serious settlement efforts early in the case (McEwen 2014).

Physical Access to Justice

One of the most essential aspects of access to justice is the ability to reach the locations where justice is administered. Physical barriers such as distance can have a severe impact on persons with disabilities and the poor (Ortoleva 2011; Carmona 2012). Distance makes it more difficult for victims to report crimes, for police to respond to reports of crime, for those seeking justice to secure legal representation, and for persons who are disabled

or poor to do something as seemingly simple as physically appear at court proceedings (Ortoleva 2011; Carmona 2012). For the poor, securing transportation to the locations where justice is carried out can be prohibitively expensive (Carmona 2012). Even when transportation is affordable, lengthy travel to physical locations often means losing valuable time at work or at home (Carmona 2012). The relative wealth or poverty of a country obviously contributes to these difficulties, particularly in those parts of the world where transportation and communications infrastructures may be underdeveloped.

Certain disabilities can make physical access to justice especially difficult. In areas without comprehensive legal requirements for building accessibility, people who cannot walk are either unable to enter courthouses and police stations or find it very difficult or humiliating (Ortoleva 2011). Those who are blind or deaf may be physically present, but often do not enjoy the full benefit of that presence without interpreters or Braille materials (Ortoleva 2011).

Institutional Access to Justice

The enforcement of the rights that underlie access to justice requires institutions that work for everyone. Such institutions must provide timely, fair, and effective resolution of questions about people's rights. Institutional barriers that hinder access to justice include ponderous or biased court systems, the expense of using legal institutions, and political interference with judicial processes (Agrast 2014).

While the justice systems in the world's most developed nations promise fairness and accessibility, they do not always fulfill those promises. Despite guarantees of fairness, usually included in these countries' laws and constitutions, marginalized groups still face problems when trying to access justice that can include physical access challenges, discriminatory police conduct, or judicial bias (Agrast 2014).

The right to legal counsel, generally regarded as a valuable guarantee, in fact creates some of the tension that exists when we think about access to

justice, especially with respect to the poor. This right often applies only for those accused of a crime, for example, not to those involved in a civil damages dispute (Agrast 2014). But a civil damages lawsuit potentially may be financially devastating, and excellent legal representation may be required to avoid that devastation. That level of civil legal representation, however, can be prohibitively expensive (Bloch 2008). And even though one may be guaranteed legal counsel in criminal cases, the quality of that legal counsel may again depend upon one's financial resources (Ogletree and Sapir 2004).

The right to counsel in criminal proceedings is certainly an important consideration in defense of one's rights, but the primary mechanisms for enforcing rights that impact access to justice often are civil proceedings. Therefore, enforcing rights attendant to access to justice often is a pay-to-play proposition, and the poor may find it difficult or impossible to challenge policies and laws that restrain their access to justice when they do not have legal counsel (Ortoleva 2011). This problem is compounded in nations where the government is either unable or unwilling to provide adequate funding to support the legal needs of the poor (Agrast 2014). These problems have been partially addressed by the European Convention on Human Rights, which has been interpreted to provide a limited right to counsel in civil cases, but enforcement of that right remains problematic. Alternative routes to remedies, such as legislation, are also costly, often prohibitively so for individuals.

Class action mechanisms alleviate some of the economic strain of civil litigation, but class actions are not universally available (Watson 2001). Furthermore, the procedural barriers to initiating a class action lawsuit require even more expenses at the beginning of the lawsuit, which may prove insurmountable even if the cost is shared by the class (Watson 2001). Nonetheless, class actions can be a powerful tool once the procedural requirements are overcome. Aggregating the claims of many individuals into a single lawsuit can make viable claims for amounts that otherwise would have been less than the cost of obtaining them individually (Watson 2001).

Class actions can promote judicial economy by consolidating what otherwise would be a multitude of separate lawsuits heard by many judges into a single lawsuit (Watson 2001). In both class actions and smaller-scale litigation, however, individuals living in poverty may have their remedies limited by fee shifting provisions. So-called "loser pays" systems of fee shifting, where the party who loses the lawsuit must pay the winning party's fees, increase the potential cost of litigation (Watson 2001; Hodges 2001). While this has the commonly desired effect of discouraging frivolous or weak claims, it may also discourage strong claims by those who simply cannot afford to pay the other party's legal fees (Hodges 2001). Some jurisdictions, recognizing that simple consumer claims need not be subject to the entire trial process, have attempted to simplify the process for these small claims by eliminating the need for legal representation and removing some procedural requirements from the claims process itself. Both class actions and simplified processes may reduce the expense of pursuing a claim, making redress more accessible to those who may otherwise lack the financial means to vindicate their rights.

But cost is not the only reason why class actions may not be very helpful. There are significant procedural hurdles. In order to proceed with a class action, for example, the complaining parties must share a common claim. Change a few facts, change the alleged wrongful actor, or change the location and it may be impossible to proceed as a class. This requirement can prove especially problematic for persons claiming disability discrimination because, given the seemingly infinite range of possible disabilities, claimants may not be able to demonstrate that they are similarly situated.

At a more basic level, it is obvious that in order to enforce substantive rights, people must first be aware of those rights. This is of special concern when it comes to persons with intellectual, visual, or auditory disabilities (Ortoleva 2011). Justice institutions such as courts and police stations often lack interpreters for the deaf, materials available in a format accessible to the blind, and guardianship for those with intellectual disabilities (Ortoleva 2011). Groups who have limited

education (often the poor and women) and who are not allowed by social custom or law to access justice institutions for themselves (women and children) share similar challenges (Turquet et al. 2015). Even when there are no court proceedings in the foreseeable future, the right to counsel clearly is valuable. Counsel can inform and advise individuals regarding their rights (Ortoleva 2011). Even something as simple as programs that provide informative materials for those with certain disabilities, or that encourage public education about rights and the law, can help bridge the knowledge gap.

The public also must be able to trust that justice institutions will serve their purpose. Minority groups often believe that these institutions serve the majority at the expense of the minority or that they even act in direct opposition to the minority. The public may not only believe that procedural access to justice is compromised, they may also believe that substantive access to justice is lacking. Both claimants and respondents may have this belief. For example, after the passage of the Americans with Disabilities Act (ADA) in the United States, news coverage and public discourse revealed a concern that courts responsible for handling ADA cases awarded excessive amounts to plaintiffs at the expense of their employers (Krieger 2003). The fact is that a significant number of reasonable accommodations for persons with disabilities cost very little or nothing at all, however, and the average cost of reasonable accommodation is only in the hundreds of dollars (Krieger 2003). Public perception does not always reflect what actually occurs in justice institutions, but a lack of public trust may prevent people from relying on justice institutions to settle disputes and resolve problems. Thus even when institutional processes are available, individuals may decide that it is not worth their time or money to pursue those processes.

Although generally available, the institutions central to access to justice may be particularly vulnerable during periods of crisis, such as war, natural disaster, or civil unrest (Francioni 2007). Financial resources may be diverted to address the crisis and the institutions themselves may be so damaged or compromised that they are ineffective.

Legal Rights and Access to Justice

We still may be at a point in time where access to justice requires the existence of enforceable legal rights that protect both procedural and substantive access to justice. As explained earlier, procedural rights govern the ability to access and use justice institutions. Procedural rights include formal rules that explicitly protect one's ability to testify before a court (competency), for example, or rules that protect the ability to bring a claim or respond to a criminal charge or rules of evidence and formal procedure.

Substantive rights identify the types of rights that we hope to protect, such as rights to life, liberty, or prosperity. These rights can adapt to specific circumstances and populations. They can be broad pronouncements or more specific to fit a particular situation, such as the right to physically access courthouses that have a wide variety of architectural design. Majority groups are often the most protected, while minorities may find that they lack some of the substantive rights and privileges that the majority possesses. Women, for example, may lack the right to own or inherit property, restraints that are not usually experienced by men (Turquet et al. 2015). Some rights, such as freedom of expression, may be reserved to a privileged social or cultural group, or they may be enforced unevenly. In that latter circumstance, the rights would not be protected procedurally, resulting in *de facto* censorship or chilling of speech. Thus the challenge remains twofold: what are the substantive rights that we need to protect and what procedures are available that are effective and efficient? The answers to these two questions will help guide local, national, and international efforts to improve and secure access to justice.

References

- Agrast MD (2014) World Justice Project Rule of Law Index 2014. Available at http://worldjusticeproject.org/sites/default/files/files/wjp_rule_of_law_index_2014_report.pdf
- Bloch FS (2008) New directions in clinical legal education: access to justice and the global clinical movement. *Wash Univ J Law Policy* 28(1):111–139

- Carmona MS (2012) Report of the Special Rapporteur on extreme poverty and human rights. U.N. Doc. No. A/HRC/23/36 (11 Mar 2012). Available at <http://daccess-dds-ny.un.org/doc/UNDOC/GEN/G13/117/94/PDF/G1311794.pdf?OpenElement>
- Francioni F (2007) Access to justice as a human right. Oxford University Press, New York
- Hodges C (2001) Multi-party actions: a European approach. *Duke J Comp Int Law* 11(2):321–354
- Krieger LH (2003) Backlash against the ADA. University of Michigan Press, Ann Arbor
- McEwen C (2014) Determining whether to initiate a mediation program and how its structure affects results – costs for parties, 1 Mediation: Law, Policy and Practice § 14:5, Cole S, McEwen C, Rogers N, Coben J, Thompson P, Thomson Reuters
- Ogletree CJ Jr, Sapir Y (2004) Keeping Gideon’s promise: a comparison of the American and Israeli public defender experiences. *N Y Univ Rev Law Soc Change* 29(2):203–235
- Ortoleva S (2011) Inaccessible justice. *ILSA J Int Comp Law* 17(2):281–320
- Turquet L et al (2015) Progress of the world’s women: in pursuit of justice. U.N. Women. Available at <http://progress.unwomen.org/pdfs/EN-Report-Progress.pdf>
- Watson GD (2001) Class actions: the Canadian experience. *Duke J Comp Int Law* 11(2):269–288

Further Reading

- Friedman LM (2009) Access to justice: some historical comments. *Fordham Urban Law J* 37(1):3–15
- Johnson E (1985) The right to counsel in civil cases: an international perspective. *Loyola Law Rev* 19(2):341–436

Accountability

- [Codes of Conduct](#)

Act-Based Sanctions

Marie Obidzinski
CRED, Université Panthéon-Assas Paris II, Paris, France

Abstract

An act-based sanction punishes actions independently of the actual occurrence of harm. On the opposite, harm-based sanctions are

imposed when harm occurs. When harm is certain, the distinction between act-based and harm-based sanctions is not relevant. On the opposite, when harm is probable, it is worth distinguishing act and harm based sanctions. In a basic public law enforcement framework, optimal deterrence can be achieved by both types of sanctions. This statement does not hold when potential offenders either have limited assets or are risk adverse or can invest in avoidance activities. Again, this assertion is not true when punishment is costly.

Definition

An act-based sanction punishes actions directly. It does not depend on the actual occurrence of harm.

Act-Based Sanctions Versus Harm-Based Sanctions

Act-based sanctions are a tool to control undesirable acts. Undesirable acts are those whose private gains are lower than the expected social harm (Shavell 1993). Act-based sanctions are generally opposed to harm-based sanctions. While the latter are imposed on the basis of the actual occurrence of harm, act-based sanctions directly punish actions, irrespective of the actual occurrence of harm (Polinsky and Shavell 2007). The distinction between harm- and act-based sanctions refers to the timing of legal intervention, before any harm or after. This distinction between act- and harm-based sanctions is relevant when harm is probabilistic, rather than certain. In the case where an action creates harm with certainty, there is no difference between setting the sanction on the basis of action or harm. Examples where harm is probabilistic are numerous in safety regulation, in environmental law (Rousseau and Blondiau 2014), or in traffic law. In crime law, attempts are a typical case where the act did not result in harm.

Shavell (1993) analyzes the choice between prevention, act-based sanctions and harm-based sanctions are in relation with the structure of law

enforcement. Prevention refers to the use of force to forbid the action. This is the case when a firm has no right to sell a good, when a person is put in jail, or when fences are built. No choice is left to the person; law is directly enforced. On the other hand, act-based and harm-based sanctions tend to influence the choice of the actors. It is worth noting that in criminal law, all three methods are used by law enforcers. A policeman can prevent the crime and arrest on the basis of an attempt or on the basis of harm. On the other hand, safety regulation is mostly characterized by act-based sanctions (Shavell 1993). The timing of legal intervention has also been analyzed by the ex post liability versus ex ante regulation literature since Wittman (1977). Act-based sanctions refer to input monitoring and harm-based sanctions to output monitoring.

The aim of the threat of sanctions is to provide the right incentives in order to separate desirable and undesirable acts, as actors differ in the private benefits they receive from acts. The threat of sanctions induces decision-makers to internalize the expected costs of their acts. Law enforcement theory therefore balances the advantages and benefits of act- versus harm-based sanctions according to circumstances.

Theoretically, in a basic law enforcement framework, optimal deterrence can be achieved by both types of sanctions. Assume that the probability of detection and conviction is exogenously set and equal under act- and harm-based sanctions. The monetary fine should be higher under harm-based sanctions in order to reach the same deterrence level as under act-based sanctions, given that the harm is probabilistic. This balance can be tilted in favor of act-based sanctions when individuals or firms have limited assets or are risk adverse (Polinsky and Shavell 2007). The reverse is true when punishment is costly. Harm-based sanctions can also incite people to engage in harm avoidance activities, where possible.

Another significant determinant of the choice between act- and harm-based sanctions lies in the information possessed either by the government or by the potential offenders on the expected harm (Garoupa and Obidzinski 2011). The level of deterrence of act-based sanctions depends on the

belief of the agents making the law (the authority). The level of deterrence of harm-based sanctions is more dependent to the beliefs of the people to whom the law applies (Friedman 2000). Beliefs regarding the probability of harm may considerably differ in the population. This might justify the use of act-based sanctions.

Cross-References

- [Criminal Sanctions and Deterrence](#)
- [Public Enforcement](#)

References

- Friedman DD (2000) Law's order: what economics has to do with law and why does it matter? Princeton University Press, Princeton
- Garoupa N, Obidzinski M (2011) The scope of punishment: an economic theory. *Eur J Law Econ* 31(3):237–247
- Polinsky AM, Shavell S (2007) Handbook of law and economics, vol 1. Elsevier, Amsterdam
- Rousseau S, Blondiau T (2014) Act-based Versus Harm-based Sanctions for Environmental Offenders. *Env Pol Gov* 24:439–454
- Shavell S (1993) The optimal structure of law enforcement. *J Law Econ* 36:255–287
- Wittman D (1977) Prior regulation versus post liability: the choice between input and output monitoring. *J Leg Stud* 6:193–211

Administrative Corruption

Maria De Benedetto
Department of Political Science, Roma Tre
University, Rome, Italy

Abstract

The widespread interest at national and international level in combating administrative corruption is strictly connected with the idea that it produces many negative effects, distorts incentives and weakens institutions. On the other hand, administrative corruption has been also considered as an extra-legal institution which –

under certain conditions – could even produce positive effects.

Anticorruption strategies have been developed with reference to a Principal-Agent-Client model or using an incentive-disincentive approach as well as an ethical perspective.

However, preventing corruption needs a toolbox: good quality regulation, also when regulation determines sanctions; controls, which should be sustainable and informed to deterrence and planning; administrative reforms, in order to reduce monopoly and discretionary powers, to strengthen the Civil Service and to ensure transparency and information.

Definition

Abuse of public power for private gain.

Administrative Corruption: The Definition Debate

Defining administrative corruption is not a simple task. There is large agreement about the idea that corruption crosses legal systems, history, and cultures and that it is “as old as government itself” (Klitgaard 1988, p. 7) and a “persistent and practically ubiquitous aspect of political society” (Gardiner 1970, p. 93). At the same time, there is an agreement about the separate idea that corruption is a relative concept, that it should be understood only inside a specific cultural context, and that a behavior which is considered to be corrupt in one country (or at one time) could be considered not to be corrupt elsewhere (or at different times): in other words, “corruption is the name we apply to some reciprocities by some people in some context at some times” (Anechiarico and Jacobs 1996, p. 3).

Despite this difficulty, scholars have provided a number of definitions starting from various points of view.

A first approach has focused on the moral stance of corruption (Banfield 1958) and “tends to see corruption as evil” (Nye 1967, p. 417), requiring changes in “values and norms of

honesty in public life,” because without “active moral reform campaigns, no big dent in the corrosive effects of corruption is likely to be achieved” (Bardhan 1997, p. 1335).

The second approach, however, has considered corruption also from an economic point of view, highlighting that under certain conditions, it might produce positive effects. As a consequence, corruption should be considered more objectively because it represents an “extra-legal institution used by individuals or groups to gain influence over the action of the bureaucracy” and, moreover, because “the existence of corruption *per se* indicates only that these groups participate in the decision-making process” (Leff 1964, p. 8).

Important contributions to the definition debate (Klitgaard 1988; Rose-Ackerman 1999; Ogus 2004) recognized that – in any case – “economics is a powerful tool for the analysis of corruption” (Rose-Ackerman 1999, p. xi).

Furthermore, corruption “involves questions of degree” (Klitgaard 1988, p. 7): in this light, “petty” administrative corruption has been distinguished from political (or “grand”) corruption which “occurs at the highest level of government and involves major government projects and programs” (Rose-Ackerman 1999, p. 27). There is also systemic corruption when it is “brought about, encouraged, or promoted by the system itself. It occurs where bribery on a large scale is routine” (Nicholls et al. 2006, p. 4).

One of the most important definitions of corruption is “behavior which deviates from the formal duties of a public role because of private-regarding (personal, close family, private clique) pecuniary or status gains; or violates rules against the exercise of certain types of private-regarding influence” (Nye 1967, p. 419).

However, the most used definition has been provided by transnational actors, such as the World Bank, which refers to a concept of corruption vague enough to be used in every national context and to include all kinds of corruption: “the abuse of public power for private benefit” (World Bank, Writing an effective anticorruption law, October 2001, Washington, 1; World Bank. Helping countries combat corruption: progress at the World Bank since 1997, Washington, 2000).

Corruption reveals a rent-seeking activity (Lambsdorff 2002), an effort to achieve an extra income (J. Van Klaveren, *The Concept of Corruption*, in Heidenheimer et al. 1993, 25) by circumventing (as in the case of creative compliance, R. Baldwin et al., *Understanding Regulation: Theory, Strategy and Practice*, Oxford University Press, 2012, 232) or directly by breaking the law.

Effects of Administrative Corruption on Administrative Performance

Economic effects of administrative corruption are controversial, so are the effects of corruption on administrative performance (D.J. Gould and J.A. Amaro-Reyes, *The Effects of Corruption on Administrative Performance. Illustrations from Developing Countries*, World Bank Staff Working Papers, number 580, Management and Development Series, number 7, 1983; see also Nye 1967).

On one side, there is a point of view which tends to overestimate the positive effects of corruption. Many aspects have been mentioned in this regard: positive effects have been recognized especially when corruption is “functional” to the agency’s mission (Gardiner 1986, p. 35) or when it secures, in some cases, economic development (Leff 1964) or when it corrects “bad” (inefficient) regulation (Ogus 2004, pp. 330–331). In other words, corruption “may introduce an element of competition into what is otherwise a comfortably monopolistic industry” (Leff 1964, p. 10).

On the other side, there is a different point of view which recognizes that corruption “can determine who obtains the benefits and bears the costs of government action” (Rose-Ackerman 1999, p. 9) and, in so doing, that “distorts incentives, undermines institutions, and redistributes wealth and power to the undeserving. When corruption undermines property rights, the rule of law, and incentives to invest, economic and political development are crippled” (Klitgaard 2000, p. 2).

However, even though some economic and bureaucratic benefits of corruption have been recognized, “in the large majority of cases intuition suggests that there will be significantly outweighed by the costs” (Ogus 2004, p. 333).

In particular, corruption has a “destructive effect [...] on the fabric of society [...] where agents and public officers break the confidence entrusted to them” (Nicholls et al. 2006, p. 1; see also O.E. Williamson, *Transaction-Cost Economics: The Governance of Contractual Relations*, in “*Journal of Law and Economics*”, Vol. 22, No. 2, 1979, 242: “Governance structures which attenuate opportunism and otherwise infuse confidence are evidently needed”).

Furthermore, corruption has been regarded as a “sister activity” of taxation, but it has been considered to be more costly: in fact, it presupposes secrecy which “makes bribes more distortionary than taxes” (Shleifer and Vishny 1993, p. 600) and which represents the greatest threat to the integrity of public officials.

Combating Corruption: Why?

High levels of corruption have been econometrically associated with lower levels of investments as a share of Gross Domestic Product (GDP) (Mauro 1995) even if “the connection between corruption and the lack of growth is more often assumed than demonstrated” (Ogus 2004, p. 229). This is one of the reasons which justifies the increasing national interest in combating corruption.

There is, also, a wider interest in anticorruption policies characterized by an international effort which presents important practical consequences, e.g., the case for the anticorruption prerequisites in World Bank loans to developing countries (Guidelines on “Preventing and Combating Fraud and Corruption in Projects Financed by IBRD Loans and IDA Credits and Grants”, 2006-revised 2011) which require the putting in place of regulatory and institutional mechanisms to fight corruption (Ogus 2004, p. 329). Anticorruption policies are, in such cases, a sort of condition for obtaining a loan.

In order to clarify some aspects relevant to combating corruption, it would be important to make note of the lack of data in this area (Gardiner 1986, p. 40) as well as of the difficulty in

measuring it (Anechiarico and Jacobs 1996, p. 14). Furthermore, it is also important to remember that a large part of the institutional debate is focused on corruption perception rather than on corruption reality and (finally) that the only sustainable institutional goal is reducing corruption because corruption is considered impossible to eradicate (Ogus 2004, p. 342): “anti-corruption policy should never aim to achieve complete rectitude” (Rose-Ackerman 1999, p. 68).

As a consequence, “the optimal level of corruption is not zero” (Klitgaard 1988, p. 24). Anti-corruption controls, in fact, are expensive (Anechiarico and Jacobs 1996), so it could be necessary to decide the extent to which we should combat corruption: the point of intersection of the curves – which describe the quantity of corruption and the marginal social cost of reducing corruption – identifies “the optimal amount of corruption” (Klitgaard 1988, p. 27). In other words, when we say “why combat corruption?”, in some way we simply mean “why keep corruption under control?”

Looking closely at the question, one of the most important reasons for which corruption should be controlled (a reason characterized at the same time by a moral and an economic stance) is that corruption represents a “form of coercion, namely economic coercion” (C.J. Friedrich, *Corruption Concepts in Historical Perspective*, in Heidenheimer et al. 1993, 16) which produces a fundamental distortion in the economic process, artificially separating economic activity and its result into two abstract concepts (M. De Benedetto, *Ni ange, ni bête*. Qualche appunto sui rapporti tra morale, economia e diritto in una prospettiva giuspubblicistica, in “Nuove autonomie”, no. 3, 2010, 657, quoting the Italian legal philosopher Giuseppe Capograssi). The result benefits someone else even though it belongs to others (see A.G. Anderson, *Conflicts of Interest: Efficiency, Fairness and Corporate Structure*, in “UCLA Law Review”, vol. 25, 1978, 794). In so doing, corruption changes the value (i.e., the result of the economic activity, at the same time moral and economic value) into mere advantage and the economic process in itself is corrupted.

Preventing Corruption: How?

Anticorruption could be considered as a “comprehensive strategy” which should be built with a number of tools (Rose-Ackerman 1999, p. 6; J.A. Gardiner and T.R. Lyman, *The Logic of Corruption Control*, in Heidenheimer et al. 1993, 827).

The first point in preventing corruption is to recognize that there are different kinds of transaction which directly produce (or indirectly stimulate) corrupt behavior and that these should be considered as separate battlefields which need specific tools.

In this regard, some contributions have used the “Principal-Agent-(Client)” model (Banfield 1975; Rose-Ackerman 1978; Klitgaard 1988; Della Porta and Vannucci 2012). In ordinary cases of corruption, there is a bribe giver and a bribe taker, but corrupt transactions involve three actors: the Principal (the State and its citizens, always considered the victims), the Agent (the civil servant in charge of administrative tasks), and the Client (the enterprise or the citizen, e.g., as tax payer).

Corrupt transactions in strict sense can be performed in the Agent-Client relationship (e.g., bribery, extortion). The Principal-Agent relationship as well as the Principal-Client relationship could present behavior oriented toward illicit rent seeking (e.g., in the first case, internal fraud and theft of government properties; in the second case, tax frauds or illegal capital transfers), very often facilitated by widespread conflicts of interest (Auby et al. 2014), but “this is not considered to be corruption since it does not include the active (or passive) collusion of an agent of the state” (Robinson 2004, p. 110).

The institutional response to prevent direct or indirect corruption in these different kinds of transaction involves a tool-kit: internal or external controls over administrative action (P/A), inspections on private economic activities (P/C), and criminal investigations into specific cases of corruption (A/C).

The second point in order to prevent corruption regards the opportunity to adopt an incentive/dis-incentive approach, changing the system of

rewards and penalties (Klitgaard 1988, p. 77; Gardiner 1986, p. 42) in a mix of “carrots and sticks” (Rose-Ackerman 1999, p. 78). There is large agreement about the opinion that “corrupt incentives exist because state officials have the power to allocate scarce benefits and impose onerous costs” (Rose-Ackerman 1999, p. 39). Reducing incentives to corruption and increasing its costs could involve structural reform (see *infra* 4.3) as “the first line of attack in an anticorruption campaign” (Rose-Ackerman 1999, p. 68).

The third point implies a problematic analysis of the large recourse – in institutions as well as in business, at national level as well as at the international one – to ethical codes and other similar tools in order to ensure ethical responses to corruption and to strengthen anticorruption policies by stimulating individual morality. However, this tendency to regulate ethics is more effective in some cultural contexts than in others but could produce side effects. Ethics is typically free, while law is characterized by coercion (and by sanctions). If we use legal provisions (or any kinds of sanction) to induce ethical behavior, we are transforming a free behavior into a legal obligation, reducing the moral involvement of the individual, even if the legal provision is established by soft laws (such as ethical codes): this is a real paradox in regulating ethics (M. De Benedetto, *Ni ange, ni bête* cit., 656; see also Anechiarico and Jacobs, 1996, p. 202). In other words, tools should be consistent with the objective: law can establish incentives for moral behavior but cannot either impose or produce a moral (free) behavior by coercive means.

Regulation

As we have seen, the problem of corruption has in part a moral and a social stance, so a first step in preventing corruption would be for regulation to accept its own limits and to recognize that not only legal but also extralegal norms operate (on this point see R. Cooter, Expressive Law and Economics, in “The Journal of Legal Studies”, vol. XXVII, June 1998, 585).

Furthermore, anticorruption policies have better chances of success if legal provisions and public opinion converge. The problem is particularly

relevant in cases of “gray” corruption (A.J. Heidenheimer, Perspectives on the Perception of Corruption, in Heidenheimer et al. 1993, 161) in which there is a mismatch between what is considered corruption by public opinion and what is corruption by law. So, if regulation wants to achieve anticorruption objectives, it should take into account the social and moral context in which it will be applied and – in this way – it will strengthen enforcement and increase compliance: “the majority of government employees are honest, not because of rules, monitoring or threats but because of value and personal morality” (Anechiarico and Jacobs 1996, p. 202).

Moreover, it has long been clear that the functioning of market economy needs “a firm moral, political and institutional framework,” which implies “a minimum standard of business ethics”. The market economy, indeed, is not capable of increasing the “moral stock” by itself because “competition reduces the moral stamina and therefore requires moral reserves outside the market economy” (W. Röpke, *The Social Crisis of our Time*, Transaction Publishers, 1952, 52). This seems to be even more true when the individual choice on ethical rules takes place inside large groups (J.M. Buchanan, *Ethical Rules, Expected Values, and Large Numbers*, in “Ethics”, vol. 76, 1965, 1).

Secondly, far from being a solution, regulation is recognized as a direct factor that promotes corruption (Tanzi 1998, p. 566). Overregulation could increase bureaucratic power and multiply the opportunities for creative compliance, nurturing a corruptible social environment and allowing more and more corruption: “the possibility of its transgression or perversion is always already inscribed into the law as hidden possibility. This, then, is the secret of law” (Nuijten and Anders 2007, p. 12).

Thirdly, since regulatory processes are fragile, special attention should be paid to sensible steps in the procedures: consultations, for example, could “increase the opportunity for corrupt transactions” (Ogus 2004, p. 341).

Starting from these premises, the problem seems to be not anticorruption regulation but good regulation in itself, regulation capable of

making rules effective, of ensuring enforcement, and of increasing compliance (in general on this point, Becker and Stigler 1974). Regarding the content of such good regulation, anticorruption objectives could be achieved thanks to reducing monopoly and discretion (Rose-Ackerman 1999) as we will see later (par. 4.3).

Another important regulatory matter in preventing corruption concerns sanctions (Klitgaard 1988, p. 78). They should be well calibrated because they respond to an intrinsically economic logic – indispensable to making laws effective – and because they can even influence the amount of the bribe: “penalizing the official for corruption changes the level of the bribe he demands, but does not change the essence of the problem” (Shleifer and Vishny 1993, p. 603; Ogus 2004, p. 336; see also Svensson 2003). On the other hand, it should be clear that “in the presence of corruption, it is optimal to impose (or at least threaten to impose) nonmonetary sanctions more often” (Garoupa and Klerman 2004, p. 220) as well as to reward enforcement (Becker and Stigler 1974, p. 13) and compliant groups (Gardiner and Lyman, *The Logic of Corruption Control* cit., in Heidenheimer et al. 1993, 837; Ogus 2004, p. 337): this idea is not new and was, in a similar form, already proposed in 1766 by Giacinto Dragonetti in his “Treatise on Virtues and Awards” (first English translation 1769).

Controls

In order to prevent corruption, controls are needed because human behavior is fallible and corruptible, because human behavior changes when subject to controls, and because corruption lives in the dark and controls may constitute the most important tool in rebalancing the asymmetric information between corrupt people and institutions (in this regard, compensating whistleblowers has been considered critically by Anechiarico and Jacobs 1996, p. 199 and Ogus 2004, p. 338). Among the different kinds of controls, inspections constitute the strongest tool, because they are characterized by coercive power.

On the other hand, traditional corruption controls have been considered inadequate and even

“outdated and counterproductive” (Anechiarico and Jacobs 1996, p. 193).

Furthermore, it should be taken into account that controls (e.g., inspections) have a hybrid nature: not only are they a way to combat or prevent corruption but also they are real occasions for corrupt transactions because they represent a concrete contact between the Agent and the Client, particularly sensitive and dangerous when the Client has an interest to maintain (in any case and at any condition) the extra income which comes from illicit activities or when the Agent (who want to achieve an illicit extra income) has the opportunity to extort the Client.

The system of controls should be, indeed, sustainable from an administrative point of view, also in the field of anticorruption. In particular, it would be important to reduce the number of controls, because they represent a cost (not only for public administration but also for enterprises and citizens; see in this regard the Hampton Report, H.M. Treasury, *Reducing administrative burdens: effective inspection and enforcement*, March 2005) and because they are (as we have seen) occasions for corruption.

At the same time, it is important to increase their effectiveness in preventing corruption cases: for this purpose, anticorruption controls as a system should be informed by deterrence. The most relevant general contribution on this topic (Becker 1968) suggests that the individual decision about compliance is a result of an economic reasoning which connects the cost of compliance, the size of the penalty, and the risk of incurring the penalty. This reasoning, on the same grounds, contributes to establishing the eventual size of the bribe (Ogus 2004, p. 336). Furthermore, planning controls in anticorruption policies should be guided by a risk-based approach (in general, R. Baldwin et al., *Understanding Regulation* cit., 281), capable of mapping the most dangerous areas of administrative activity (in terms of probability of corruption) and capable of focusing – between the possible objects of control – on cases in which it is more probable to find evidence of corruption.

Administrative Reforms

There is a sort of conflict – observed by some scholars – between anticorruption policies and administrative reform. When anticorruption prevails, administrative reforms seem to be reduced in importance or to become marginal: “the logic of antibureaucratic reform leads to a model of public administration that ignores corruption, while the logic of anticorruption reform ignores public administration” (Anechiarico and Jacobs 1996, p. 204). It could even be possible that governments’ anticorruption effort produce further costs “[...] not only in terms of the funds spent to control corruption, but in the deflection of attention and organizational competence away from other important matters” (Klitgaard 1988, p. 27).

At the same time, it could be useful to approach the problem of administrative reforms (in the perspective of anticorruption) in a practical manner, because “in terms of economic growth the only thing worse than a society with a rigid, overcentralized, dishonest bureaucracy is one with a rigid, overcentralized, honest bureaucracy” (Huntington 1968).

The first idea, in this regard, is to reduce the public sector: “the only way to reduce corruption permanently is to drastically cut back government’s role in the economy” (Becker 1997).

The second relevant aspect is to reduce monopoly and discretionary powers (Ogus 2004, p. 331): “insofar as government officials have discretion over the provision of these goods [licenses, permits, passports and visas], they can collect bribes from private agents” (Shleifer and Vishny 1993, p. 599).

The third aspect regards the civil service (Rose-Ackerman 1999, p. 69) also because there are bureaucracies which seem to be less corruptible than others (S. Rose-Ackerman, *Which Bureaucracies are Less Corruptible?*, in Heidenheimer et al. 1993, 803). The question should be analyzed both from the point of view of civil service independence (Anechiarico and Jacobs 1996, p. 203) and from the point of view of civil service incentive payments (“often cited as one of the most effective ways of fighting corruption”, Bardhan 1997, p. 1339).

The fourth aspect involves procedural and organizational design (Ogus 2004, p. 338) as well as administrative cooperation, crucial in order to enforce regulation and to effectively prevent corruption cases: in fact, “internal organisation of institutions influences their members’ propensity to corruption” (Carbonara 2000, p. 2). Among other aspects, increasing international cooperation in combating corruption is indispensable: this is clear at the EU level, where it was recently affirmed that “anti-corruption rules are not always vigorously enforced, systemic problems are not tackled effectively enough, and the relevant institutions do not always have sufficient capacity to enforce the rules” (EU Anti-corruption Report, COM 2014, 38 final, 2). Furthermore, this is confirmed at further levels, as in the case of GRECO, Group of States against Corruption, established in order “to improve the capacity of its members to fight corruption” (Statute of the GRECO, Appendix to Resolution (99) 5, art. 1) or as in the case of the OECD Convention on Combating Bribery of Foreign Public Officials in International Business Transactions.

Overall, indeed, there is a general problem of transparency and information. Transparency reduces the opportunities for corruption “through procedures which make the process and content of decision-making more visible” (Gardiner and Lyman, *The Logic of Corruption Control* cit., in Heidenheimer et al. 1993, 830). In the same way, information on what the Agent and the Client are doing allows the principal “to deter corruption by raising the chances that corruption will be detected and punished” (Klitgaard 1988, p. 82).

However, every anticorruption project needs a fine-tuning (Anechiarico and Jacobs 1996, p. 198) which implies both a legal and an economic approach, but which should by now be open to the contributions of other disciplines (e.g., behavioral sciences). In any case, “scholars of law and economics” will continue developing studies in the area of corruption also “because it raises fascinating issues about the enforcement of law in the broadest sense” (Bowles 2000, p. 480).

References

- Anechiarico F, Jacobs JB (1996) The pursuit of absolute integrity. How corruption control makes government ineffective. The University of Chicago Press, Chicago
- Auby JB, Breen E, Perroud T (eds) (2014) Corruption and conflicts of interest. A comparative law approach. Edward Elgar, Cheltenham
- Banfield EC (1958) The moral basis of backward society. Free Press, New York
- Banfield EC (1975) Corruption as a feature of government organization. *J Law Econ* 18:587–605
- Bardhan P (1997) Corruption and development: a review of issues. *J Econ Lit* 35:1320–1346
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169
- Becker GS (1997) Want to squelch corruption? Try passing out raises. *Business Week*, no. 3551, 3 Nov 1997, p 26
- Becker GS, Stigler GJ (1974) Law enforcement, malfeasance, and compensation of enforcers. *J Leg Stud* 3:1–20
- Bowles R (2000) Corruption. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics*, volume V. The economics of crime and litigation. Edward Elgar, Cheltenham, pp 460–491
- Carbonara E (2000) Corruption and decentralisation, dipartimento scienze economiche, Università di Bologna, Working papers no. 342/83
- Della Porta D, Vannucci A (2012) The hidden order of corruption. An institutional approach. Ashgate Publishing, Farnham, Surrey
- Gardiner JA (1970) The politics of corruption. Russel Sage Foundation, New York
- Gardiner JA (1986) Controlling official corruption and fraud: bureaucratic incentives and disincentives. *Corrupt Reform* 1:33–50
- Garoupa N, Klerman D (2004) Corruption and the optimal use of nonmonetary sanctions. *Int Rev Law Econ* 24:219–225
- Heidenheimer AJ, Johnston M, Levine VT (eds) (1993) Political corruption. A handbook. Transaction Publishers, New Brunswick (third printing, first printing 1989)
- Huntington SP (1968) Political order in changing societies. Yale University Press, New Haven, p 69
- Klitgaard R (1988) Controlling corruption. University of California Press, Berkeley and Los Angeles
- Klitgaard R (2000) Subverting corruption. *Financ Dev* 37:2–5
- Lambsdorff JG (2002) Corruption and rent-seeking. *Public Choice* 113:97–125
- Leff N (1964) Economic development through bureaucratic corruption. *Am Behav Sci* 8:8–14
- Mauro P (1995) Corruption and growth. *Q J Econ* 110(3):681–712
- Nicholls C, Daniel T, Polaine M, Hatchard J (2006) Corruption and misuse of public office. Oxford University Press, Oxford
- Nuijten M, Anders G (eds) (2007) Corruption and the secret of law. A legal anthropological perspective. Ashgate Publishing, Aldershot, Hampshire
- Nye JS (1967) Corruption and political development: a cost-benefit analysis. *Am Polit Sci Rev* 61(2): 417–427
- Ogus A (2004) Corruption and regulatory structures. *Law Policy* 26:329–346
- Robinson M (ed) (2004) Corruption and development. Routledge, London
- Rose-Ackerman S (1978) Corruption. A study in political economy. Academic press, New York
- Rose-Ackerman S (1999) Corruption and government: causes, consequences and reform. Cambridge University Press, Cambridge
- Shleifer A, Vishny RW (1993) Corruption. *Q J Econ* 108:599–617
- Svensson J (2003) Who must pay bribes and how much? Evidence from a cross section of firms. *Q J Econ* 118(1):207–230
- Tanzi V (1998) Corruption around the world: causes, consequences, scope, and cures. *IMF Staff Pap* 45(4): 559–594

Administrative Courts

Sofia Amaral-Garcia

Center for Law and Economics, ETH Zurich,
Zürich, Switzerland

Definition

Administrative courts are courts that specialize in and apply administrative law, a branch of public law that focuses on public administration. In other words, administrative courts adjudicate cases of litigation involving the state and private citizens (both individuals and corporations) and apply the law that governs the activities of administrative agencies of government. There is no homogeneous model for administrative courts across jurisdictions. Even if one considers a broad distinction of legal families, it is possible to find many distinct models. Every jurisdiction has a way of solving litigation with the state, independently of having a specialized administrative court.

Introduction

Administrative courts are courts that specialize in administrative law, a branch of public law that

focuses on public administration. In other words, administrative courts adjudicate mainly cases of litigation involving the state (both institutions and public officials) and private citizens (both individuals and corporations) and apply the law that governs the activities of administrative agencies of government. The role of these types of courts is directly related to the existence and functions of a state or government: in the one hand, it would make no sense for these courts to exist without a state; on the other hand, these courts may have distinct roles depending on the specific functions of the state, which might vary across jurisdictions and time.

Constitutional courts (if existent) also adjudicate cases related to the government. For this reason, one might argue that it is hard to draw a clear-cut line dividing the types of cases that each court hears. Assigning the tripartite division of government branches to administrative courts and constitutional courts might help somehow drawing a division, even though it is never an exclusive one. Generally speaking, administrative courts adjudicate primarily cases related to the executive function of the state, and constitutional courts adjudicate cases related to the legislative, judicial and executive functions.

Institutions in general, and courts in particular, matter for economic growth and development (e.g., Mahoney 2001). Scholars from law and economics, institutional economics and legal studies have acknowledged their role and importance, especially in recent decades. Many answers are still unanswered, in particular if we narrow the analysis down to administrative courts as extensive empirical analysis in administrative courts is virtually nonexistent. The emerging field of comparative administrative law has comprehensively examined and explained substantial differences of administrative law across countries. Nevertheless, the collection of administrative courts' models (even within the same legal families) embedded in highly complex legal, political, and economic systems create a challenge to comparative analysis. In fact, it is difficult to assess which model is the best or to quantify how much one model is better than the other because an extensive number of factors play a role as well. The transplant of

other models does not necessarily result in better outcomes precisely for this reason.

"The core idea of government under law requires an independent institution willing to enforce that law against the administration – both the political and the bureaucratic administration – whenever the law is broken" (Bishop 1990, pp. 492–493) – this is where administrative courts come in. Administrative courts hear a variety of cases similar to those involving individuals, such as contracts, torts or property, as long as the state is one of the parties involved. Nevertheless, the state has particular functions that private citizens do not have. Declaring war, issuing passports, collecting taxes, having the monopoly of legitimate coercion (which also imposes the duty of procedural fairness), or having a monopoly over some activities and provision of certain goods or services are a few examples of state's specific functions (e.g., Cane 2011). Generally speaking, some of the most common areas of administrative law are taxes, immigration, and licensing which tend to make up for a significant proportion of cases that administrative courts hear. Moreover, these areas are more relevant in civil law countries given their inclination for *ex ante* control mechanisms rather than *ex post* control mechanisms as it tends to be the case in common law systems. To sum up, administrative courts also resolve disputes that are specific to the role of the state and protect citizens from state overreaching.

Historical Origins and Institutional Differences

Although it is not the purpose of this work to make an extensive analysis of all administrative courts in Europe or elsewhere, some basic notions of the models adopted in England and France are useful at this stage. France and England deeply influenced many legal systems across the world, even though these countries opted for two different ways of dealing with administrative law. In England, ordinary courts were in charge of hearing administrative cases, whereas this role was assigned to specialized state officials connected

to the executive branch in France. This is generally perceived to be the major difference between administrative law systems (Bignami 2012, p. 148). There were important historical differences in those two countries with respect to the rise of bureaucracy and administrative law that explain the option for “ordinary” or specialized courts.

In France, the seventeenth and eighteenth centuries were marked by intense conflicts between *intendants* (officers responsible to the Crown in charge of the provinces’ administration) and *parlements* (powerful courts controlled by local elites). In fact, the “paralysis of the royal administrative of the ancient régime, caused by the *Parlements* (which were, in fact, judicial bodies), is often considered one of the causes of the French revolution and was a prime impetus for the French conception of separation of powers, in which the judicial courts lack jurisdiction over administrative acts of the State” (Massot 2010, p. 415). Napoleon created the *Conseil d’État* (Council of State), which became the successor of the *Conseil du Roi* (King’s Council, a specialized review body) and the first specialized administrative court. The Council of State still maintains nowadays the dual function of adjudicating cases against the French administration and drafting government laws and rules. At the time of its creation, judicial review made by ordinary courts (i.e., the judicial power of the government) represented an intrusion on the executive power and these courts were not allowed to adjudicate claims against the government. This was the main rationale for assigning that task to a specialized body: the executive, not the judiciary (Cane 2011, p. 43; Bignami 2012, p. 149; see also Mahoney 2001).

In England, the distinction between public and private law has been historically less sharp not because England lacks public-law courts but because ordinary courts have jurisdiction over all types of disputes (Cane 2011). A constitutional struggle developed in the seventeenth century between the Stuarts and judges with respect to the judges’ right to decide cases related to the royal power or to decide cases in which the king had an interest (Page and Robson 2014). At that

time, the Stuart kings attempted to create separate courts to deal with cases related to the government in order to expand royal control. However, the victory of the Parliament established the independence of judges and everyone should obey the law. Courts developed judicial review, i.e., the revision of administrative decisions, as a way of supervising inferior government bodies (Cane 2011). Local elites with little central involvement would administer the business of government and appeals against government officers would be taken to courts of general jurisdiction (Bignami 2012, p. 149). Administrative tribunals developed during the twentieth century (Shapiro 1981, p. 111 & seqs.) which, in a very simplistic way, can be thought of as an adjudicatory body that is not a court (Cane 2010).

The German model of judicial review is an alternative to the French and English models. During the nineteenth century, German liberals endeavored to implement legal structures that would limit state power by the monarchs of German states. At the time, the ideas of the jurist Rudolf von Gneist strongly influenced German administrative law. Gneist was a strong advocator of an independent judicial review system that would ensure the protection of citizens’ rights and contended that there should be a generalist administrative court independent of the executive. Moreover, the composition of the court should allow an independent judicial control of administrative power, so it would be staffed by professional administrators and respected citizens (Feld 1962, p. 496; Nolte 1994, p. 199). In 1872 Gneist’s ideas were implemented with the Prussian Supreme Administrative Court, which was the highest judicial body of a three-tier system of administrative courts. This court exerted a great influence on the development of German administrative law (Feld 1962). According to the German model, currently the most widespread model in Europe, a specialized branch of the judiciary specializes in administrative law (Fromont 2006, p. 128). Civil judges and administrative judges follow the same recruitment process and guarantee of independence. The main difference is that administrative judges specialize in administrative law.

The spirit of the French, English, and German administrative models has been implemented elsewhere. Several countries adopted the French administrative model, having a Council of State separated from the judiciary. In continental Europe, some of these countries are the Netherlands, Belgium, Italy, and Greece. However, whereas cases of government liability are adjudicated by the Council of State in France, these cases are adjudicated by courts in Italy, Belgium, and the Netherlands. The British model of a generalist court has been implemented, among others, in Ireland, the USA, Australia, and New Zealand. Austria, Finland, Poland, Portugal, Spain, Sweden, and Switzerland have implemented a model of the German type.

The Evolution of the State and the Role of Administrative Courts

The current role of administrative courts is necessarily intertwined with the functions of the modern state, which are complex and diverse across jurisdictions. Moreover, these functions are not static and evolve continuously, together with economic and political development (the development of the welfare state has been particularly relevant). For instance, the provision of health, housing, education, electricity and transport, among others, went through diverse degrees of public ownership and control (Cane 2011). During the 1980s and the 1990s several Western countries experienced a shift in the boundary between the public and private sectors. With the aim of reaching a single market, the European Commission liberalized numerous network industries, such as gas, air transport, electricity, postal and railroad services (Custos 2010, p. 279). In some countries, there was a shift from public to private in many economic sectors due to privatizations (for example, in the UK, Portugal, Italy, France, and Spain). Meanwhile, a deregulation process took place in the USA, where government regulation was reduced in a number of sectors. The government opted for contracting-out some services to private companies and public-private partnerships emerged for performing functions

that have previously been carried out by governmental bodies. A few examples are road repairs, highways constructions, and garbage collection.

In particular, nowadays the state is a constant presence in many spheres of daily life, its powers are vast, and its functions are considerably complex. An enormous machinery must be in place so that the state can perform its complex functions. The legislature approves laws and statutes that will be executed by government agencies, run by nonelected civil servants. A democratic government running under the law must provide a way of monitoring and supervising the performance of government agencies and bureaucrats. Therefore, administrative courts can be asked to perform judicial review of administrative decisions (e.g., if a public body acted beyond its powers or if a public body failed to act or perform a duty statutorily imposed on it).

The actions of state officials might also impose harm to citizens, very similarly to what happens in contract or tort. Some examples include medical liability of a doctor practicing in a public hospital or accidents caused by state-owned cars and driven by public employees. The difference in these cases is that the state is one of the parties, most likely the party supposedly causing the harm. Administrative courts might be called to adjudicate state liability and award damages in case the state is found liable, but the reliance on administrative courts to perform this task depends on each specific jurisdiction. In reality, there is no unique model for administrative courts or for adjudicating litigation involving the state. In terms of procedure, administrative courts in civil law tradition countries have exclusive jurisdiction over tort litigation against the state as sovereign (Dari-Mattiacci et al. 2010). Moreover, administrative courts follow rules of administrative procedure that tend to treat the state as a nonordinary defendant (differences might be found with respect to statute of limitation, liability standards or the possibility of having out-of-court settlements, to name only a few). Cases in which the state acts as a private entity might lead to jurisdictional ambiguity, and depending on the country these cases might end up in ordinary courts. In some countries, cases of jurisdictional conflict

might be addressed to a court of conflict (e.g., France, Italy, and Portugal).

Some Examples of Institutional Differences

It is nevertheless worth mentioning that a jurist trained in the USA might face some difficulties when trying to understand the functioning of administrative courts and the profession of administrative judges in Continental European civil law tradition countries. In the USA, much of administrative review is vested in public authorities and independent agencies, often described as administrative tribunals. The agency officials, whose task is to adjudicate cases according to the Administrative Procedure Act (APA) of 1946, are administrative law judges and administrative judges (Cane 2010, p. 427). These administrative officials are typically recruited by the Office of Personnel Management to become adjudicators in agencies that are part of the executive branch of government. The APA did not introduce specialized courts to deal with judicial review of administrative acts but it effectively created “a special form of jurisdiction that governs the review of agency decisions in ordinary courts. Especially to continental jurists, then, it is often worth emphasizing that the US legal system, in some sense, also distinguishes between administrative law questions and other legal disputes” (Halberstam 2010, p. 187). Moreover, in the US, the Court of Appeals for the District of Columbia (DC) Circuit has specialized in administrative law and is frequently the final court for administrative matters.

Recently in the UK there were important judicial reforms, namely the Constitutional Reform Act 2005 and the Tribunals, Courts and Enforcement Act 2007. According to this later reform, administrative courts are effectively being created. Additionally, the jurisdiction of some subject-specific tribunals (such as social security, education, taxation, pensions, and emigration) has been transferred to the new First-tier Tribunal and Upper Tribunal. The First-tier Tribunal comprises seven chambers has as main function deciding general appeals against a decision made by a Government agency or department. The Upper Tribunal comprises four chambers (one of which

being the Administrative Appeals Chamber) and hears appeals from the First-tier Tribunal in points of law. In a way, recent times have introduced a sharper distinction between administrative and private English law.

Significant transformations also took place in French administrative law, namely with the creation of the administrative courts of appeal in 1987 and various reforms that extended the powers of administrative judges (in particular with respect to injunctions and the possibility of issuing urgent judgments to prevent illegal administrative behavior) (see Auby and Cluzel-Métayer 2012, p. 36). Currently, there are 42 administrative tribunals, 8 administrative courts of appeal and the highest level is still the *Conseil d'État*. French administrative courts do not adjudicate all administrative cases. The view is that the administration is only partly subjected to special rules and to public law, which translates in some cases ending up in ordinary courts even if they involve the administration. Delimiting the precise jurisdiction of French administrative courts is extremely complex (for more on this see Auby and Cluzel-Métayer 2012, p. 25).

With respect to Germany, administrative courts are one of the five branches of the judiciary, which besides administrative courts includes ordinary courts (civil and criminal), labor courts, fiscal courts, and social courts. The last three courts comprise the so-called special courts (e.g., Schröder 2012, p. 77). There are three levels of administrative courts: lower administrative courts, higher administrative courts, and the Federal Administrative Court. The total number of lower administrative courts depends on the population and size of each *Land* (state), there is at least one administrative court in each State but no more than one higher administrative court (Singh 2001, p. 187 & seqs.; Schröder 2012, p. 78 & seqs.).

Generally speaking, civil law systems tend to hear cases of judicial review within one administrative court (with a few exceptions, such as social security cases in some jurisdictions or taxes in Germany) whereas common law systems tend to have a different appeal tribunal for each agency (Bishop 1990, p. 525). However, the few examples presented above illustrate the complexity of

the organizational setting of administrative courts. Homogeneity cannot be found, not even at the highest court level of Continental European jurisdictions where some systems have one Administrative Supreme Court (e.g., Portugal), others have one Administrative Section within the Supreme Court (e.g., Spain) and others have none of these. Independently of institutional and legal differences, every jurisdiction has a way of solving litigation between the state and citizens and of appealing these decisions. The main differences are generally with respect to whether there is a specialized administrative court, whether there are particular procedural rules, and whether there are specialized courts to deal with conflicts of jurisdiction.

Do We Need Specialized Administrative Courts?

The debate on specialized administrative courts has raised several contentious questions that have been asked for decades (see, e.g., Nutting 1955; Revesz 1990). For instance, should the adjudicative function of administrative cases be vested in separate bodies and, if so, are courts the most appropriate institutions? Should administrative courts have original and appellate jurisdiction? Is it more beneficial to have specialized administrative courts to adjudicate cases involving the state or leave it to generalist courts? The answers to these questions remain controversial, with several arguments being found in favor and against specialized courts (see, among others, Dreyfuss 1990; Revesz 1990; Baum 2011).

Some of the potential benefits of specialized courts are extensive to administrative courts as well. The “neutral virtues” of judicial specialization are the quality of decisions, efficiency, and uniformity in the law (Baum 2011, pp. 4 & 32). Specialization should result in more correct decisions in complex areas of the law, precisely because the adjudicator will become an expert in a certain type of decisions. This can be particularly relevant in fields of law that involve complex technical skills. Moreover, specialization might

allow decisions to become more uniform and coherent. One potential advantage of specialized courts on administrative matters might be that the judges feel more confident given their expertise, which may make them more willing to go against administrative decisions. If judges have the incentives and training to become experts in administrative law, it is also possible to have tailored procedures in court to deal with the particular features of the state as defendant (Dari-Mattiacci et al. 2010, p. 28). A major advantage of separate administrative courts is the possibility to develop a set of principles that accepts the specific nature of the state as defendant (such as access to information, evidence produced by the administration and control over administrative discretion) balancing the interests of citizens and the ability of the administration to pursue the public interest (Bell 2007, pp. 291–293 and p. 299). There are possible “nonneutral effects” of specialized courts on the substance, namely a change in the ideological content of judicial policy or the support for competing interests in a given field (Baum 2011, p. 4 & seqs.).

As for the arguments against specialized courts, there are also many. Courts and judges that become specialized in some particular area of the law might apply the law in a narrower way, and have fewer skills in applying concepts from other areas of law when necessary. Moreover, some arguments against specialized administrative courts are particularly relevant. Specialized administrative courts might be more easily captured by the state, which becomes more likely with the separation of jurisdictions. The marginal cost for the judge in deciding against the state is higher in administrative than in ordinary judicial courts (e.g., Mahoney 2001). Specialization makes accountability more difficult, because the knowledge of administrative law becomes a specific asset on human capital for administrative judges, which might become more dependent on the government precisely for this reason (see Dari-Mattiacci et al. 2010, p. 28). Hence, it might be more difficult for administrative judges to realize that the state has overreached either because judges exhibit systemic biases or because judges want to maintain their place as state officials. This

would be an undesirable outcome in case it would result in administrative judges issuing biased decisions in favor of the state. However, and in spite of the debate on this issue, the lack of empirical evidence makes it impossible to draw rigorous and extensive conclusions on this claim.

The recruitment of administrative judges becomes naturally a critical piece of the well-functioning of administrative courts. Traditionally, commonwealth systems have recognition judiciaries which, together with ordinary courts for administrative review, tends to result in a higher degree of autonomy in comparison with continental systems (Garoupa and Mathews 2014, p. 12). In the majority of civil law countries in Continental Europe, administrative judges follow a career as generalist judge and specialize in administrative law when serving in administrative courts. An exception is France, where most of the members of the *Conseil d'Etat* are civil servants, recruited from the *École Nationale d'Administration*, the elite school for training French government executives. In both civil and common law systems judges are appointed for life: in civil law, judges are appointed for life as judges (i.e., career tenure) and in common law judges are appointed for life in a specific court (i.e., court tenure).

It is not possible to properly assess the pros and cons of specialized administrative courts without considering the legal, administrative lawmaking and political systems in which a particular court is located in. In the end, judicial specialization may affect courts output in a very complex way. Whether the effects are good or bad depend on the particular system where the court is located. There are still limited empirical findings which makes it difficult to assess the effects of specialized administrative courts. Moreover, even if there were strong evidence that the benefits of specialized administrative courts outweigh the disadvantages, it would still be unclear what the best institutional model for administrative courts would be. Indeed, there are many different ways to structure specialized courts, namely specialized courts with generalized judges, generalized courts with exclusive special jurisdiction, specialization at the trial and/or appellate level, specialized trial

courts with general appellate courts or general trial courts reviewed by special appellate courts (Dreyfuss 1990, p. 428). Furthermore, as the brief comparison among different jurisdictions has shown, it is possible to have distinct formats even at the highest court levels. Additionally, the interactions between judges and the state may have important implications which might be translated in more or less politicized courts (see Garoupa et al. 2012).

Concluding Remarks

Administrative courts are part of the system of checks and balances of the government system. Therefore, administrative courts have an important institutional role in modern societies, given that they still keep the task of protecting citizens from the powerful state. In the same way that modern societies have institutions and mechanisms to enforce the law and solve conflicts among private parties, it is fundamental to allow citizens to review government decisions and to be compensated if they have been harmed by government action or inaction. The function of administrative courts and the quality of their decisions can also have relevant economic impacts. All in all, it is not only important to have a way of making a claim against the state: it is also important that the decisions being held by the institutions in charge of the adjudication are not prostrate biased.

Recent decades have introduced changes in the functions of the state, with the development of the welfare state being eventually the most relevant. Naturally, the role of administrative courts has also been affected by these changes. Concomitantly, important developments took place with repercussions in the administrative sphere, such as the creation of the European Union and the establishment of the European Convention on Human Rights. It is still unclear what the implications of global administrative on national administrative courts will be. Contributions from political science and law and economics would be fruitful and would bring important insights to the debate.

References

- Auby J-B, Cluzel-Métayer L (2012) Administrative law in France. In: Seerden R (ed) *Administrative law of the European Union, its Member States and the United States*, 3rd edn. Intersentia, Antwerpen
- Baum L (2011) *Specializing the courts*. University of Chicago Press, Chicago
- Bell J (2007) Administrative law in a comparative perspective. In: Öricü E, Nelken D (eds) *Comparative law: a handbook*. Hart Publishing, Oxford
- Bignami F (2012) Comparative administrative law. In: Bussani M, Mattei U (eds) *The Cambridge companion to comparative law*. Cambridge University Press, Cambridge, pp 145–170
- Bishop W (1990) A theory of administrative law. *J Leg Stud* XIX:489–530
- Cane P (2010) Judicial review and merits review: comparing administrative adjudication by courts and tribunals. In: Rose-Ackerman S, Lindseth PL (eds) *Comparative administrative law*. Edward Elgar, Cheltenham/Northampton, pp 426–448
- Cane P (2011) *Administrative law*, 5th edn. Oxford University Press, Oxford
- Custos D (2010) Independent administrative authorities in France: structural and procedural change at the intersection of Americanization, Europeanization and Gallicization. In: Rose-Ackerman S, Lindseth PL (eds) *Comparative administrative law*. Edward Elgar, Cheltenham/Northampton, pp 277–292
- Dari-Mattiacci G, Garoupa N, Gómez-Pomar F (2010) State liability. *Eur Rev Private Law* 18(4):773–811
- Dreyfuss RC (1990) Specialized adjudication. *Brigham Young Univ Law Rev* 1990:377–441
- Feld W (1962) The German administrative courts. *Tulane Law Rev* XXXVI:495–506
- Fromont M (2006) *Droit administrative des États européens*. Thémis droit, Paris
- Garoupa N, Mathews J (2014) Strategic delegation, discretion, and deference: explaining the comparative law of administrative review. *Am J Comp Law* 62:1–33
- Garoupa N, Gili M, Gómez-Pomar F (2012) Political influence and career judges: an empirical analysis of administrative review by the Spanish Supreme Court. *J Empir Leg Stud* 9(4):795–826
- Halberstam D (2010) The promise of comparative administrative law: a constitutional perspective on independent agencies. In: Rose-Ackerman S, Lindseth PL (eds) *Comparative administrative law*. Edward Elgar, Cheltenham/Northampton, pp 185–204
- Mahoney PG (2001) The common law and economic growth: Hayek might be right. *J Leg Stud* XXX:503–525
- Massot J (2010) The powers and duties of the French administrative judge. In: Rose-Ackerman S, Lindseth PL (eds) *Comparative administrative law*. Edward Elgar, Cheltenham/Northampton, pp 415–425
- Nolte G (1994) General principles of German and European administrative law – a comparison in historical perspective. *Mod Law Rev* 57(2):191–212
- Nutting CB (1955) The administrative court. *N Y Univ Law Rev* 30:1384–1389
- Page EC, Robson WA (main contributors) (2014) Administrative law. In: *Encyclopædia Britannica Online*. Retrieved 20 Aug 2014, from <http://www.britannica.com/EBchecked/topic/6108/administrative-law>
- Revesz RL (1990) Specialized courts. *Univ PA Law Review* 138:1111–1174
- Schröder M (2012) Administrative law in Germany. In: Seerden R (ed) *Administrative law of the European Union, its Member States and the United States*, 3rd edn. Intersentia, Antwerpen
- Shapiro M (1981) *Courts, a comparative and political analysis*. The University of Chicago Press, Chicago
- Singh MP (2001) *German administrative law in common law perspective*, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht. Springer, Berlin

Administrative Law

Giulio Napolitano

Law Department, Roma Tre University, Rome, Italy

Abstract

The law and economics literature has traditionally paid very little attention to administrative law history and rules. Economic analysis of law, however, can provide a useful explanation of the logic of administrative law, beyond the purely legal top-down approach. On one side, administrative law provides public bodies with all the needed powers and prerogatives to face and overcome different types of market failures. On the other side, administrative law is a typical regulatory device aiming to face some structural and functional distortions of bureaucracy, as a multi-principal agent. From this economic (and political) point of view, administrative law is much less stable than what it is usually thought to be. Frequent changes in both substantive and procedural rules can be explained as the outcome of repeated interactions among the legislator, the bureaucrats and the private stakeholders.

Definition

A set of rules governing public bodies and their interactions with private parties.

Logic and Ratios of Administrative Law

The law and economics literature has traditionally paid very little attention to administrative law history and rules. Notwithstanding its great expansion also in nonmarket fields, it is still an “unexpected guest” in the public law context (Ulen 2004). Major contributions come from other scientific approaches, like public choice (Farber and Frickey 1991) and political economy of law (McCubbins et al. 2007). They are different from the law and economics movement, of course, but they all share individualistic methodology and the rational interpretation of human behaviors. Unfortunately, on one side, economic and political literature has little confidence with highly technical and detailed provisions of administrative law. On the other side, administrative law scholars represent a close community, usually unwilling to open their minds to methodologies different from legal positivism. That’s why, at least until today, the capacity of law and economics to shed new light into the field of administrative law has been very limited: a missed opportunity both for the law and economics movement and for the traditional scholarship of administrative law (Rose-Ackerman 2007).

According to the still dominant legal scholarship, administrative law is a coherent set of rules, ordered by some general principles, like the rule of law, impartiality, transparency, and proportionality, and characterized by its own specific features, such as the existence of public law entities, the special prerogatives of the Executive and its related branches, the decision-making procedure, and the judicial review. From this perspective, administrative law is represented as a quite stable institution, notwithstanding the frequent change of its specific rules.

Economic analysis, however, can play a very important role in explaining the intimate essence and the distinctive functions of administrative

law. Two ratios explain the logic of administrative law, outside the purely legal top-down approach. On one side, administrative law provides public bodies with all the needed powers and prerogatives to face and overcome different types of market failures. On the other side, administrative law is a typical regulatory device aiming to face some structural and functional distortions of bureaucracy, as a multi-principal agent. From this economic (and political) point of view, administrative law is much less stable than what it is usually thought to be. Frequent changes in both substantive and procedural rules can be explained as the outcome of repeated interactions among the legislator, the bureaucrats, and the private stakeholders (Napolitano 2014).

Administrative Powers as a Mean to Correct Market Failures

The market failures theory offers a powerful explanation of the tasks that modern bureaucracies accomplish in modern societies and of the powers they exercise. Roles and prerogatives of public bodies, of course, are the outcome of complex social and political processes. But the market failures theory shows the existence of a rational basis underlying those historical developments and provides a useful test to verify the persisting need for the public intervention (Barzel 2002).

Many of the so-called sovereign functions of modern governments (like the protection of public security or the defense against external attacks) represent a solution to the problem of public goods. Non-excludability and non-rivalry make the private provision inefficient. Only governments, through the compulsory power of taxing, can solve the free-riding problem, which impedes the proper working of markets.

The public regulation of many economic activities aims to face other market failures (Posner 1997). Entry controls, technical requirements, and pollution standards address the problem of negative spillovers generated especially by industrial and noxious facilities. The regulation of prices and quality standards (in many countries enacted by independent authorities) limits the

market power of network operators in utilities like electronic communications and electricity. Obligations of disclosure in financial markets balance the informational asymmetry between financial institutions, investors, and savers.

Moreover, the government can make compulsory the consumption of merit goods, like education (at least for early stages) and health care (e.g., to prevent contagion, in case of epidemic diseases). The public provision of those services on a larger scale, however, is coherent also with the protection of constitutional rights and the design of the welfare state existing especially in Europe.

In all these cases, administrative law provides public authorities with the prerogative powers that are necessary to make citizens comply with regulations, orders, and duties to pay. This way, public authorities can overcome the high transaction costs that they would otherwise face in order to reach an (often impossible) agreement with private parties. From this point of view, administrative law plays an empowerment function in favor of public bodies, giving them all the prerogatives that are necessary to pursue the public interest. According to the rule of law, those powers must be assigned by the legislator through specific provisions. This is coherent from an economic perspective, because unilateral decisions are not Pareto efficient, reducing the welfare of affected parties. Their sacrifice, then, is justified only in the name of collective preferences aggregated by the legislator, as representative of the community.

Bureaucracy as a Multi-principal Agent

In all the contemporary societies and democracies, public policies and regulations that are needed to solve market failures and to satisfy the collective preferences must be implemented at the administrative level. Bureaucrats do not simply execute the law, like an automatic machine. On the contrary, they exercise a wide discretionary power, making choices among different possible alternatives, all coherent with the law. Delegation from elected bodies to expert agents is a world strategy, which responds to a principle of efficient division of labor (Mashaw 1985). The benefits of

such a strategy, however, can be reduced by the emergence of opportunistic behavior, as it happens in every principal-agent relation (Horn 1995). Many provisions of administrative law pursue exactly the purpose of keeping under control the bureaucratic behavior in the attempt of preventing and punishing drift, capture, and corruption.

More precisely, bureaucracy is a multi-principal agent. At the national level, different political actors compete in order to assume the guidance of public administrations. Even if citizens are supposed to be the original patrons of public policies and of their administrative implementation, they have no direct voice. All the relevant powers are in the hands of elected representatives. But they can embody different political visions, trying to make them prevail in the administrative arena, through specific administrative law devices. This happens particularly in a system of divided government (Mueller 1996). In the USA, the President, especially through the cost-benefit analysis review (Posner 2001; Kagan 2001), and the Congress, enabling “fire alarm” by citizens and through special committees monitoring (McCubbins and Schwartz 1984; Ferejohn and Shipan 1990; Bawn 1997), adopt different legal strategies in order to assume the control over bureaucracy.

Other principals of public administration emerge at supranational level. A growing number of public policies are designed at global or macro-regional level by supranational institutions. They require national administrations to implement them in a coherent way, even though that can create a conflict with elected bodies at state level. Regulation and directives on the administrative implementation and enforcement of common policies represent a fundamental tool in order to achieve compliance against national drifts.

Public administrations also have multiple stakeholders. Individual citizens (and future generations) cannot easily act for the satisfaction and protection of their interests, unless they are specifically struck by an administrative decision. On the contrary, economic actors of different nature and dimension usually organize themselves to influence the exercise of administrative powers,

especially by regulatory agencies. But they compete and fight, one against the other, to obtain *ex ante* from legislators legal tools to play future games before agencies starting from an advantageous position (Macey 1992; Holburn and Vanden Bergh 2006).

All these repeated interactions among competing (and often in opposition) actors explain the existence of different strategies and conflicts in shaping administrative law and in using it as a powerful tool to keep bureaucracy under control.

The Industrial Organization of Bureaucracy

Public administrations, in the exercise of their tasks, produce goods and services in the interest of the entire community. Those goods and services cannot be delivered simply on the basis of a case-by-case bilateral agreement between a public agent and a private recipient. The production of those goods and services, on the contrary, requires the establishment of a stable organization, governed by a set of well-defined rules and codes of conduct.

The institutional design of bureaucracy responds to some relevant regularity in time and space. Everywhere, the Executive is at the center of the stage and exercises, directly or through dedicated units, the core functions of government (the protection of public security, the defense, the administration of justice, the collection of tax). According to a principle of efficient division of labor, agencies and independent authorities accomplish specialized, technical, and regulatory tasks, being separated and somehow insulated from the Executive. State-owned corporations are engaged in business, when considered to be the best way to manage market failures. In federal or decentralized countries, states, regions, and municipalities are autonomous entities and provide directly many goods and services to citizens, being closer and more accountable to them.

The legislator is usually free to change the institutional design of bureaucracy. Name and tasks of ministries can be changed. Agencies, independent authorities, and state-owned

corporations can be established, merged, broken up, and abolished at any time. Tasks and autonomy degree of local governments can be modified or fine-tuned. However, all these operations are limited by path dependence. Legal and bureaucratic structures tend to be stable. Their existence is often protected by different stakeholders (employees, clients, pressure groups). That's why adaptation to changing needs and to institutional reforms may be slow or ineffective.

Bureaucracies work very differently from enterprises. Many of the goods and services they produce have no market value. Competition, which plays a very important role in pushing enterprises toward efficiency, doesn't exert any pressure in raising the quality of public services and in reducing the costs of government. Customer satisfaction, which is fundamental for every enterprise, is not so relevant to top managers in the public sector, whose main concern is to comply with regulations and political instructions. This happens also because public bodies have no market revenues. On the contrary, they are financed through the tax system and the subsequent budgetary decisions, which are assumed by the Parliament and the Executive. Unfortunately, the electoral cycle and instability in government make it difficult for politicians to know in depth the workings of bureaucracy and how to develop long-term cooperative strategies with public sector managers. Finally, bureaucracies must satisfy collective preferences that are dispersed, changing, and sometimes conflicting, (which is different from earning high profit aspirations of companies' stakeholders). That's why it becomes very difficult to assess the bureaucratic performance, while the private sector's managers are kept under control, even if in a very partial and imperfect way, looking at the value of stocks and at the market share (Tirole 1994).

All that explains why bureaucracies suffer from different forms (productive, economic, and industrial) of inefficiency. Many constitutional and administrative law provisions aim to reduce such inefficiency. Political directives are separated from every day management. The legal and economic treatment of civil servants is progressively becoming equal to that of private

employees. Artificial indicators of bureaucratic performance are created in order to keep under control bureaucratic behavior and to introduce performance-related pay systems also in the public sector (Rose-Ackerman 1986).

Strategies of Administrative Action

In order to allow administrative agencies and other public bodies to perform their tasks, administrative law provides them with different legal means of action. From this point of view, especially in the continental European tradition, the main distinction is between unilateral and bilateral means of action. In the first case, the law gives public bodies the power to adopt decisions, which produce binding effect on third parties, even if they weren't consented. Those unilateral decisions are considered an expression of the supremacy of public authorities over the citizen, due to their linkage to the State. In the second case, public bodies conclude contracts on the basis of their general capacity of private law. Given this general distinction, the legal scholarship describes the typology of the most important administrative decisions, according to their proper effect on the recipient, some of them enlarging the private sphere (grants, licensing, other permissions), others restricting it (orders, takings, fines, and administrative sanctions). A similar approach is followed in relation to contracts. Public bodies can make use of different procurement schemes and conclude other kinds of transactions.

The economic analysis of law offers the opportunity to go further. Means of action of public bodies are analyzed not only in a static way but in a dynamic way too, in order to catch the different strategies that rational (at least to a certain extent) public bodies and officials pursue. From this point of view, administrative law empowers the public bodies, leaving them the choice between different legal instruments and the proper design of their content. First of all (and from a very general point of view), in many cases, public bodies evaluate whether it is more convenient to adopt unilateral decisions (e.g., a taking) or to find an agreement with a private party (a sale). Such an

evaluation can be made on the basis of the transaction costs' theory. If it is not too costly to find such an agreement, the consensual solution will satisfy both the public and the private interest, therefore avoiding any case for judicial review. On the contrary, when transaction costs are too high (e.g., because the private party holds a monopolistic position), the unilateral decision might be the most efficient solution.

Economic analysis provides a useful insight also in understanding strategies of acting through law making ("wholesale procedures") or through adjudicatory decisions ("retail procedures") (Cooter 2000, pp. 164–165). It gives precious suggestions about the proper design of licenses and concessions (time, price, contractual conditions). It's helpful in the fixing of the amount of compensation in the case of a taking and in the determination of the optimal level of administrative sanctions. Economic analysis reminds us the importance of the proper design of public contracts. The general interest can be satisfied only if the transaction is convenient for the private party too; otherwise the procurement procedure will fail or the private contractor will be unable to execute the contract.

Finally, game theory provides a useful insight in the evaluation of the interactions between public bodies and private parties, explaining success and failure of public policies (as in the case of the so-called simplification of the administrative procedures of licensing, often vanished by the absence of trust between agencies and private parties (Von Wangenheim 2004)).

The Regulation of Administrative Action

Administrative action is characterized by two distinctive features. On one side, public bodies, in executing the law, act as agents of different principals. On the other side, public bodies exercise a legal and economic power against private parties. The regulation of administrative action, then, pursues two purposes. The first is to ensure the fulfillment of the public interest, in coherence with the guidelines issued by the principals. The second is to protect the firms and the citizens

interacting with the public bodies in a subordinate position. From this double perspective, it becomes possible to understand the whole set of the legal provisions related to the administrative procedure, the exercise of discretionary powers, and contractual activity.

Many countries have enacted a legislation to regulate the administrative procedure in general (the USA in 1946 with the Administrative Procedure Act, Germany in 1976, Italy in 1990). In other countries, like France and the UK, relevant rules can be drawn from the principles stated by the courts and from sector by sector legal provisions. The regulation of administrative procedures, on one side, structures the decision-making process in the public interest. On the other side, it recognizes the rights of participation in favor of affected parties. Such participation is not only in the private interest. It's also a powerful tool of decentered "fire-alarm" control, in order to reduce informational asymmetry of political principals and to prevent bureaucratic drift (McCubbins et al. 1987).

The limitation of administrative discretion is very important too in order to avoid decisions contrary to the public interest or unduly negative for the private recipients. Legal provisions select the protected interests, fix the prerogatives of the acting bodies, and determine the criteria which must be followed in the exercise of power. However, a certain degree of flexibility is necessary to allow the administrative agencies to manage the specificity of every case and to face unpredictable circumstances (Cooter 2000, p. 91).

A detailed regulation is needed also when public bodies conclude contracts. On one side, the waste of public money must be prevented. On the other side, there is the risk of discrimination in the adjudication of contracts. That's why, international, European, and national regulations require fair and open tendering procedures to select the best offer. From this point of view, the way in which public bodies select their contractual partners is very similar to a marriage by mail. This explains the importance of a formal and detailed contract in which all future circumstances are covered by an *ex ante* agreement.

The efficiency of the rules devoted to the regulation of administrative action, anyhow, must be put under scrutiny. There are good reasons which explain the need for those rules and their growth in recent times. However, those regulations don't always prove to be really useful or to ever serve their intended purpose upon enactment. The final outcome is often an over-regulation of administrative action, which makes it excessively rigid and unfit to meet the quest for flexibility required to satisfy the public interest.

The Judicial Review and the Non-expert Control Paradox

The regulatory system of administrative law in every country is completed by the existence of a judicial review against the decisions adopted by public bodies. Private parties, which are affected by an administrative decision, can go before a court to obtain the declaration of invalidity of that decision, if retained as illegitimate. In this way, while protecting their private interests, they launch "fire-alarm" signals to allow a third party independent oversight of the bureaucratic behavior.

Courts play a very important role in controlling the respect of the legal provisions that limit administrative discretion. They also check the exercise of effective investigations on facts and interests and the disclosure of all the information at the basis of the administrative decision. Finally, they enforce the respect of all the procedural requirements that allow the participation and the external control of administrative action by private parties (Bishop 1990).

Different models of judicial review and of control on the proper behavior of bureaucracy exist in the world (Josselin and Marciano 2005). Common law countries give to ordinary courts the competence to review administrative decisions. Many continental European countries, on the contrary, established special administrative courts, following the French model of the *Conseil d'Etat*. Administrative courts are supposed to be more sensible to the public interest and more generous

in recognizing and protecting power and prerogatives of the bureaucracy. However, differences among ordinary and administrative courts are greatly reduced over time. Special benches of ordinary courts were established, e.g., in the UK, in order to address administrative law cases. Administrative courts, at the same time, became much more effective in protecting private interests.

The judicial review of administrative action by courts can appear a paradox. Judges, who have only a legal training, are asked to check the decisions adopted by expert bureaucrats. This way, the system of judicial review runs the risk of frustrating the very reason of delegation to public administration, which is made exactly in name of an efficient division of labor to highly specialized and expert bodies. The existence of specific administrative courts, different from ordinary courts, however, reduces the danger of scarce preparation. By focusing their attention on disputes between agencies and citizens, they can develop a specific knowledge of the different fields of administrative action. Moreover, administrative judges, especially in France and Italy, act as consultant of public bodies and are often appointed as heads of staff in the Cabinet. This way they can reach a deeper comprehension of the way in which bureaucracies work.

In any case, the judicial review can ameliorate the quality of administrative decisions for at least three reasons. First, procedural rules before court would more likely guarantee the due process than administrative procedure rules in isolation would do. This way, courts may have access to a greater deal of information than that available to bureaucracy at the moment of the administrative decision. Second, due to self-selection made by private parties, the judicial review can be focused on really suspicious cases rather than having to check every administrative decision. And third, public agencies have an incentive to make *ex ante* investments in the preparation of the administrative decisions, through a more detailed evaluation of facts and interests, and in legal protection, in order to prevent the judicial review made by courts (Von Wangenheim 2005).

References

- Barzel Y (2002) A theory of the state. Economic rights, legal rights, and the scope of the state. Cambridge University Press, Cambridge
- Bawn K (1997) Choosing strategies to control the bureaucracy: statutory constraints, oversight, and the committee system. *J Law Econ Organ* 13(1):101–126
- Bishop W (1990) A theory of administrative law. *J Legal Stud* XIX(2):489–530
- Cooter RD (2000) The strategic constitution. Princeton University Press, Princeton
- Farber DA, Frickey PP (1991) Law and public choice. A critical introduction. Chicago University Press, Chicago
- Ferejohn J, Shipan C (1990) Congressional influence on bureaucracy. *J Law Econ Organ* 6:1–20
- Holburn GLF, Vanden Bergh R (2006) Consumer capture of regulatory institutions: the creation of public utility consumer advocates in the United States. *Public Choice* 126:45–73
- Horn MJ (1995) The political economy of public administration. Cambridge University Press, Cambridge
- Josselin JM, Marciano A (2005) Administrative law and economics. In: Backhaus JG (ed) *The Elgar companion to law and economics*, 11th edn. Edward Elgar, Cheltenham, pp 239–245
- Kagan E (2001) Presidential administration. *Harv Law Rev* 114:2245–2385
- Macey J (1992) Organizational design and the political control of administrative agencies. *J Law Econ Organ* 8(1):93–110
- Mashaw JL (1985) Prodelegation: why administrators should make political decisions. *J Law Econ Organ* 1(1):81–100
- McCubbins M, Schwartz T (1984) Congressional oversight overlooked: police patrols versus fire alarms. *Am J Polit Sci* 28(1):165–179
- McCubbins M, Noll R, Weingast B (1987) Administrative procedures as instruments of political control. *J Law Econ Organ* 3(2):243–277
- McCubbins M, Noll R, Weingast B (2007) The political economy of law. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*, 11th edn. Elsevier, Amsterdam, pp 1651–1738
- Mueller DC (1996) Constitutional democracy. Oxford University Press, New York/Oxford
- Napolitano G (2014) Conflicts and strategies in administrative law. *Int J Const Law* 12(2):357–369
- Posner RA (1997) The rise and fall of administrative law. *Chicago-Kent Law Rev* 72:953–963
- Posner EA (2001) Controlling agencies with cost-benefit analysis: a positive political theory perspective. *Univ Chicago Law Rev* 68(4):1137–1199
- Rose-Ackerman S (1986) Reforming public bureaucracy through economic incentives? *J Law Econ Organ* 2:131–161
- Rose-Ackerman S (2007) Introduction. In: Ead (a cura di) *The economics of administrative law*. Edward Elgar, Cheltenham, pp XIII–XXVIII

- Tirole J (1994) The internal organization of government. *Oxf Econ Pap* 46(1):1–29
- Ulen TS (2004) The unexpected guest: law and economics, law and other cognate disciplines, and the future of legal scholarship. *Chicago-Kent Law Rev* 79(2):403–429
- von Wangenheim G (2004) Games and public administration. The law and economics of regulation and licensing. Edward Elgar, Cheltenham
- von Wangenheim G (2005) Should non-expert courts control expert administrations? In: Josselin J-M, e Marciano A (eds) *Law and the state. A political economy approach*. Elgar, Cheltenham/Northampton, pp 310–332

Administrative Procedure

Mitja Kovac

Department of Economic Theory and Policy,
Faculty of Economics, University of Ljubljana,
Ljubljana, Slovenia

Abstract

This chapter examines the economics of administrative procedure and shows how law and economics insights can be used to understand fundamental features of administrative procedure. Moreover, it offers a survey of law and economics literature on administrative procedure and addresses three general aspects of administrative procedure. The first is the administrative decisions and access to information. The second general aspect is administrative procedure and agency capture problem. The third general aspect addressed in this chapter is the issue of administrative procedure and international trade.

Definition

Administrative procedure is a set or system of rules that govern the procedures of public bodies for managing organizing bureaucratic actions inside bureaucracies and in their interactions with private parties. These procedures are generally meant to establish efficiency, consistency, responsibility, and accountability.

Introduction

In all industrialized societies, there is a tension between market and collectivist system of economics organization. The latter one is the system where state seeks to direct or encourage behavior which would not occur without such an intervention. The aim of such an intervention is to correct perceived or real market failures in meeting collective or public interest goals. Such goals are generally pursued by a plethora of executive and legislative branches of governments. In this respect, modern law and economics analysis of public law divides into two strings of work. One dealing with causes and consequences of bureaucratic action inside bureaucracies and the other focusing on external interaction, such as between legislators and the bureaucratic institutions comprising executive branch, between the latter institutions and the citizens and enterprises (Weigel 2006). The latter field of research is in law and economics known as the issue of regulation (Ogus 2004). Ogus (2004) actually argues that civil legal cultures have specific concepts and instruments to govern those “external relationships,” for example, German law operates with “*Wirtschaftsverwaltungsrecht*” (literally: “Administrative Law of the Economy”) and the French with “*Droit Public Economique*,” whereas English law employs imprecise expressions, such as “regulation” and “regulatory law.” Although perplexing problems and issues of regulation have triggered some attention of law and economics scholars (Weigel 2003; Ogus 2004; Rose-Ackerman 1995; Schuck 1994; Johnson and Libcap 1994; Rose-Ackerman and Lindseth 2010; Lewish and Parker 2017) and despite its great expansion to nonmarket fields, the proper interpretation of administrative procedure has received relatively little law and economics analysis in recent years. One may even argue that law and economics are still an unexpected guest in the field of administrative law (Ulen 2004; Napolitano 2014). Napolitano (2014) shows that indeed the law and economics literature has traditionally paid very little attention to administrative law, history, and rules. While civil and criminal procedures have been thoroughly addressed and produced a striking amount of

literature, the assessment of public/administrative law actually reveals surprising gaps in the law and economics literature of administrative procedure.

Moreover, the law and economics literature has generally paid very little interest toward procedural issues, and the bulk of work done in that field contains hardly any economics (Schuck 1994). Weigel (2006), for example, stresses that general administrative procedure acts (which play central roles in civil law systems) have actually rarely been economically thoroughly addressed or even investigated whether those essential systems of procedural rules are efficient tools for the promotion of effective, wealth-maximizing outcomes.

Identified lack might indeed be contributed to the limited confidence that the law and economics scholars might have with highly technical and vastly detailed provisions of administrative procedure. Furthermore, Rose-Ackerman (2007) points out that administrative law scholars represent a close community, usually unwilling to open their minds to methodologies different from legal positivism. This isolation might also explain why the capacity of law and economics to shed new light into the field of administrative procedure has been very limited. Hence, this lack of literature represents a missed opportunity both for the law and economics and for the traditional legal scholarship of administrative procedural law (Ogus 1998; Rose-Ackerman 2007).

However, the limited amount of literature might be contrasted with a Bishop's (1990) theory that the essential function of administrative law and of administrative procedure is thought to come to grips with economic slack created by bureaucrats. Moreover, there is, for example, a vast law and economics literature on production of legal rules by agencies and bureaucracies (von Wangenheim 2004).

Administrative Decisions and Access to Information

Classical law and economics literature conceives legal procedure as an institution designed to minimize the sum of "error costs" (the social costs

generated when a judicial system fails to carry out the allocative or other social functions assigned to it) and "direct costs" (such as lawyers', judges', and litigants' time) of operating legal dispute resolution machinery (Posner 1973). According to this classical law and economics framework, the rules and other features of any procedural system can be analyzed as efforts to maximize efficiency (Posner 1973). In line with this observation, Posner (1973) actually already, as part of his case studies, addresses the deterrence issues of administrative sanctions, argues against any judicial review of agency fact-finding, and suggests that the characteristic combination of prosecution and adjudication in the same agency may be a source of inefficiency. Stewart (1975) supports this information disclosure function by arguing that administrative decision should be made transparent through requirements for public comment and open meetings and that administrative procedures should give citizens and organized interests the ability to represent their views in the administrative process. This proposition is in line with Stigler's (1971) path-breaking insights that open procedures are thought to foster pluralist politics that protect against regulatory capture, the danger that an industry will come to control an agency's decision-making to secure private benefits.

Posner (2011) also argues that administrative procedure will be less consistent over time than judicial one and also stresses that precedents will play a smaller role in administrative than in judicial decision-making. He shows that a heralded innovation of the administrative procedure in the USA was the administrative agency's looseness structure where agency is able to issue rules, bring cases, decide cases, conduct studies, issue advice, and even propose a legislation (Posner 2011).

However, one of the first field-specific law and economics contributions in the field of administrative procedure was Rose-Ackerman's seminal article on "The Progressive Law and Economics- And the New Administrative" on the lowering information cost function of modern administrative procedural laws. Rose-Ackerman (1988) suggests that if administrative procedures lower the cost of access to information about administrative

decisions, then interested parties will be more likely to be able to find out about an upcoming decision. The more time an interested party has to mobilize financial and political resources against a particular decision, the more likely it will be able to force the administrative decision-maker, through his or her more politically accountable superiors, to influence that decision. The impact on substantive policy comes from the greater benefit to less organized interest groups of lowering this cost of access to information (Rose-Ackerman 1988).

Cooter (2002) employs game theoretical tools and shows, while assessing the US Administrative Procedure Act discusses the court-agency relationships. He provides evidence that the relatively high transaction costs of formal decisions increase the agency's discretionary power over a class of decisions, whereas the relatively low transaction costs of informal decisions reduce the agency's discretionary power (Cooter 2002). Furthermore the court will have less concern over the agency's discretionary power when agency preferences are close to court preferences. Thus, Cooter's model predicts the courts will favor formal decision-making when agency preferences are close to court preferences (Cooter 2002).

In addition, Verkuil (1978) analyzed the emerging concept of the US administrative procedure and argued that administrative procedure should be concerned with the overall fairness and accuracy of decisions, with their efficient and low-cost resolution and, in a democratic society, with participant satisfaction with the process.

Administrative Procedure and Agency Capture Problem

One of the central problems of the representative democracy is how to ensure that decisions are responsive to the interests or preferences of citizens. In the 1980s scholars began to utilize law and economics insights to study public law and examined how rules and administrative procedures shape policy and evaluated these outcomes from the perspective of economics efficiency (McNollgast 2007). This stream of

research conceptualizes the decision-making procedures of government as rationally designed by elected official to shape policies arising from decisions by executive agencies and to align incentives of agencies with the ones of citizens. McCubbins et al. (1987, 1989) study argue that administrative procedure is actually designed to solve the agency problem between citizens and agencies and that the mechanics of fire-alarm oversight are imbedded in the administrative procedures of agencies that are within the jurisdiction of the US Administrative Procedure Act. They actually see administrative procedures as another mechanism to control bureaucracy as kind of a check-and-balance system that prevents the opportunism and moral hazard of bureaucratic system and its institutions. For example, all of the procedural features of the US Administrative Procedure Act economically speaking actually reduce agency's information advantage (McNollgast 2007). McNollgast also points out that congress employs administrative procedure to delegate some monitoring responsibility to those who have standing before an agency and who have a sufficient stake in its decisions to participate in its decision-making process and to trigger oversight by pulling the fire alarm (McNollgast 2007). In addition, administrative procedures create a basis for judicial review that can restore the status quo without legislative correction, and as a result administrative procedures may cope with the first-mover advantage of agencies (McNollgast 2007). He also states that administrative procedures might facilitate the political control of agencies in five different ways:

- (a) administrative procedures ensure that agencies cannot secretly conspire against elected official;
- (b) agencies must solicit valuable information;
- (c) the administrative proceedings are public, thereby enabling political principals to identify not just potential winners or losers of the policy but also their views;
- (d) the sequence of procedure (notice, comment, deliberation, collections of evidence, and construction of records) enables elected representatives to respond when administrative

agencies seek to move toward a goal that is not aligned with citizens ones; and

- (e) it serves to indicate the stakes of a group in an administrative proceeding (McNollgast 2007).

Moreover, McCubbins et al. (1987) argue that procedural requirements also affect the institutional environment in which agencies make decisions and thereby limit an agency's range of feasible policy actions. In recognition of this, elected officials may design procedures to solve two prototypical problems of political control. First, procedures can be used to mitigate informational disadvantages faced by politicians in dealing with agencies (McCubbins et al. 1987). Second, procedures can be used to enfranchise important constituents in agency decision-making processes, thereby assuring that agencies are responsive to their interests (McCubbins et al. 1987). The most subtle and the most interesting aspect of procedural controls is according to McCubbins et al. (1987) the possibility to assure administrative agency's compliance without specifying, or even necessarily knowing, what substantive outcome is most in their interest. Namely, by controlling processes, political leaders assign relative degrees of importance to the constituents whose interests are at stake in an administrative proceeding and thereby channel an agency's decisions toward the substantive outcomes that are in the interest of those who are intended to be benefited by the policy. Hence, political leaders can be responsive to their constituencies without knowing the details of the policy outcomes that these constituents want (McCubbins et al. 1987, 1989). Rose-Ackerman (2007) supports this "fire-arm" proposition and adds that even though the constraints are nominally procedural, they have also substantive effects. Namely, instead of requiring close legislative control, the required procedures assure the enacting coalition that the agency will respond to changes in the preferences of those interests served by the legislation (Rose-Ackerman 2007).

Contradicting McCubbins et al. (1987), Moe argues that the role administrative procedures play must be understood in a dynamic context

(Moe 1989). According to Moe's argument, these procedures create a "lock-in" effect – constraining both future politicians and their agents – and therefore act as a mechanism whereby current majorities can ensure, at a cost, that future majorities will not upset their policy intentions (Moe 1989).

Recently, De Figueiredo and Vanden Bergh (2004) provide the first empirical investigation of previously discussed theories. They empirically, by employing an administrative procedures enactment dataset for the period 1941–1981, test whether organizational procedures that are enacted by public officials have any impact on the nature of both bureaucratic control and performance. They have also examined the set of conditions under which the US federal states adopt administrative procedures and found that the likelihood of adopting administrative procedural acts are increased when (a) democratic legislative supermajorities (veto-proof majority) face a Republican governor and (b) when democratic control is perceived to be temporary and when they fear the future loss of power (De Figueiredo and Vanden Bergh 2004). Their results indicate that existing theories emphasizing agency and dynamic effects are empirically valid, albeit with an important qualification: there is a distinctive partisan bias in the usefulness of administrative procedures for these purposes (De Figueiredo and Vanden Bergh 2004). De Figueiredo and Vanden Bergh (2004) also show that despite a dual set of conditions under which they might be adopted, administrative procedure acts seem to have a markedly Democratic bias.

Administrative Procedure and International Trade

Another stream of law and economics research focuses on the importance and on the impact of administrative procedural law on trade policy. Rosenbaum (1998), for example, argues that administrative procedures that lower the cost of obtaining information about and notice of agency decision-making and that therefore increase the amount of time between an interest group's

learning of a potential decision and the final administrative decision shift the balance of power between more and less organized groups. Namely, by giving less organized interest groups a greater opportunity to mobilize political resources to influence bureaucrats through the politicians who oversee them, administrative procedures might shift the balance of power between constituencies (Rosenbaum 1998). While more organized groups may have the political resources, connections, money, and staff – to learn of agency decision-making without procedural constraints – less organized groups will be able to benefit from the increased possibility of notice and opportunity to mobilize. Therefore, by lowering the cost of access to information about administrative decisions, and by thus shifting the balance of power among constituencies from more to less organized groups, administrative procedural law gives according to Rosenbaum (1998) consumers greater power in determining trade policy. Hence, since consumers of a given good are interested in lower trade barriers while producers are interested in greater protection, this shift of power will lower trade barriers (Rosenbaum 1998).

Concluding Remarks

While civil and criminal procedures have been thoroughly addressed and produced a vast amount of law and economics literature, the assessment of public/administrative law actually reveals surprising gaps. Identified gap in the law and economics of administrative procedure calls for further empirical, behavioral, experimental, and theoretical research that would shed additional light upon the most triggering questions and would provide additional insights, normative policy implications for legislators in improved daily decision-making and lawmaking. Coglianese (2002) actually suggests that empirical research on administrative law has the potential for evaluating and ultimately improving prescriptive efforts to design administrative procedures in ways that contribute to more effective and legitimate governance. Hence, identified missed opportunity for the law and economics scholarship of administrative procedural law

opens doors for an extensive law and economics treatment that would offer sets of legal and economics arguments for an improved regulatory response.

References

- Bishop W (1990) A theory of administrative law. *J Leg Stud* 19:489–530
- Coglianese C (2002) Empirical analysis and administrative law. *University of Illinois Law Review*. University of Illinois, Urbana Champaign, pp 1111–1137
- Cooter R (2002) *The strategic constitution*. Princeton University Press, Princeton
- De Figueiredo RJP Jr, Vanden Bergh RG (2004) The political economy of state-level administrative procedure acts. 47. *J Law Econ* 2:569–588
- Johnson RN, Libcap G (1994) *The federal civil service system and the problem of bureaucracy*. University of Chicago Press, Chicago
- Lewish Parker JS (forthcoming 2017) Procedure in American and European law: a general economic analysis. In: Hatzis AN (ed) *Economic analysis of law: a European perspective*. Edward Elgar, Cheltenham
- McCubbins MD, Noll RG, Weingast BR (1987) Administrative procedures as instruments of political control. *J Law Econ Org* 3:243–277
- McCubbins MD, Noll RG, Weingast BR (1989) Structure and process, politics and policy: administrative arrangements and the political control of agencies. *Virginia Law Rev* 75:431–482
- McNollgast (2007) The political economy of law. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*. North-Holland, Amsterdam, pp 1707–1715
- Moe TM (1989) The politics of bureaucratic structure. In: Chubb JE, Peterson PE (eds) *Can the Government Govern?* Brookings Institution Press, Washington, DC, pp 267–330
- Napolitano G (2014) Administrative law. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York, pp 1–8
- Ogus AI (1998) Regulatory appraisal: a neglected opportunity for law and economics. *Eur J Law Econ* 6:53–68
- Ogus AI (2004) *Regulation: legal form and economic theory*. Hart Publishing, Oxford
- Posner RA (1973) An economic approach to legal procedure and judicial administration. *J Leg Stud* 2:399–458
- Posner AR (2011) *Economic analysis of law*. Kluwer, Austin
- Rose-Ackerman S (1988) The progressive law and economics-and the new administrative law. 98. *Yale Law J* 98:341–368
- Rose-Ackerman S (1995) *Controlling environmental policy*. Yale University Press, New Haven
- Rose-Ackerman (2007) Public choice, public law and public policy. Keynote address, First World Meeting of the Public Choice Society Amsterdam, Mimeo

- Rose-Ackerman S, Lindseth PL (2010) Comparative administrative law. Edward Elgar, Cheltenham
- Rosenbaum RD (1998) Domestic bureaucracies and the international trade regime: the law and economics of administrative law and administratively imposed trade barriers. Harvard Law School Discussion Paper.
- Schuck PH (1994) Foundations of administrative law. Foundation Press, New York
- Stewart RB (1975) The reformation of administrative law. *Harv Law Rev* 88:1669–1813
- Stigler GJ (1971) The theory of economic regulation. *Bell J Econ Mgt Sci* 2:3–21
- Ulen TS (2004) The unexpected guest: law and economics, law and other cognate disciplines, and the future of legal scholarship. *Chicago-Kent Law Rev* 79(2):403–429
- Verkuil PR (1978) The emerging concept of administrative procedure. *Columbia Law Rev* 78(2):258–329
- Wangenheim GV (2004) Procedural rules and administrative decisions. In: *Encyclopedia of law and economics*. Edward Elgar, Cheltenham
- Weigel W (2003) *Rechtsökonomik*. Vahlen Verlag, München
- Weigel W (2006) Why promote the economic analysis of public law? *Homo Econ* 23(2):195–216

Adoption

Martha M. Ertman

Francis King Carey School of Law, University of Maryland, Baltimore, MD, USA

Abstract

Historical and recent developments in legal economic analysis of adoption the United States reveal changing supply and demand of children and the emergence of submarkets for adoption through agencies (public or private) and independent adoptions. Current legal rules against baby-selling and adoption agency practices mask the existence of adoption markets by banning payments to birth parents yet exempting payments to agencies and other adoption professionals. Economic proposals seek to narrow gaps between supply and demand by creating incentives (or removing disincentives) through substitutes, subsidies, and reduced transactions costs. These solutions could prevent a good number of would-be parents from remaining childless, provide homes

for many children who currently languish in foster care or group homes, and better recognize the costs that adoption imposes on birth parents.

Synonyms

[Children](#); [Family](#); [Parenthood](#)

Introduction

This entry tracks historical and recent developments in legal economic analysis of adoption. Focusing on the United States, it first traces historical market changes as evidenced by changing supply and demand of children and the emergence of submarkets for adoption through agencies (public or private) and independent adoptions. It then identifies current legal rules against baby-selling and adoption agency practices that mask the existence of adoption markets by banning payments to birth parents yet exempting from that ban payments to agencies and other adoption professionals. It concludes by describing economic proposals to narrow current gaps between supply and demand by creating incentives (or removing disincentives) through substitutes, subsidies, and reduced transactions costs. These solutions, if implemented, could prevent a good number of would-be parents from remaining childless, provide homes for many children who currently languish in foster care or group homes, and better recognize the costs that adoption imposes on birth parents.

Adoption Markets

History

The laws of supply and demand have long governed adoption. In the nineteenth century, norms against unwed motherhood and the lack of a social safety net made adoption a buyer's market. The "price" of children whose parents could not raise them was so low that many birth mothers had to pay people known as "baby

farmers” to take the children off their hands. In the custody of the baby farmers, that low value translated to low survival rates. Children lucky enough to survive were put to work, getting hired out for farm or domestic work. Gradually, adoptive parents became willing to pay to adopt. In the late nineteenth and early twentieth century, newspaper classified advertisements commonly listed children for adoption along with a price (Zelizer 1985; Herman 2008).

Gradually, adoption law and practice came to reject market rhetoric as society and legal rules began to view childhood as a vulnerable period spent in school instead of laboring in fields or factories. In 1851, Massachusetts passed the first adoption statute, which required that a child’s interests be accounted for in an adoption. Other states followed suit, eventually establishing standards for determining a child’s adoptability and adoptive parents’ suitability. Gradually, states completely replaced the birth family with the adoptive family courtesy of new birth certificates and sealed records. Along the way, agencies and social workers acquired a nearly exclusive gate-keeping function over both the adoption and information passing between birth and adoptive families. Limits on child labor and increased life spans due to industrialization and medical advances all contributed to these social and economic changes that justified higher investments in children’s human capital. Technologically, the advent of safe, affordable infant formula in the 1920s and 1930s further encouraged couples to adopt. By 1937, a magazine article could exclaim, “The baby market is booming . . . the clamor is for babies, more babies . . . We behold an amazing phenomenon: a country-wide scramble on the part of childless couples to adopt a child” (Zelizer 1985).

Adoption rates continued to rise until the early 1970s, when several phenomena contributed to both a rapid drop in the supply of infants available for adoption and increase in demand from new sources. On the supply side, more unmarried mothers could collect child support from the children’s fathers once the Supreme Court struck down as unconstitutional rules that relieved non-marital fathers of financial responsibilities.

Poverty was less likely to force a woman to relinquish her child for adoption as federal welfare programs expanded to cover more poor, often unmarried, women and their children. Rises in nonmarital births and single parenthood after divorce decreased the stigma of single motherhood.

Some changes in legal rules and cultural practices influenced both supply and demand. As women could get effective birth control and legal abortions, they became less likely to conceive or carry an unplanned pregnancy to term. That increased control made motherhood more of a choice than it had ever been, allowing some women to take full advantage of newly expanded opportunities to develop their human capital by pursuing a career. On the supply side, that increased access to well-remunerated work meant that more single women who did bear children could keep them. On the demand side, many women who had delayed conception while they established careers found their fertility diminished when they got around to having children.

Consequently, in 1972, nearly 20% of white single mothers placed their babies in adoptive homes, but by 1995, that rate decreased to 1.5% and has since dipped to about 1%. Black mothers, in contrast, have never relinquished in high numbers. Even before cases like *Griswold v. Connecticut* and *Roe v. Wade* expanded reproductive choices by decriminalizing birth control and abortion, only about two percent of black single mothers surrendered their infants for adoption. That low rate was due to many agencies’ unwillingness to place African-American babies and the willingness of extended kinship networks in the black community to take in those children. Since the 1970s black women’s relinquishment rate has dipped to nearly zero. By the late 1980s, this low supply of healthy infants translated to a hundred would-be adoptive parents competing to adopt each healthy infant available for adoption, according to the National Council for Adoption (Joyce 2013).

Agencies and adoptive parents continue to express preferences for some children over others. Prior to World War II, most agencies considered children of color and those with disabilities unfit

for adoption, relegating them to institutions. While more agencies began to accept these children in the 1950s – increasing the supply of adoptable infants – many would-be adoptive parents retained their preference for white children. In the early twenty-first century, most would-be adoptive parents – a majority of whom are white – continue to prefer to adopt white over African-American children, though adoptive parents who are single or gay are more willing to create an interracial family through adoption (Baccara and Collard-Wexler 2012).

Adoption Markets Today

Today, adoption continues to function as a market, though rhetoric of both law and adoption professionals persists in masking its market characteristics. About 2.5% of American children are adopted, 1.6 million in 2000. Adoption professionals – agencies, attorneys, and social workers – generate fees of \$2–3 billion annually (Baccara 2010). Agencies refer to adoption as a “gift” instead of an exchange, and statutes criminalize or otherwise prohibit payment of fees in connection with an adoption. But those statutes also exempt fees paid to professionals, and states enforce the statutes largely to penalize birth parents receiving any payment. Some states permit adoptive parents to pay for a birth mother’s maternity clothes, psychotherapy, and living expenses, while others prohibit those expenditures or impose strict time or monetary limits. A black or gray market persists in the shadow of these prohibitions, with adoptive parents or agencies paying for birth mothers’ expenses (Ertman 2003, 2015).

The legal adoption market is fragmented. A child could be placed through a private agency, through a public child welfare agency (“foster care”), or without agency involvement (“independent adoption”). An adoption may also be subject to rules imposed by a religiously affiliated agency, may involve relatives or nonrelatives, and could be domestic or international. While accurate data on adoption is scarce and rates vary over time, as of 2002, about 84% were domestic – evenly divided among familial and nonfamilial – with 16% international (which are nearly always non-familial). That same year, over half of the children

adopted by nonrelatives were adopted out of foster care. Many of the children adopted out of foster care are older; suffer from physical, mental, and emotional disabilities; and/or are children of color (Bernal et al. 2007; Moriguchi 2012). This entry focuses on three categories that exhibit different market patterns: unrelated domestic adoption, adoption out of foster care, and independent adoption.

Within the first category – unrelated domestic adoptions – demand for healthy white infants far outstrips supply (Landes and Posner 1978). Consequently, adoptive parents commonly pay agencies higher fees to adopt those babies than to adopt children whom adoption professionals describe as “hard to place.” (Ertman 2015). Another common fee structure charges adoptive parents on a sliding scale, with higher-income adopters paying higher fees. Wealth and income gaps between whites and African-Americans, coupled with most adoptive parents’ preference for a child of their own race, combine to make it more expensive in general to adopt a healthy white infant than to adopt children who are older, of color, and/or disabled.

Adoptions in the second category – foster care – exhibit the opposite pattern: high supply of children and low demand by adoptive parents, generating what Elisabeth Landes and Richard Posner called a “glut” of children in foster care, a situation that they compared to “unsold inventory stored in a warehouse” (Landes and Posner 1978). Consequently, children in foster care may wait 2 years for adoption and get shuttled between 10 and 20 homes during those years. Only around ten percent – 50,000 – get adopted into permanent homes each year, and African-American children can wait twice as long as white kids to get adopted out of foster care (Beam 2013).

In the third category – independent adoption – comprehensive data is even more scarce than in agency adoptions. Many independent adoptions are stepparent adoptions – as when a divorced woman’s new spouse adopts a child from the prior marriage – which do not require an intermediary to match the children to their new families nor a study of suitability of adoptive families. The fees are largely for lawyers, rather than agencies, and do not depend on the

characteristics of either parent or child. Some monetary payments are not legally binding, but nevertheless occur with regularity. A noncustodial birth father often demands a “price” for consenting to the adoption in the form of the mother agreeing not to pursue him for unpaid child support (Hollinger 2004). Another common payment – this one entirely nonmonetary – involves the birth parent retaining visitation rights in an agreement known as a Post Adoption Contact Agreement. These agreements are legally binding in about half the states (Ertman 2015).

Proposals to Narrow Gaps Between Supply and Demand

The market for adoption evidences inefficiencies. At least since the 1970s, supply and demand for two major types of adoption have been mismatched, with high demand coupled with low supply of healthy white infants and low demand coupled with high supply of children in foster care. Consequently, adoptive parents who are only willing to adopt a healthy, white infant may well remain childless, while as many as 125,000 adoptable children languish in foster care. At the fiscal level, adoption is less expensive than foster care. Between 1983 and 1986, adoption decreased government expenditures on foster care by \$1.6 billion. Non-fiscal but nevertheless valuable benefits of adoption over foster care include the better health, behavioral, educational, and employment outcomes of children who are adopted over those who remain in foster care (Hansen 2008).

Foundational Work by Becker, Landes, and Posner

Economic proposals have sought to increase the supply for in-demand babies and increase the demand for hard-to-place children. Nobel Laureate Gary Becker’s 1981 book *A Treatise on the Family* established the relevance of economic tools to understand family forms. On the supply side of adoption, he suggested that birth parents are more likely to put what he calls “inferior” children “up for sale or adoption.” On the demand

side, he postulated a “taste for own children, which is no less (and no more) profound than postulating a taste for good foods or any other commodity entering utility function” and that women are reluctant to commit “effort, emotion and risk” to children “without considerable control over rearing” as well as information ahead of time about the children’s “intrinsic characteristics” (Becker 1981). Legal economists Elisabeth Landes and Richard Posner further developed the theoretical model of adoption as a market in their highly influential 1978 article “The Economics of the Baby Shortage.” They identified the “potential gains in trade from transferring the custody of the child to a new set of parents” and cataloged the pros and cons of a “free baby market” in comparison to a likely black market that flourishes in part because of legal constraints on payments in adoption. Their proposal was quite modest. As “tentative and reversible steps toward a free baby market,” they proposed that agencies use fees charged to higher-income adoptive parents to make “side payments” to pregnant women to induce them to relinquish the child for adoption instead of terminating the pregnancy (Landes and Posner 1978). Many readers mistook this proposal for an open market in children – akin to slavery – and criticized its tendency to treat children as commodities (Radin 1996; Williams 1995), despite Posner’s clarification about the proposal’s narrowness (Posner 1987).

Recent Applications

Recent market proposals further develop the idea of openly marketizing adoption.

Increase Supply of In-Demand Infants

Legal reforms aimed to increase the supply of the most sought-after babies who are available for adoption would focus on birth parents – especially birth mothers – because they are the ones who generally make the decision whether to keep a child or place him or her for adoption. Legal economists have argued that legal rules and adoption professionals could make birth parents’ substitutes for adoption more expensive, use subsidies to encourage relinquishment, and lower transactions costs.

Make Birth Parents' Substitutes More Expensive Despite the apparent causal relationship between legal decisions granting women rights to birth control and abortion and plummeting number of infants available for adoption, empirical research has yielded mixed results regarding the effect of reproductive freedoms on adoption. One study found that states that legalized abortion before *Roe v. Wade* also had a 34–37% decrease in adoption of unrelated white children (Bitler and Zavodny 2002). However, other studies either found no statistically significant relationship between adoption and the price and availability of abortion (Medoff 1993) or a causal relationship between adoption and abortion availability but also found an unexpected effect of restrictive abortion laws decreasing the number of unwanted births (Gentian 1999).

One proposal would impose barriers to abortion to increase the supply of healthy, white infants by making adoption a “two-sided market clearing institution” that remedies information asymmetries. It would increase the supply of sought-after babies through two legal changes: (1) allowing payments to birth mothers and (2) requiring women considering an abortion to hear a pitch from an adoption provider about the “economic incentives” should the birth mother “produce a child for the market” (Balding 2010).

Subsidize by Decreasing Birth Parents' Disincentives to Relinquish Birth parents, like other parents, generally prefer to raise the children they bear rather than surrender them for adoption. Legal economic scholars since the 1970s have proposed lifting the ban on payments to birth parents to help overcome the disincentive to relinquish. One scholar proposes that “baby market suppliers” – birth parents – “share in the full profits generated by their reproductive labor” (Krawiec 2009). Other scholars propose indirect payments, such as adoptive parents paying a birth mother’s legal expenses to ensure she fully understands her rights or paying the birth mother’s college tuition or job training expenses after the

placement. These payments would have both monetary and nonmonetary benefits. Tuition payments in particular would reduce the chance of the birth mother having to relinquish another child since her education and training should improve her ability to support any child she gives birth to in the future (Hasday 2005).

Birth parents may deem one type of non-monetary exchanges as more valuable than any monetary payments since it reduces somewhat the emotional cost borne by birth parents who lose all contact with their children. As adoption rates plummeted in the 1970s, that decreased supply and increased demand enabled birth mothers to exercise more bargaining power. Agencies began to let the birth mothers select the adoptive parents and also negotiate with prospective adoptive parents to get them to make promises to raise the child in a particular way (i.e., Catholic or with music education) and to provide the birth mother with periodic letters, pictures, email contact, and even in-person visits as the child grows up. These latter agreements, known as Post Adoption Contact Agreements, are increasingly common and increasingly enforced by courts (Ertman 2015).

Lower Transactions Costs Other scholars seek to increase adoptions by decreasing transactions costs. One scholar critiques the uncertainty created by statutes that allow a long period for birth parents to revoke their consent to an adoption on the grounds that courts could determine voluntariness of that consent at an early point, creating certainty for a child’s place in his or her new family (Brinig 1994).

Increase Demand for Hard-to-Place Children Legal reforms aimed to increase the demand for children in foster care focus on adoptive parents because they are the ones who decide whether and whom to adopt. Legal economists have argued for making adoptive parents’ substitutes for adopting a foster care child more expensive, subsidizing foster care adoptions to reduce the price of adopting a special needs child, and lowering transactions costs.

Make Adoptive Parents' Substitutes More Expensive

Many, if not most, adoptive parents turn to adoption only after infertility treatments fail. Since the 1990s, reproductive technology techniques – especially in vitro fertilization (IVF) – have improved so that they more effectively help women become pregnant. In that same period, rates of domestic adoption have continued to decline. Accordingly, economists have asked whether IVF is a substitute for adoption. One study found that between 1999 and 2006, a 10% increase in adoption correlated with a 1.3–1.5% decrease in the number of IVF cycles performed. The correlation was higher when the researchers focused on infant adoptions, older adoptive mothers, and international adoption. Unsurprisingly, since adoption by relatives occurs in specialized circumstances such as stepparent adoption, prevalence of IVF does not correlate with those adoptions (Gumus and Lee 2012).

Legal reforms could make IVF more expensive to pull more adoptive parents toward children in foster care. A state legislature could remove subsidies to IVF such as state-mandated insurance coverage for the expensive procedure or require a formal legal process akin to adoption in IVF procedures using donor eggs (Appleton 2004).

Some scholars contend that race matching reduces demand for foster care children, since many adoptive parents are white and most children in foster care are either African-American or Latino. If the race of one child substituted for the race of another, the reasoning goes, then federal laws and adoption policies that prohibit delaying or denying an adoption because of the child or adoptive parents' race help the market clear faster. Yet empirical research indicates that race matching does not reduce the number of adoptions (Hansen and Hansen 2006).

Subsidize Adoption out of Foster Care

Subsidies have been among the most effective tools to spur demand for children in foster care. Because many prospective adoptive parents cannot afford the \$30,000 or more required to adopt

an infant through a private agency, adopting a child through a public agency may be the only available path to parenthood. Since the 1970s the federal government has enacted subsidies to encourage these adoptions, and empirical research has demonstrated the positive and statistically significant effect of subsidies on the rate of those adoptions (Hansen 2007).

Inefficiencies persist, however. Subsidies that continue after the adoption offset the costs borne by adoptive families who care for children with special needs (often due to early adverse experiences that landed them in foster care). But these adoption assistance programs are administered by states rather than the federal government, which has caused variation among states. Some states issue payment based on the type of adoptive family rather than the needs of the child, redirecting the subsidy away from its intended purpose (Hansen 2008).

One proposal to increase demand for children in foster care suggests an “all-pay simultaneous ascending auction with a bid cap” with prospective parents submitting sealed bids. Profits generated by placing healthy white infants would be used to place the children they call “less-desirable.” (Blackstone 2004; Blackstone et al. 2008).

Lower Transactions Costs

Some legal economists have critiqued the agency gatekeeping function as extracting surplus (Blackstone et al. 2008) and increasing transactions costs (Brinig 1994). One proposal would reduce search costs by replacing that agency function with a national database, detailing the characteristics and requirements of prospective adoptive and birth parents and children on a platform akin to the Multiple Listing Service used by real estate agents (Balding 2010). A second proposal would impose a 10% tax on adoption expenses and channel the funds generated in high-price adoptions to subsidize adoption of children out of foster care (Goodwin 2006). A third proposal, already in place in many states, reduces search costs borne by both agencies and the children shuttled through foster care placements while they wait to be adopted. Instead of limiting the search for appropriate parents to married

heterosexuals, this approach expands the definition of suitable parents to include single people and same-sex couples. This expansion shortens the search because single and gay adoptive parents are more willing to create interracial families than their married, heterosexual counterparts (though whites in all three groups generally express a preference for white children) (Baccara and Collard-Wexler 2012).

Conclusion

Despite sharp criticism of early legal economic literature proposing economic frameworks for viewing adoption and remedying the two problems of queues of adoptive parents waiting to adopt healthy infants and queues of foster care children waiting to be adopted, economic and legal researchers continue to propose market solutions to remedy the situation.

Cross-References

- [Becker, Gary S.](#)
- [Contract, Freedom of](#)
- [Gender Diversity](#)
- [Posner, Richard](#)

References

- Appleton S (2004) Adoption in the age of reproductive technology. *University of Chicago Legal Forum* 393–450
- Baccara M, Collard-Wexler, A (2012) Child-adoption matching: preferences for gender and race, NBER Working paper 16444
- Balding C (2010) A modest proposal for a two-sided market clearing institution under asymmetric supply constraints with skewed pricing: the market for adoption and abortion in the United States. *J Public Econ Theory* 12:1059–1080
- Beam C (2013) To the end of June: the intimate life of American foster care. Houghton Mifflin Harcourt, Boston
- Becker GS (1981) *A Treatise on the family*. Harvard University Press, Cambridge
- Bernal et al (2007) Child adoption in the US, manuscript
- Bitler M, Zavodny M (2002) Did abortion legalization reduce the number of unwanted children? *Persp. Sex & Repro Health* 34:25–33
- Blackstone EA (2004) Privatizing adoption and foster care: applying auction and market solutions. *Child Youth Serv Rev* 26:1033–1049
- Blackstone EA et al (2008) Market segmentation in child adoption. *Int Rev Law Econ* 28:220–225
- Brinig MF (1993–1994) The effect of transactions costs on the market for babies. *Seton Hall Legis J*, 18:553
- Ertman M (2015) *Love's promises: how formal and informal contracts shape all kinds of families*. Beacon Press, Boston
- Ertman M (2003) What's wrong with a parenthood market. *N C Law Rev* 82:1
- Gentleman LA (1999) The Supply of Infants Relinquished for Adoption. *Econ Inq* 37:412–431
- Goodwin M (2006) The free market approach to adoption. *Boston Coll Third World Law J* 26:61–79
- Gumus G, Lee J (2012) Alternative paths to parenthood: IVF or child adoption. *Econ Inq* 50:802–820
- Hansen ME (2007) Using Subsidies to Promote the Adoption of Children from Foster Care. *J Fam Econ Iss* 28:377–393
- Hansen ME (2008) The distribution of a federal entitlement. *J Socio-Econ* 37:2427–2442
- Hansen ME, Hansen B (2006) The Economics of Adoption of Children from Foster Care. *Child Welfare* 85:559–583
- Hasday J (2005) Intimacy and economic exchange. *Harv Law Rev* 119:491–530
- Herman E (2008) *Kinship by design: a history of adoption in the modern United States*. Chicago University Press, Chicago, p 37
- Hollinger (2004) State and federal adoption law. In: Cahn NR, Hollinger JH (eds) *Families by law: an adoption reader*. New York University Press, New York
- Joyce K (2013) *The child catchers: rescue, trafficking, and the new gospel of adoption*. Public Affairs, New York
- Krawiec KD (2009) Altruism and intermediation in the market for babies. *Washington & Lee Law Rev* 66:203–257
- Landes EM, Posner RA (1978) The economics of the baby shortage. *J Legal Stud* 7:323–348
- Medoff MH (1993) An Empirical Analysis of Adoption. *Econ Inq* 31:59–70
- Moriguchi C (2012) The evolution of child adoption in the United States, 1950–2010: an economical analysis of historical trends. Discussion Paper Series A No. 572
- Posner RA (1987) The regulation of the market in adoptions. *Boston U L Rev* 67:59–72
- Solinger R, Susie WUL (1992) *Single pregnancy and race before Roe v Wade*. Routledge, New York
- Williams PJ (1995) *The rooster's egg: on the persistence of prejudice*. Harvard University Press, Cambridge, MA, pp 218–225
- Zelizer V (1985) *Pricing the priceless child: the changing value of children*. Basic Books, New York

Further Reading

- Statutes regulating payment of birth mother expenses: Ind. Code § 35–46–1–9 (2013); Md. Code Ann., Fam. Law § 5–3A–45 (2013); In re Adoption No. 9979, 591 A.2d 468 (Md. 1991)

Alcohol Prohibition

Nicholas A. Snow

Department of Economics, Indiana University,
Bloomington, IN, USA

Definition

Alcohol prohibition was a period from 1920 to 1933 in the United States of America where the manufacture, transportation, and sale of intoxicating beverages was made illegal.

Alcohol Prohibition

The national prohibition of alcoholic beverages in the United States of America lasted from January 16, 1920 to December 5, 1933. This period of time has become known as the “Noble Experiment,” a phrase coined by President Herbert Hoover. Alcohol prohibition was instituted by the 18th Amendment to the United States Constitution, which stated, “the manufacture, sale, or transportation of intoxicating liquors within, the importation thereof into, or the exportation thereof from the United States and all territory subject to the jurisdiction thereof for beverage purposes is hereby prohibited.” The Volstead Act was put into place for the purpose of enforcing this constitutional amendment. The experiment was meant to alleviate the various social ills that many perceived alcoholic beverages created such as violence, crime, corruption, and poor health.

In theory, prohibition would reduce consumption, which, in turn, would be followed by a reduction of these social problems. To judge this policy from an economic perspective, we look at whether or not this end goal was achieved. From this perspective, the “Noble Experiment” is often seen as failure.

It is interesting to note, however, that economists right before alcohol prohibition in the United States were primarily *for* prohibition. Their arguments rested on a belief that a sober society would outcompete heavy-drinking societies. Economist

Irving Fisher’s three books on the subject, 1927s *Prohibition at its Worst*, 1928s *Prohibition Still at its Worst*, and 1930s *The Noble Experiment*, all provided statistical evidence for the increased efficiency of prohibition and argued that the failures are a result of suboptimal enforcement. Fisher even claimed he could not find a single economist opposed to the prohibition of alcohol. Perhaps it was hindsight that changed this perception, as few would look back on alcohol prohibition as a success. The overwhelming empirical evidence and a look at economic theory clearly illustrate the difficulties alcohol prohibition had in successfully achieving its ends.

While alcohol consumption did decrease in the 1920s, prohibition did not help achieve the stated goals of prohibition. In fact, in a 1991 essay entitled “Alcohol Consumption During Prohibition,” Miron and Zwiebel estimate that while consumption fell to 30% of the pre-prohibition level in the 1920s, it steadily rose throughout the rest of the prohibition era, reaching about 70% of the pre-prohibition level. And this occurred despite increased enforcement efforts.

From a theoretical standpoint, this is not surprising. Prohibition is a supply-reduction policy, which shifts the supply curve up and to the left. This leads to a reduction in quantity demanded; however, it also leads to an increase in price. The higher price offers a compensation for the increased risk prohibition creates in continuing to operate in the market. In other words, the increases in enforcement efforts do lead more risk adverse individuals to exit the market; however, the higher prices provide incentives for more daring entrepreneurs, and those more adept at using violence and secrecy, to remain or enter the market.

The prohibition also had a profound effect on the quality and nature of the goods within alcohol markets. Quality tended to fall dramatically as consumers turned to the black market or to produce their own in order to meet their demand. The largest effect is what Richard Cowan referred to as the “Iron Law of Prohibition,” which states that the greater the level of enforcement, the greater to potency of the goods in question will become. Within the alcohol market, this meant a shift away from lower alcohol by volume drinks, and

thus bulkier, such as beer and wine, toward higher alcohol by volume drinks, such as whiskey.

Tellingly, economist Mark Thornton in a 1991 study entitled “Alcohol Prohibition was a Failure” analyzed the effectiveness of alcohol prohibition at achieving its stated goals of increase public health and decreasing crime. He found that prohibition had no discernable effect on public health, and perhaps more shockingly, he found that there was an increase in the crime rate, particularly among homicides. Thus, we must judge alcohol prohibition as having failed to accomplish its goals. These findings echo the work done by Clark Warburton at the tail end of prohibition with his 1932 book *The Economic Results of Prohibition*.

Fortunately, America’s experiment with alcohol prohibition ended with the passage of the 21st amendment. After which, there was a dramatic reduction in crime, including organized crime, and an increase in health, jobs, and prosperity for all Americans.

Cross-References

► [Prohibition](#)

References

- Cowan RC (1986) A war against ourselves: how the narcs created crack. *Natl Rev* 5:26–31
- Fisher I (1927) Prohibition at its worst. Alcohol Information Committee, New York
- Fisher I (1928) Prohibition still at it worst. Alcohol Information Committee, New York
- Fisher I (1930) The noble experiment. Alcohol Information Committee, New York
- Thornton M (1991) Alcohol prohibition was a failure. *Policy Anal* 157
- U.S. Treasury Department (1930) Statistics concerning intoxicating liquors. Bureau of Industrial Alcohol December, 1930
- Warburton C (1932) The economic results of prohibition. Columbia University Press, New York

Further Reading

- Okrent D (2010) Last call: the rise and fall of prohibition. Scribner, New York
- Thornton M (1991) The economics of prohibition. University of Utah Press, Salt Lake City

Alimony

Cécile Bourreau-Dubois and Myriam Doriat-Duban

Department of Economics, BETA UMR 7522, Université de Lorraine, Nancy, France

Abstract

Alimony or spousal support is a transfer of income between spouses intended mainly to reduce inequality in living standards following a divorce. Economic analysis provides two justifications for maintaining it in societies where women are now less financially dependent on their husbands and fault is no longer a decisive factor in divorce: the efficiency of marriage and the prevention of opportunism. It also provides some theoretical justifications for the methods used to calculate it.

Synonyms

[Spousal support](#)

Introduction

In many Western countries, the law provides that in the event of divorce, one of the spouses (generally the husband) will pay a sum of money to the ex-spouse (generally the wife) in the form of a lump-sum or regular payments, in particular when the separation leads to a significant difference in living standards. This transfer has different names in different countries: *pension alimentaire pour époux* in Quebec, *spousal support* in Canada, *alimony* in the United States, *spousal maintenance* in the United Kingdom, and *prestation compensatoire* in France. For a long time, the justification for it was based on the wife’s financial dependence on the husband and/or adultery. In the past decades, two phenomena have raised questions about whether it is still justified to maintain this obligation between former spouses. The first one is the mass entry of women into the job market. The second one is the

no-fault divorce revolution (Leeson and Pierson 2017; Bracke and Mulier 2017), which leaves less room to the fault in divorce law. The economic analysis of alimony has provided new justifications, based on notions of efficiency and the prevention of opportunism; it also provides methods of calculation compatible with those justifications.

Alimony and the Economic Efficiency of Marriage

The first model of alimony was developed by Landes (1978), following on from the work done by Becker (1973, 1974). The approach is a normative one: alimony is supposed to encourage spouses to maximize their joint domestic production (income, upbringing of children, preparation of meals and activities contributing to the family's well-being).

The model rests on three hypotheses: the household's production is specific in the sense that it loses all value in the event of divorce, the domestic assets and wife's income depend on the amount of time she devotes to the home, and the husband's earnings depend on the time he devotes to his work but also on the time his wife devotes to domestic tasks. These hypotheses therefore fix a framework in which only the wife takes part in housework. The optimum degree of the wife's specialization is determined in a two-period model, where the variable to be maximized is the present value of the couple's total income. In this context, alimony is justified by the existence of transaction costs (costs of negotiating and implementing the transfers of wealth between the members of the couple). More precisely, if the household's revenue is totally divisible, if the transfers take place at no cost, and if the estimated likelihood of divorce is identical for both spouses, the optimum level of specialization is obtained without a judge having to decide alimony when divorce occurs. On the other hand, positive transaction costs (due to information asymmetries, indivisible public goods like children) can prevent the optimum level of domestic production being reached by "voluntary" transfers of income. Because it guarantees the wife that she will recoup the fruits of her investment in her household, alimony encourages her to specialize in domestic

tasks, thereby maximizing the couple's expected income. Alimony is therefore nothing other than the implicit price, paid to the wife upon divorce, that encourages the spouses to behave in optimum fashion during the marriage.

This concept has aroused a considerable amount of criticism, in particular from feminist legal experts who object that it encourages women to specialize in domestic tasks and helps to perpetuate gender inequality (Singer 1994). Landes's hypothesis according to which only the wife can divide her time between market activities and non-market (domestic) activities is decisive in this result. Although Landes does not justify it and it can appear somewhat dated, Ellman (1989) points out that it can still be seen to correspond to a certain reality. Singer (1994) adds that this hypothesis also lacks any internal and external foundation. Thus, it ignores specialization between women rather than between spouses: nowadays the most productive household is one in which both spouses work full time and entrust care of the home and children to low-paid other people (most often women) (Carbone 1990; Brinig 1993). In addition, for many couples, a marriage is efficient when both members of the couple are committed to their career and devote a significant amount of time to their children (Brinig 1993).

Alimony and the Prevention of Opportunism

Marriage, in some ways, is quite similar to a long-term contractual relationship (Cohen 1987, p. 267). Alimony can then be justified by the concern to protect the weaker spouse who has made specific non-redeployable investments (i.e., domestic activities) from the opportunistic behavior of the other spouse.

Marriage is considered here as a long-term partnership, during which the spouses enter into reciprocal commitments, both explicit (imposed by law such as respect, faithfulness, moral assistance, duty of support) and implicit (promises). Nevertheless, marriage is different to a long-term commercial contract. Even if marriage has an instrumental value in the sense that the spouses can be considered as inputs enabling the production of an output (Becker's approach), above all it has an intrinsic value of its own (religious,

spiritual, proof of love) that a commercial contract does not have (Cohen 1987). By this specific dimension, marriage is a way of exchanging promises, whose value depends on the long-term behavior of each spouse. This is where a risk arises that is common to all long-term contracts: the risk of opportunism. The problem comes from the fact that the gains from the specific investments made generally by the wife are unequally spread over time. At the beginning of the union, the wife makes specific investments (bringing up children, cleaning, etc.) in return for the financial support of her husband, including once the children have left the parental home. The risk is then that the husband will adopt an opportunistic behavior consisting of breaking off the contract by a divorce just at the time when the spouse should be receiving her due, once the children have grown up. The husband thus appropriates what contract theory economists call the “quasi-rent” belonging to his wife. Geddes and Zak (2002) confirm that the husband’s optimum strategy effectively consists of leaving his wife once she has made her contribution to the marriage. Dnes (1999) assimilates this behavior with a “greener grass effect” which encourages the husband to terminate the union before having to “pay” his wife for her investment in the family and to go to live with another – younger – woman. The consequences for the wife are all the greater when the investments made in the household cannot be redeployed on the labor and/or remarriage market. The change from fault-based divorce to no-fault divorce has increased this risk of opportunism because one of the spouses can now terminate the marriage at any time, virtually without compensation for the other spouse (Brinig and Crafton 1994).

The solution consists of making more costly for the potentially opportunistic party to break the contract costly for the potentially opportunistic party. Alimony can play this role since, on the one hand, the wife knows that in the event of a divorce, she will be compensated for her investment in domestic life (here we find Landes’s incentive-based approach once again), while the husband is encouraged not to behave in an opportunistic way as divorce will not enable him to escape his commitments to his wife.

Alimony as a Means of Compensation

In both of these analyses, alimony constitutes a form of compensation justified by the recognition of a domestic investment made by the creditor during the marriage. This raises the question of how to compensate the party adversely affected by the ending of the marriage. By analogy with the notion of contract damages as compensation proposed in American contract law (Fuller and Perdue 1936), three forms of compensation are envisaged in the literature on alimony (Starnes 2011).

According to logic based on *restitution*, when one of the parties has made an investment that has led to the enrichment of the other party, but from which she cannot benefit, then it is necessary to pay her compensation. In this perspective, relying on the idea that marriage increases men’s productivity, and thereby their income, Carbone and Brinig (1991), like Landes (1978), consider that divorce would deprive the wife of her due in terms of what she has contributed to her husband’s success, especially when she has enabled him to continue his studies (Rea 1995) or when she has sacrificed her own career to create conditions more favorable to her husband’s professional success (Ellman 1989). The logic of restitution then leads alimony to be calculated by placing the party that has benefited professionally from the marriage (generally the husband) in the same situation as if the marriage had never existed.

According to the *reliance interest* logic, the compensation aims to compensate the injured party for the loss incurred (e.g., expenses incurred) following the termination of the contract. According to Brinig and Carbone (1988), the application of this logic is relevant to divorce because the domestic investment can lead to opportunity costs, in this case a loss of human capital due to the slowing of the professional career (Ellman 1989). This involves placing the injured party (generally the wife) in the same situation as if she had never been married.

Finally, according to the logic based on *expectation*, the injured party is compensated for the benefits that could have been expected if the marriage had not been terminated. As Cohen (1987) emphasizes, the gains of a marriage are often asymmetrically divided between the spouses.

The spouse who works rapidly enjoys the benefits of the marriage (comfort of home and family life, facilitated professional career), while the spouse who invests in the domestic sphere at the beginning of the marriage reaps the gains in the long term, once the children have grown up. The loss caused by the divorce then resides in the loss of those gains, which is, according to Cohen (1987), more a loss of conjugal services (affection, sex, complicity, etc.) than a loss of career opportunities or a deterioration in standard of living. This therefore involves placing the “victim” in a situation identical to that which would have prevailed if the contract has been met. More precisely, the assessment of the loss incurred then involves the determination of the minimum sum the spouse will have to pay the other spouse to accept the divorce, if only divorce by mutual consent were possible. In other words, this involves determining the minimum sum that the spouse must pay to the other spouse so that it makes no difference whether they divorce or remain married. The expected loss criterion therefore guarantees that only efficient marriage terminations take place.

Bolin (1994) establishes a link between Landes’s model (Landes 1978) and the three logics of compensation identified by American contract law. According to Bolin, this model, in which alimony guarantees an efficient specialization in the two spouses’ roles, is compatible with an infinity of amounts of alimony, which depend on the parties’ powers of negotiation. Nevertheless, these amounts are bounded by two extreme values: the highest amount that the husband would agree to pay and the lowest amount that the wife would agree to receive. More precisely, Bolin shows that the highest amount would correspond to the gain made by the husband as a result of the socially optimal investment of his wife in the household, which in fact would correspond to compensation according to the *restitution* principle. For its part, the minimal amount of alimony would correspond to the minimal amount accepted by the wife. This would be such that the wife is just compensated for the losses she has incurred due to her increased investment in domestic tasks. Such an amount would be very close to compensation based on the *reliance*

interest logic. On the other hand, according to Bolin, compensation in terms of *expectation* could not be integrated into the efficiency model proposed by Landes insofar as it would lead to a level of investment on the part of the wife that would be higher than the optimal level.

Summary

Economic analysis provides justifications for the existence of alimony, including in modern societies where it no longer corresponds to a sanction for misconduct or a prolongation of the duty of support toward a financially dependent wife. The justifications are based first of all on an efficient division of tasks between the spouses and secondly on the prevention of a risk of opportunism to the detriment of the spouse that has made the specific investments. In addition, economic analysis proposes criteria for calculating alimony.

References

- Becker GS (1973) A theory of marriage: part I. *J Polit Econ* 81(4):813–846
- Becker GS (1974) A theory of marriage: part II. *J Polit Econ* 82(2):S11–S26
- Bolin K (1994) The marriage contract and efficient rules for spousal support. *Int Rev Law Econ* 14:493–502
- Bracke S, Mulier K (2017) Making divorce easier: the role of no-fault and unilateral revisited. *Eur J Law Econ* 43:239–254
- Brinig M (1993) The law and economics of no-fault divorce, family. *Law Quart* 26(453):456–58.
- Brinig M, Crafton S (1994) Marriage and opportunism. *J Leg Stud* 23(2):869–894.
- Brinig M, Carbone J (1988) The reliance interest in marriage and divorce. *Tulane Law Rev* 62:855–884
- Carbone J (1990) Economics, feminism, and the reinvention of alimony: A reply to Ira Ellman. *Vanderbilt Law Rev* 1463:1463–1501.
- Carbone J, Brinig M (1991) Rethinking marriage: feminist ideology, economic change, and divorce reform. *Tulane Law Rev* 65:953–1010
- Cohen L (1987) Marriage, divorce and quasi-rents: or I gave him the best years of my life. *J Leg Stud* 16:267–272
- Dnes A (1999) Marriage contracts. In: B. Bouckaert and De G. Geest (Eds), *Ency Law Econ*, Edward Elgar 2000:864–885.
- Ellman I (1989) The theory of alimony. *Calif Law Rev* 77(1):1–81

- Fuller LL, Perdue WR (1936) The reliance interest in contract damages. *Yale Law J* 46:52–96
- Geddes R, Et Zak P (2002) The rule of one-third. *J Leg Stud* 31(1):119–137.
- Landes E (1978) Economics of alimony. *J Leg Stud* 7(1):35–63
- Leeson PT, Pierson J (2017) Economic origins of the no-fault divorce revolution. *Eur J Law Econ* 43: 419–439
- Rea S (1995) Breaking up is hard to do: the economics of spousal support, working paper number UT-ECIPA-REAS-95-01. <http://www.epas.utoronto.ca:5680/wpa/wpa.html>
- Singer J (1994) Alimony and efficiency: the gendered costs and benefits of the economic justification for alimony. *Georgetown Law J* 82:2423–2460
- Starnes CL (2011) Alimony theory. *Fam Law Q* 45(2): 271–291

Allowance Trading

► Emissions Trading

Alternative Dispute Resolution

Nathalie Chappe
University of Franche Comté CRESE, Besançon,
France

Definition

Alternative dispute resolution (ADR) refers to any mode of dispute resolution that does not utilize the court system, such as arbitration, neutral assessment, conciliation, and mediation. Methods of ADR are different from one another, but share common points, notably the feature that a third party is involved and a less formal and complex framework than courts. The third party offers an opinion about the dispute to the disputants or chose a binding decision. In recent decades, many countries have adopted rules requiring parties to go through some form of ADR before resorting to trial. ADR programs currently operate in a wide variety of contexts: union-management

negotiations, commercial contract disputes, divorce negotiations, etc. The utilization of ADR mechanisms is championed by parties and lawyers, as well as by politicians or judges. Parties and lawyers hope to withdraw from ADR a benefit in terms of time and costs. Politicians and judges seek to relieve congestion in the courts by providing new methods of dispute resolution to meet the increasing demand for justice.

The economic analysis of the ADR mechanisms returns to the following question: is ADR actually an efficient and cost-reducing system? To answer the first point about efficiency, we need to know the incentives created by ADR: parties' incentives and third party's incentives. The second argument concerns the choice of an ADR mechanism by the parties. Once these answers are given, we will be able to determine if the ADR should be subsidized, provided, or mandated by the state. The economic approach is based on the assumption that individuals are rational and act by comparing costs and benefits.

ADR can be binding or nonbinding (or hybrid), voluntary or mandatory, and ex ante or ex post. Ex ante ADR concerns arrangements to use ADR made before a dispute arises, while ex post ADR refers to the use of ADR once the dispute has arisen. Secondly in binding ADR, the parties will be bound and abide by the decision of the third party. In nonbinding ADR, each party is free to reject the decision of the neutral and either takes the matter to court or does nothing. A hybrid procedure could include a nonbinding procedure followed by a binding procedure if needed. Thirdly, ADR can be voluntary or mandatory. These distinctions are important because the costs and benefits of different kinds of ADR may be different, hence a different impact on the parties' decision to turn to ADR, on the third party's behavior and on the incentives to settle.

Why Parties Turn to ADR to Resolve Disputes?

The incentives to turn to ADR are different between ex ante and ex post agreements.

Ex Ante Agreements

Ex ante ADR may be adopted because it would lead to mutual advantages. ADR may lower the cost of resolving disputes. Mediation and arbitration are indeed less expensive methods than court. ADR may also lower the risks attending disputes. Arbitration may be chosen because the arbitrator will be an expert whose award may be anticipated. Note, though, that the two previous benefits are equally advantages of ex post agreements.

Ex ante ADR may act as a quality signal. Chappe (2002) shows that in a contractual relationship between one buyer and two sellers (one of which offers a high-quality product, the other a low-quality product), the high-quality seller offers an arbitration clause, which then serves as a quality signal. Holler and Lindner (2004) analyze mediation as a signal; it possibly gives away information about the party who call for it. More precisely, to reject mediation can be interpreted as a “negative signal,” while the interpretation of accepting or proposing mediation is ambiguous and provides no clear information.

Moreover, since ADR provides greater social benefits and information than litigation, the dynamics of the process should tend to induce the parties to include a clause submitting future disputes to ADR (e.g., an arbitration clause). Such a clause has a deterrent effect on behavior that may prevent the dispute. Defendants are induced to take a more efficient level of care since their anticipation in case of dispute is better.

Such benefits cannot be obtained by ex post agreement since it is too late to disclose information or to take care.

Ex Post Agreements

Shavell (1995) follows the standard model of litigation which assumes that disputing parties choose ADR or not depending on the potential outcomes. “Parties are risk neutral, bear their own legal costs, and know the judgment amount that would be awarded, but may hold different beliefs about the likelihood of a plaintiff victory.” The expected outcome is defined as the probability of winning multiplied by the judgment amount won minus the costs. Parties will choose an agreement if and only if it is Pareto superior to their

alternatives. Parties determine the net expected values (expected value for the plaintiff minus the expected value for the defendant) for ADR, settlement, and trial. Depending on this value, parties will opt for different method of dispute resolution. Shavell (1995) also holds that parties have probabilistic beliefs about how ADR influences the outcome of the trial: ADR perfectly predicts trial outcomes, ADR does not predict trial outcomes, and ADR imperfectly predicts outcomes. If ADR perfectly predicts trial outcomes, parties will opt for ADR which have a positive net expected value. If ADR does not predict trial outcome, nonbinding ADR will never occur.

Third Party

The role of the third party differs between binding and nonbinding ADR. In nonbinding ADR, as mediation, the third party never explicitly recommends how a dispute should be resolved. The third party assists parties in the resolution of their dispute and acts as a settlement facilitator. In binding arbitration, the neutral evaluates the merits of the case and issues a decision which is binding.

Binding

The most famous binding ADR is arbitration. Compared to court adjudication, the advantages and drawback of arbitration come from three differences. Firstly, unlike a judge (who is assigned), the arbitrator is a private person chosen by the parties. He is chosen for his expertise in the subject matter of the dispute. Consequently, the arbitrator’s award may be more predictable (there is less risk) and faster (Ashenfelter 1987; Bloom and Cavanagh 1986).

In conventional arbitration, arbitrators make awards that are weighted averages of some notion of what is appropriate based on the facts and the offers of the parties where the weights depend on the quality of offers as measured by the difference between the offers. In final offer arbitration, the arbitrators choose the offer that is closest to some notion of an appropriate award based on the facts (Bazerman and Farber 1985). Farber and Bazerman (1986) find strong results which support this framework.

Nonbinding

The mediator is not charged to solve the litigation, but simply to make proposals and to facilitate agreement. The mediator helps parties clarify issues, explore settlement options, and evaluate how best to advance their respective interests. The role of the mediator is thus above all to avoid the failure of agreement. Settlement may fail because of informational and psychological barriers (Arrow et al. 1995).

Models of settlement negotiations explain the occurrence of costly trials by assuming that parties have private information about the outcome of a trial. The role of the mediator will be to collect and distribute information in order to make sure that parties become aware of the mutual benefits that the mediation allows. Two points distinguish the mediator from the judge: distribution of information and confidentiality of the exchanges. Indeed, contrary to the judge who is held by the principle of contradictory, the mediator is free to disseminate or not informations. Ayres and Brown (1994) study how mediator resolve dispute by “caucusing” privately with the individual disputants. Mediators can create value by controlling the flow of information to mitigate moral hazard and adverse selection. Moreover, the information disclosed to the mediator cannot be used at trial. If such were not the case, the parts would not be encouraged to reveal information knowing that in the event of failure, the court could be put in possession of information which could be used for its decision, whereas in the absence of mediation, it would not have mentioned these facts within the framework of the lawsuit.

The main cognitive barriers are the attitude toward risk, the aversion for the losses and the framing effects. The mediator will limit the cognitive and the “reactive devaluation” barriers to conflict resolution (Arrow et al. 1995).

What Incentives to Settle Are Created by ADR?

Adding ADR to the traditional litigation process has complex and important effects on several ways. Notably, ADR will have an effect on the

incentives to settle. What is the incentives for parties to settle their disputes if ADR makes them less costly? We consider firstly binding ADR and secondly nonbinding ADR. In the two cases, incentives depend on the third party’s role which we have described above.

Binding ADR: The Case of Arbitration

It is often claimed that the final offer arbitration (FOA) system is more likely than the conventional arbitration (CA) system to lead the parties to settle. Under conventional arbitration (CA), the parties present their cases to an arbitrator who has the flexibility to impose any award he or she deems appropriate. The major criticism of CA is that arbitrators are perceived to compromise between the parties’ final offers, which encourages them to take extreme positions into arbitration resulting in a “chilling effect” on negotiated settlement (Farber 1981). FOA is claimed to induce disputants to stake out more reasonable positions (Farber and Katz 1979; Farber 1980). Under FOA, the arbitrator must choose either one party’s or the other party’s final offer. If the offers of the two disputants do not converge or criss-cross, then the arbitrator chooses the one that is closest to his/her notion of a fair settlement as the settlement. However, while FOA improves convergence and outperforms CA, several papers show that the design of the mechanism is not sufficient to harmonize the parties’ positions (Chatterjee 1981). To restore convergence between offers, several procedures have been presented. Brams and Merrill (1986) propose a “combined” arbitration procedure (CombA): if the arbitrator’s notion of a fair settlement lies between the disputants’ final offers, then the rules of FOA are used. Otherwise, the rules of CA are used. Under some conditions (concerning the density function of the estimate about the arbitrator’s notion of a fair settlement), this procedure succeeds in overcoming the defects of FOA and increases voluntary settlement rates. Zeng et al. (1996) propose to improve FOA by letting each disputant make double offers (double-offer arbitration, DOA). They show that the two secondary offers converge. DOA succeeds in inducing the convergence of the offers made by the

disputants without assumptions about the density function as in CombA. Zeng (2003) proposes an amendment to FOA. The final arbitration settlement is determined by the loser's offer instead of the winner's offer, if the offers diverge. The author shows that this design induces two disputants to submit convergent offers. Finally, Armstrong and Hurley (2002) propose to incorporate what they call the closest offer principle. Arbitrator may put a higher weight on the disputant's offer which is closest to his/her own assessment. Their model allows them to generalize previous models of both CA and FOA. They derive the equilibrium offers that risk neutral disputants would propose and show how these offers would vary under different arbitration procedures: FOA and CA. They show that the offers made under CA are always more extreme than those made under FOA.

Chappe and Gabuthy (2013) highlight the strategic implications of legal representation in arbitration by analyzing how the presence of lawyers may shape the disputants' bidding behavior. They show that the contingent payment mechanism has interesting properties in that it enhances the lawyer's incentives to provide effort in defending her client (in comparison with the case without legal representation), without altering the gap between the parties' claims in arbitration.

Nonbinding ADR

Doornik (2014) considers a mediator who acts as a channel to verify and communicate the informed party's beliefs about the expected value of the judgment, without sharing any supporting evidence that would advantage her opponent in court. This increases the incentive for the informed party to share their private information about the likely outcome in court and hence increases settlement rates.

Dickinson and Hunnicutt (2010) examine the effectiveness of implementing a nonbinding suggestion prior to binding dispute settlement. On the one hand, a nonbinding suggestion may serve as a focal point, thereby improving the probability of settlement. On the other hand, a nonbinding suggestion may reduce uncertainty surrounding the outcome from litigation, thereby increasing dispute rates. Their results suggest that nonbinding

recommendations improve bargaining outcomes by reducing optimism. Dispute rates are significantly lower when a nonbinding recommendation is made prior to arbitration.

Conclusion

ADR is now an integral part of the litigation system. Ex ante ADR agreements may result in mutual advantages for the parties to a contract, in superior incentives to disclose information or to avoid disputes, and in improved incentives to settle. Consequently, ex ante ADR agreement should be enforced. Nevertheless, some ADR (specifically arbitration) may be more expensive in comparison to the courts. Incorporating ex post ADR agreements to the standard litigation system has effects on incentives to settle immediately and to go to trial. Since parties settle in the shadow of ADR, their behavior is affected by the introduction of ADR. The conclusions about the desirability of ADR depend significantly on several factors: (i) the cost of ADR (mediation may be cheaper than trial, while arbitration may be more expensive), (ii) the predictive power of the ADR about the trial outcome, and (iii) the degree of coercion from the decision (binding or non-binding). Consequently it is difficult to give a general recommendation requiring compulsory arbitration.

References

- Armstrong MJ, Hurley W (2002) Arbitration using the closest offer principle of arbitrator behavior. *Math Soc Sci* 43(1):19–26
- Arrow K. et al. (1995) *Barriers to conflict resolution*. W. W. Norton & Company, 1st ed, p. 368
- Ashenfelter O (1987) Arbitrator behavior. *Am Econ Rev* 77(2):342–346
- Ayres I, Brown JG (1994) Economic rationales for mediation. *Va Law Rev* 80:323–402
- Bazerman M, Farber H (1985) Arbitrator decision making: when are final offers important? *Ind Lab Relat Rev* 39(1):76–89
- Bloom D, Cavanagh C (1986) An analysis of the selection of arbitrators. *Am Econ Rev* 76:408–421
- Brams SJ, Merrill S (1986) Binding versus final-offer arbitration: a combination is best. *Manag Sci* 32(10):1346–1355

- Chappe N (2002) The informational role of the arbitration clause. *Eur J Law Econ* 13(1):27–34
- Chappe N, Gabuthy Y (2013) The impact of lawyers and fee arrangements on arbitration. *J Inst Theor Econ* 139:720–738
- Chatterjee K (1981) Comparison of arbitration procedures: models with complete and incomplete information. *IEEE Trans Syst Man Cybern* 11(2):101–109
- Dickinson D, Hunnicutt L (2010) Nonbinding recommendations: the relative effects of focal points versus uncertainty reduction on bargaining outcomes. *Theory Decis* 69:615–634
- Doornik K (2014) A rationale for mediation and its optimal use. *Int Rev Law Econ* 38:1–10
- Farber H (1980) An analysis of final-offer arbitration. *J Confl Resolut* 24(4):683–705
- Farber H (1981) Splitting-the-difference in interest arbitration. *Ind Lab Relat Rev* 35:70–77
- Farber H, Bazerman, (1986) The general basis of arbitrator behavior: an empirical analysis of conventional and final-offer arbitration. *Econometrica* 54(6):1503–1528
- Farber H, Katz H (1979) Interest arbitration, outcomes, and the incentive to bargain. *Ind Labor Relat Rev* 33(1):55–63
- Holler MJ, Lindner I (2004) Mediation as signal. *Eur J Law Econ* 17:165–173
- Shavell S (1995) Alternative dispute resolution: an economic analysis. *J Legal Stud* 24:1–28
- Zeng DZ (2003) An amendment to final-offer arbitration. *Math Soc Sci* 46(1):9–19
- Zeng D-Z, Nakamura S, Ibaraki T (1996) Double-offer arbitration. *Math Soc Sci* 31(3):147–170

Altruism

Sophie Harnay

EconomiX CNRS UMR 7235, University Paris
Ouest Nanterre La Defense, Nanterre, France

Abstract

Until recently, most works in the economic analysis of law (EAL) have assumed that material self-interest exclusively motivates economic agents, including individuals involved in the functioning of the legal system. Altruistic choices, defined as a sacrifice that benefits others, thus remain outside the scope of the EAL. They become an issue for the EAL when the question whether to substitute external (legal) incentives for internal (moral) motives when altruism appears insufficient to

have people help each other gradually emerges as a by-product of the liability controversy and as the outcome of the development of the economics of altruism in the 1970s.

Altruism

Since Adam Smith's *Theory of Moral Sentiments* and with landmarks in works of John Stuart Mill, Francis Ysidro Edgeworth, Vilfredo Pareto, or Léon Walras, the problem of altruism and giving, and more generally of the adoption of moral norms, has been a long-standing issue and a challenge within economics, due to the obvious tension between altruism and the self-interest assumption. Thus, most works in the economic analysis of law (EAL) assume that material self-interest exclusively motivates all people. Accordingly, individuals involved in the functioning of the legal system – criminals, litigants, judges, and lawyers – are supposed to behave as rational and self-interested utility maximizers. Furthermore, rules are seen as incentives designed to induce agents to behave in a way that is consistent not only with their private interest but also with social efficiency and public interest. Within this framework, thus, altruistic behaviors have not been an issue for the EAL for a long time, as they are apparently “irrational,” and the study of helping behaviors and of a wide set of situations involving individuals who depart from the pursuit of their own interest to help others has been overlooked.

In the field of the economic analysis of law, disinterested behaviors start to become an issue in the early 1970s, as a by-product of the controversy on the economic analysis of liability and negligence that rages at the time. The questions that are raised then are the following: how is it possible to explain behaviors that are not obviously the consequence of selfishness? Can unselfish behaviors possibly be regulated? Should disinterested behaviors be encouraged and regulated? Is that possible to substitute external (legal) incentives for internal (moral) motives when altruism appears insufficient to have people cooperate and help each other?

A first step in the acknowledgment of disinterested behaviors in the law and economics literature is taken with Richard A. Posner's pioneering works on liability rules (Posner 1972). Arguing that liability in a case should be decided following a cost-benefit analysis, Posner explicitly refers to rescue situations, helping behaviors, and good and bad Samaritanism as illustrations for his economic theory of liability and support for its negligence rule within the debate on strict liability versus negligence (Posner 1973). However, no reference is made to the economic analysis of altruism at that time: within Posner's analytical framework, rescue situations do not differ from less extreme situations substantially and, therefore, they do not call for specific liability rules.

A second step focusing on the regulation of rescue behaviors more explicitly builds on the advances in the economic analyses of altruism that are made during the 1970s. During the decade, indeed, economists start to analyze altruistic behaviors with economic tools and assumptions. They approach altruism in three main ways. First, following Becker's works (1973, 1974a, b, 1976) and Dawkins (1976), a first family of works models altruism using interdependent utility functions in which the welfare of others enters an individual's utility function: thus, the higher the utility level of others, the higher the utility of the benefactor. A second family of analyses studies reciprocal altruism, as it is determined by the expectation of future gains within repeated interactions between players. A third set of models assumes dual utilities of agents: following up Harsanyi's works (1976), they consider a dual nature of man combining both an egoistic and an altruist side within a single individual. Such duality then accounts for pro-social behaviors, including pure altruism, and explains disinterested behaviors as the consequence of a moral imperative ruling out rational calculation and a natural propensity to moral behaviors. The economic debate on altruism is also fueled by the discussion of economists with biologists and the emergence of sociobiology (Wilson 1975).

In the field of law and economics, analyses concentrating on rescue behaviors mostly build on models of interdependent utility functions, in

which the altruistic rescuer's utility function becomes a function of the utility of the endangered person (Landes and Posner 1978a, b). Analyses therefore endorse Becker's claim according to which economic analysis is sufficient to explain altruistic behaviors, without the need to resort to sociobiology and models of group selection to account for altruistic behavior. Furthermore, reciprocal altruism is also considered as irrelevant to account for rescue behaviors, as rescue most often involves strangers and relies on some form of pure altruism, devoid of all expectation of compensation. Rather, altruistic behavior is justified by a "recognition factor" that rescuers receive from helping others. Consequences in terms of regulation are then drawn from the existence of altruism. If people are mainly motivated by recognition in their altruistic transactions, including their rescue activities, it may not be efficient to impose liability for non-rescue: because it reduces the public recognition allotted to the altruistic individual, liability rules may then have a pernicious effect and decrease the number of altruistically motivated actions. In that sense, altruism thus provides a substitute for costly legal rules intended to internalize external benefits of rescues in emergency situations, when transaction costs are too important to allow an efficient use of legal rules.

References

- Becker GS (1973) A theory of marriage: part I. *J Polit Econ* 81(4):813–846
- Becker GS (1974a) A theory of social interactions. *J Polit Econ* 82(6):1063–1093
- Becker GS (1974b) A theory of marriage: part II. *J Polit Econ*, Part 2. Marriage, family human capital and fertility 82(2):S11–S26
- Becker GS (1976) Altruism, egoism and genetic fitness. *J Econ Lit* 14(3):817–826
- Dawkins R (1976) *The selfish gene*. Oxford University Press, Oxford
- Harsanyi JC (1976) *Essays on ethics, social behaviour, and scientific explanation*. Reidel, Dordrecht/Boston
- Landes WM, Posner RA (1978a) Salvors, finders, good Samaritans, and other rescuers: an economic study of law and altruism. *J Leg Stud* 7(1):83–128
- Landes WM, Posner RA (1978b) Altruism in law and economics. *Am Econ Rev* 68(2):417–421

- Posner RA (1972) A theory of negligence. *J Leg Stud* 1(1):29–96
- Posner RA (1973) Strict liability: a comment. *J Leg Stud* 2(1):205–221
- Wilson EO (1975) *Sociobiology, the new synthesis*. Belknap/Harvard University Press, Cambridge, MA

Amnesty Programs

► Leniency Programs

Analytical

► Rationality

Anarchy

Peter T. Leeson
Department of Economics, George Mason
University, Fairfax, VA, USA

Abstract

Anarchic environments are ubiquitous. Unable to rely on government, individuals in such environments develop private institutions of governance to promote social cooperation. Private governance consists of privately created social rules and mechanisms of their enforcement. Anarchy may secure better socioeconomic development than government in least-developed countries.

Synonyms

Statelessness

Definition

Anarchy is the absence of government. Government is more difficult to define. Typically

government refers to a governance institution with a territorial monopoly on force. However, when government is defined this way, the presence or absence of government, and thus the presence or absence of anarchy, depends on how one construes the territory in question. Other attempts to define government in terms of characteristics with which government is often associated, such as compulsory recognition of a governance institution's authority, or a "high" cost of exiting a governance institution, are similarly problematic. Many governance institutions ordinarily characterized as nongovernmental share these characteristics, confounding efforts to define government, and thus anarchy, in a way that consistently reflects common intuitions.

Despite this conceptual difficulty, there is widespread agreement about whether or not government, and thus anarchy, prevails in concrete cases. Although a criminal gang, for example, may enjoy a monopoly on force in a neighborhood, few would consider the gang a "government." On the contrary, most would characterize the gang's presence as an outcome of government's effective absence and thus the environment in which the gang operates as anarchic. The distinction between government and anarchy is therefore useful in practice and poses little problem for applied research that seeks to study "government" or "anarchy" (Leeson 2014a).

Anarchic Environments

A large number and wide variety of social environments worldwide were or are anarchic. Most of the world for most of its history lacked government. And, if one goes back far enough, all historical environments are anarchic.

In the contemporary world, dysfunctional governments in least-developed countries have resulted in large populations' inability to rely on government to protect their lives and property. Indeed, in such countries, government is often the chief depredator of citizens' lives and property. In one least-developed country, Somalia, there has been no government at all for several decades.

Despite having high-functioning governments, the contemporary developed world also hosts numerous anarchic environments. Even where government is highly functioning, the state's eye cannot be everywhere all the time. Moreover, appealing to governmental institutions to enforce contracts and property rights is often expensive and slow (Stringham 2015). Many social and commercial activities in the developed world thus occur outside government's *de facto* reach. Black markets, which are criminal and, even in developed countries, large, also present anarchic environments. Such markets' participants cannot rely on government to secure their property rights or promote cooperation between them (Skarbek 2014).

Internationally, the world has always existed in, and continues to exist in, anarchy. Various supranational organizations, such as the United Nations, seek to facilitate international cooperation. However, such organizations' member states remain sovereign. There is therefore no supranational agency with the authority to create or enforce binding laws on nation states and no agency with the authority to act as the final arbiter of disputes between them.

Interactions between private international traders also occur in an anarchic international environment. Before 1958, private international commercial contracts could not be enforced in government courts. Even after 1958, when the first international treaty was created for this purpose, the vast majority of international commercial contracts continue to be enforced without government. Moreover, like all international treaties, those which promise government enforcement of international commercial agreements are not themselves enforceable by government, since no supranational government exists (Leeson 2008).

Anarchy and Private Governance

In the seventeenth century, Thomas Hobbes famously characterized life in anarchy as "solitary, poor, nasty brutish, and short." Contemporary conventional wisdom sees anarchy similarly:

anarchy is widely believed to be chaotic, violent, and impoverishing. A growing body of research, however, finds that Hobbes' characterization of anarchy, and the contemporary conventional wisdom that echoes it, is wrong (Powell and Stringham 2009).

Although anarchic environments lack government, they do not lack governance: institutions that create and enforce social rules. In anarchy, such governance is provided privately – i.e., through the activities of private individuals seeking to make themselves better off. Private governance institutions in anarchy can be highly effective at producing social order, regulating violence, and more generally facilitating social cooperation.

Private governance institutions in anarchy take myriad forms. The particular forms they take reflect the particular problems of social cooperation that the people who find themselves in particular anarchic environments face. Private social rules may emerge in anarchy "organically" via the interactions of individuals, which, over time, generate customs or social norms that govern behavior. Private international commercial law, for example, emerged in this fashion via the interactions of medieval international merchants who, in the absence of supranational government, required rules to facilitate exchange between them (Benson 1989). Alternatively, private social rules may emerge in anarchy via individuals' deliberate construction. Nineteenth-century American settlers in the so-called "wild West," for instance, established private land clubs and cattle associations with the express purpose of creating laws to protect their property rights prior to the United States' government's appearance in frontier areas (Anderson and Hill 2004).

Private rule-enforcement mechanisms may also emerge in anarchic environments both "organically" and through private individuals' deliberate efforts. The most basic of such mechanisms is the "discipline of continuous dealings," according to which a social rule breaker is shunned or otherwise cut off from the benefits of future cooperative dealings by the members of his community. The members of Gypsy communities, for instance, whose gray market economic

activities typically prevent them from relying on government, use the threat of being ousted from the Gypsy world to enforce economic agreements between one another (Leeson 2013).

Other private punishments for the enforcement of social rules in anarchy include, for example, monetary fines, physical or capital punishment, and supernatural sanctions (Friedman 1979; Leeson 2009, 2014b). Private punishments for rule breaking in anarchy are often part of more elaborate and sophisticated private enforcement institutions. Eighteenth-century Caribbean pirates, for instance, who, as criminals, could not rely on government, used fines and capital punishment to punish violations of their private law, but developed, interpreted, and administered these punishments in the context of a broader system of constitutional democracy they developed to privately govern their ships (Leeson 2007a).

Anarchy and Development

Anarchy's relationship to socioeconomic development is nuanced. On the one hand, no anarchic society has come close to the level of socioeconomic development achieved by the most successful societies governed by states, namely, the countries of contemporary North America and Western Europe. This strongly suggests that government is necessary for maximal development.

On the other hand, governments in North America and Western Europe are unusual in the world in that they are highly functional. In most of the world, governments are not nearly so high functioning; and in much of the world, governments are highly dysfunctional. For a variety of reasons, governments in such places have failed to realize the potential exhibited by governments in the wealthiest nations.

This failure raises the question of whether private governance in anarchy may be able to secure a higher level of development than dysfunctional government. The developmental superiority of high-functioning government to private governance does not imply the developmental superiority of dysfunctional government to private

governance. Somalia, whose dysfunctional government collapsed in 1991, is the only contemporary nation that has experienced both a significant period under dysfunctional government and one in anarchy and thus the only contemporary nation that permits a direct comparison of development under these alternative governance arrangements. Nevertheless, the results of such a comparison are instructive.

On nearly every available development indicator, Somalia exhibited a marked improvement in anarchy compared to under its former government (Leeson 2007b; Powell et al. 2008; Leeson and Williamson 2009). This does not mean, of course, that Somalia achieved high levels of development in anarchy (it did not), nor does it mean that Somalia could not achieve dramatically higher levels of development if it could achieve a high-functioning government. It does, however, suggest the possibility that where dysfunctional government and private governance are the only governance alternatives – as would seem to be the case in much of the least-developed world – anarchy may be superior to government for development.

References

- Anderson TL, Hill PJ (2004) *The not so wild, wild west: property rights on the frontier*. Stanford University Press, Stanford
- Benson BL (1989) The spontaneous evolution of commercial law. *Southern Econom J* 55:644–661
- Friedman D (1979) Private creation and enforcement of law: a historical case. *J Legal Stud* 8:399–415
- Leeson PT (2007a) *An-arrgh-chy: the law and economics of pirate organization*. *J Polit Econom* 115:1049–1094
- Leeson PT (2007b) Better off stateless: Somalia before and after government collapse. *J Comp Econ* 35:689–710
- Leeson PT (2008) How important is state enforcement for trade? *Am Law Econ Rev* 10:61–89
- Leeson PT (2009) The laws of lawlessness. *J Legal Stud* 38:471–503
- Leeson PT (2013) Gypsy law. *Public Choice* 155:273–292
- Leeson PT (2014a) *Anarchy unbound: why self-governance works better than you think*. Cambridge University Press, Cambridge
- Leeson PT (2014b) God damn: the law and economics of monastic malediction. *J Law Econ Organ* 30:193–216
- Leeson PT, Williamson CR (2009) Anarchy and development: an application of the theory of second best. *Law Develop Rev* 2:77–96

- Powell B, Stringham EP (2009) Public choice and the economic analysis of anarchy: a survey. *Public Choice* 140:503–538
- Powell B, Ford R, Nowrasteh A (2008) Somalia after state collapse: chaos or improvement? *J Econ Behav Organ* 67:657–670
- Skarbek D (2014) *The social order of the underworld: how prison gangs govern the American penal system*. Oxford University Press, Oxford
- Stringham EP (2015) *Private governance: creating order in economic and social life*. Oxford University Press, Oxford

Anticommons, Tragedy of the

Qianwei Ying
Business School, Sichuan University, Chengdu,
Sichuan, China

Definition

The tragedy of the anticommons is a type of coordination breakdown, in which a single resource has numerous owners, each of whom has the right to individually exclude others from its use but no effective privilege of using it independently. This coordination failure, i.e., the exercise of the veto power from any one of the owners will result in the underuse of the resource, frustrating what would be a socially desirable outcome.

The Concept of the Tragedy of the Anticommons

According to the definition by Heller (1998), an *anticommons property* is a scarce resource in which multiple owners have the right to individually exclude others from its use, and no one has an effective privilege of using it independently. The use of anticommons property requires these numerous right holders of a single resource to reach a consensus on the permission. The coordination breakdown of the owners, i.e., the exercise of the veto power from any one of the excluded right holders, will result in the underuse of the

property, frustrate what would be a socially desirable outcome, and thus lead to the “tragedy of anticommons.” The “tragedy of the anticommons” can be considered as a mirror image of the older concept of tragedy of the commons described notably by Hardin (1968), where multiple individuals are endowed with the privilege to individually use a given resource but without a right or an effective way to monitor and constrain each other’s use, leading to the overuse of the resource. In terms of efficiency, the problem of the anticommons is based on a positive externality, while the problem of the commons is based on a negative externality.

The concept of the “tragedy of anticommons” is especially important for the transitional economies during the privatization process. One solution to the traditional tragedy of the commons is to fragment the common property into pieces and distribute them to different private owners. Private owners tend to avoid overuse because they can benefit directly from conserving the resources they control. Unfortunately, in the privatization process, too many separate owners of a single resource may be created, so that each one can block the others’ use of the resource as a whole. Meanwhile, the separate use of the part of the resource that each separate owner controls may be impossible or with little value. As a result, nobody can use the resource efficiently if cooperation fails, leading to the tragedy of anticommons. For example, suppose an acre of land is distributed to 100,000 individuals who each own one square foot. The land will never be used for anything as long as thousands of people have to agree on what to do.

The anticommons is a paradox. While private ownership usually increases the efficiency of the use of scarce resources and the social wealth, too many private owners from the fragmentation of a single property right has the opposite effect: it stops efficient use of the resource as a whole, hinders future innovation based on existing numerous intellectual property rights held by different owners, and may cost our lives (Heller 2008). Anticommons property appears when new property rights are being defined and allocated without mechanisms for resolving

ownership disputes. While Heller devotes most of his attention to the underuse of commercial property in Moscow and other cities in the former Soviet Union, the concept of the anticommons has much wider implications in many other cases. Anticommons property can also emerge in developed markets wherever new property rights are being defined. This may occur when new technologies make possible uses of, for example, intellectual property and environmental rights, unanticipated by the previously existing legal mechanism.

Once anticommons property appears, it is difficult to remedy the situation either through markets or subsequent regulation. Care must be taken to avoid the accidental creation of anticommons property when new property rights are being defined by conveying core bundles of rights rather than multiple rights of exclusion.

Typical Examples of the Tragedy of Anticommons

Tragedy of Anticommons in the Transition from Marx to Markets in Moscow

A classical example of the tragedy of anticommons is proposed by Michael Heller in his famous article “The tragedy of the anticommons: Property in the transition from Marx to markets” published in *Harvard Business Review*, in which he examines a paradox in Moscow after the dissolution of the Soviet Union (Heller 1998). Storefronts remained empty even though the economy was growing and there was demand for consumer goods. In contrast, street kiosks in front of them filled with goods and customers. Why did the new merchants not come in the storefronts? One reason was that transition governments often failed to endow any individual with a bundle of rights that represents full ownership. Instead, fragmented rights were distributed to various stakeholders, including private or quasi-private enterprises, workers’ collectives, privatization agencies, and local, regional, and federal governments. The fragmentation of the property right on the storefronts leads to a tragedy of the anticommons, an underuse of the storefronts due to the allocation of

multiple new owners with the rights to exclude others from its use. An owner benefits from keeping a storefront empty, by excluding others because exclusion preserves the value of the right, perhaps for later trade to property bundlers or perhaps for use in rent seeking. The right of exclusion is valuable precisely because others want to use the resource and will pay something to collect the right. Because multiple parties may hold the same right, almost any use of the storefront requires the agreement of multiple parties. Even if only one party opposes the use, that party may be able to block others from exercising their rights.

Moving a storefront from anticommons to private property ownership requires unifying fragmented property rights into a usable bundle. In other words, creating private property requires moving from too many owners, each exercising a right of exclusion, to a sole decision maker, controlling a bundle of rights. Unfortunately, this process of collecting the fragmented property rights back together could be extremely difficult. In markets, entrepreneurial property bundlers may assemble control over stores by negotiating with all the holders of rights of exclusion. However, the market route to bundling rights might fail altogether if the transaction costs of bundling exceed the gains from conversion or if owners engage in strategic behavior such as holding out for the conversion premium.

The tragedy of the storefront anticommons is that owners waste the resource when they fail to agree on a use. Empty stores result in forgone economic opportunity and lost jobs. As of 1995, about 95% of commercial real estate in Russia remained in some form of divided local government ownership. Of this commercial real estate, a significant portion was unused.

Tragedy of Anticommons in Patents and Intellectual Property Rights

Patents and other forms of intellectual property protection for upstream discoveries may stimulate incentives to undertake risky research projects and could result in a more equitable distribution of profits across all stages of R&D. However, too many owners holding patents or intellectual

property rights in previous discoveries may constitute obstacles to future research. A researcher who may have felt entitled to coauthorship or a citation in an earlier era may now feel entitled to be a coinventor on a patent or to receive a royalty under a material transfer agreement. The result has been a spiral of overlapping patent claims in the hands of different owners. Researchers and their institutions may resent restrictions on access to the patented discoveries of others, yet nobody wants to be the last one left dedicating findings to the public domain. The tragedy of the anticommons in patents and intellectual property rights thus appears when a user needs access to multiple patented inputs to create a single useful product. Each upstream patent allows its owner to set up another tollbooth on the road to product development, adding to the cost and slowing the pace of downstream innovation.

The severity of the “tragedy of the anticommons” problem in patents and intellectual property rights depends on the difficulty in negotiating with the upstream patent holders. The larger is the number of separate patent holders, the more severe is the tragedy of the anticommons. On the other hand, if the upstream patent holders are united to one entity to make decisions on the aggregate royalty, it would be much easier for the downstream producers to obtain the right of using all these patents and thus avoid the tragedy of the anticommons. For example, building a DVD player requires using hundreds of patented inventions. No company could ever build a DVD player if it had to negotiate with all patent holders and obtain their unanimous consent. These patents would be worthless due to gridlock. Fortunately, the owners of the patents used in building DVD players have formed a single entity authorized to negotiate on their behalf. But if you’re creating something new that does not have an organized group of patent holders, there will be “tragedy of the anticommons.”

As suggested by Heller and Eisenberg (1998), an anticommons in biomedical research may be more likely to endure than in other areas of intellectual property because of the absence of organized group of patent holders, the high transaction

costs of bargaining, heterogeneous interests among owners, and cognitive biases of researchers. In this case, policy-makers should seek to ensure coherent boundaries of upstream patents and to minimize restrictive licensing practices that interfere with downstream product development. Otherwise, more upstream rights may lead paradoxically to fewer useful products for improving human health.

The same logic can be applied to other high-tech industries besides of biomedical research. It is simply impossible to create a high-tech product these days without infringing on patents. For example, a new software or electronic device may use thousands of patents. It may not be practical even to discover all the possible patents involved, and it is certainly difficult to negotiate with thousands of patent holders individually. Therefore, overprotection on the patents or intellectual property rights may easily lead to a tragedy of anticommons for future product or technology development.

Tragedy of the Anticommons in Water Markets in the United States

Bretsen and Hill (2009) proposed another typical example of tragedy of the anticommons in water markets of the United States. Water shortages are becoming an increasingly important concern in much of the American West. Urban and environmental demands for water are increasing rapidly, and both physical and institutional constraints prevent new water supplies from being developed. With increasing demands for water for municipal, industrial, and environmental uses, transfers of water from agricultural irrigation to other uses will produce economic gains. A study of water transfers between 1987 and 2005 revealed dramatic differences in the value of water in urban uses versus agricultural uses (Brewer et al. 2008). Other estimates place the marginal value of water in municipal and industrial use at three to four times its marginal value in agricultural uses (Carey and Sunding 2001). However, in the face of increasing demand for water for urban and environmental uses in the western United States, water transfers out of agriculture have been fewer than one would expect

based on price differentials, and most of those that have occurred have required extensive negotiations. The reason for the difficulty in water transfer lies in the tragedy of the anticommons, since the existence of multiple rights of exclusion unbundled from the rights of use under the prior appropriation doctrine in the American West creates an anticommons that has impeded water transactions.

The multiple exclusion rights are the result of the evolutionary path of water institutions and the expansion of certain legal doctrines, such as the public interest and public trust doctrines which were originally designed to protect the property rights of people who used navigable waters and who depended on return flows from other irrigators.

The existence of multiple veto rights over water that are not bundled with use rights makes marketing of water difficult in the American West. In some cases no water trades will occur because of the multiplicity of veto rights; in other cases the tragedy of the anticommons leads to protracted and expensive negotiations with the need for side payments to secure the approval of all concerned. Furthermore, even after water transfers have been placed under contract, there is still the strong possibility that lawsuits can be brought to invalidate such transfers. The end result is an anticommons in which exclusion rights prevent the exercise of use rights and lock water into lower-value agricultural uses rather than allowing voluntary transfers into higher-value municipal, industrial, and environmental uses.

Tragedy of the Anticommons in Enterprise Licensing in China

As a developing country in the process of transition, China has inherited a complicated bureaucratic system of regulation from its past as a centralized planned economy. For several decades (which continues on today), almost all areas of business have been more or less under tight regulation. Routine enterprise licensing procedures involved multiple sign-offs and approvals, each with an independent right to reject any application. To get an investment project approved, an entrepreneur needs the permissions from all the

related government agencies in the chain of licensing procedures, such as the local governments' administration for industry and commerce, local environmental protection bureau, local health bureau, local construction bureau (for safety work concern), fire prevention guards, and so on.

Without bribery, a government licensing agency may benefit little from approving the project but may bear great responsibilities if the project they passed caused environmental, safety, or other accidents later. The government licensing agency faced with such an asymmetry between return and responsibility may simply choose to exercise its veto right to avoid the potential responsibilities in the future. So long as any one of the licensing agencies chooses to exercise its veto right, the entrepreneur's investment project will be rejected. Therefore, with the assumption of no bribery, an entrepreneur's project may be easily rejected when there are multiple licensing agencies, leading to a tragedy of the anticommons. As shown in Ying and Zhang (2008), the larger the number of licensing agencies in the procedure is, i.e., the higher the fragmentation of the licensing right is, the higher the possibility for the occurrence of the tragedy of the anticommons is.

In reality, the entrepreneur may bribe each licensing agency to get its project approved. However, a licensing agency may deliberately hold back its excluding right or suspend the project to ask for a higher bribery later. Meanwhile, the government officials in charge of the licensing process are running at the risk of getting caught as criminals by accepting the bribes and thus may ask for more compensations of their risk. Consequently, it is still difficult for an entrepreneur to get its project approved by each licensing agency even with bribery, due to the high transaction costs of bargaining, heterogeneous interests and asymmetric information among entrepreneurs and licensing agencies, and cognitive biases of the licensing agencies. It is not an exaggeration that entrepreneurs may have to spend months, or even years, to obtain the necessary approvals to start their investment projects (if not completely rejected), leading to another case of tragedy of the anticommons.

Since the 18th National Congress of the Communist Party of China, the Chinese government has been reforming the current bureaucratic licensing system to significantly reduce the number of administrative approval and licensing items, which should be good news to ease the tragedy of the anticommons in enterprise licensing.

Other Examples of Tragedy of the Anticommons

Besides of the above typical examples, there are many other cases of anticommons in reality. As argued by Hunter (2003), the tragedy of the anticommons does not just happen in the physical world but may also occur in cyberspace. Since the opening of the Internet to commercial exploitation in 1992, commercial operators have grown exceedingly fat, relying on the public character of the Internet and their easy access to the vast public commons on the Internet that was created before they ever arrived. However, by carving out remarkable new property rights online, the commercial operators are enclosing cyberspace. They have set up barriers to the use of their online resources, which may erode cyberspace's public commons. To obtain access to some online resources that are all necessary to create a new online innovation or product, a user may need the permissions from a bunch of existing commercial operators, each of whom is endowed with an excluding right. The difficult negotiations with each excluder may lead to a tragedy of digital anticommons.

Tragedy of the anticommons may also occur in the construction of municipal infrastructures and new building or during the urbanization process in transitional economies such as China. During the process of urbanization, or for the sake of improving municipal infrastructures, it might be necessary to construct new buildings in cities, roads, railroads, and similar transportation arteries, as well as other public facilities. Although the benefit to society from these constructions may be substantial, every single property owner along the way of construction must agree on the plan. This provides conditions for the tragedy of the anticommons, as even if

hundreds of owners agree, a single landowner can stop the new construction, e.g., the road or railroad. The ability for one person to veto the construction drastically increases the transaction costs and thus may cause tragedy of the anticommons. In some cases when negotiations with each excluder are extremely difficult, even some illegal violence instead is used to drive the "nail household," i.e., the ones who insist on staying without accepting the proposed compensations, out of the way, leading to brutal conflicts and other social tragedies.

Similarly, Heller (2008) indicates that the rise of the "robber barons" in medieval Germany was the result of the tragedy of the anticommons. Nobles commonly attempted to collect tolls on stretches of the Rhine passing by or through their fiefs, building towers alongside the river and stretching iron chains to prevent boats from carrying cargo up and down the river without paying a fee. The multi blockers along the river thus provided conditions for the tragedy of the anticommons.

Types of the Tragedy of Anticommons

Heller (1998) classified the tragedy of the anticommons into two types: legal anticommons and special anticommons. In a legal anticommons, substandard bundles of rights are allocated to competing owners in a normal amount of space, such as a storefront. By contrast, in a spatial anticommons, an owner may have a relatively standard bundle of rights but too little space for ordinary use.

In another way, Parisi et al. (2004) classified the tragedy of the anticommons into simultaneous and sequential anticommons. In the simultaneous case, various right holders exercise exclusion rights at the same time, independently. This may involve two agents linked in a coincident relationship, such as multiple co-owners with cross-veto powers on other members' use of a common resource. As an illustration, we can think of the choice of multiple land owners on whether to contribute their property for a joint venture development. Each parcel of land enters at the same

level as an input of production for the joint development. If all parcels of land are necessary for the realization of the development, the multiple parcel owners would effectively hold a cross-veto power for the realization of the joint project. In the sequential case, the exclusion rights are exercised in consecutive stages, at different levels of the value chain. The various right holders exercise exclusion rights in succession. This may involve multiple parties in a hierarchy, each of whom can exercise exclusion or veto power over a given proposition. As an example, a land owner grants building rights to a third party, who, in turn, grants a partial use right to another. If reunification of the land is desired, a vertical chain of contracting would be necessary. The grantor of the building rights would have to repurchase those rights from his/her grantee; in turn, the grantee of building rights would have to repurchase the partial use rights that he/she granted, and so on. The difficulty in reaching all these sequential repurchase agreements will cause a sequential tragedy of the anticommons. Both simultaneous and sequential anticommons problems are the result of nonconformity between use and exclusion rights.

A Formal Model of the Tragedy of Anticommons

Buchanan and Yoon (2000) proposed a formal economic model to explain the tragedy of the anticommons as a mirror image of the tragedy of commons. In their model, Q measures the quantity of usage, and P measures the average value of a unit product, where there is a linear relationship between P and Q as follows:

$$P = a - bQ \quad (1)$$

where a and b are positive constants. Consider, first, the two-person case, where A and B are to be assigned either (i) usage rights or (ii) exclusion rights. In either case, explicit collusion will allow for attainment of the efficient solution. In collusion, the optimal choice of Q is to maximize the total rent:

$$\text{Max}_Q PQ = (a - bQ)Q \quad (2)$$

The solution to this optimization problem is $Q^* = a/2b$, while the corresponding maximum rent (i.e., the total value of the products) is $a^2/4b$.

Now suppose that the required mutual trust is absent and joint action is impossible. In the first case, if each person is assigned a right to use the facility but cannot exclude the other from usage, the interaction will converge on an equilibrium that is analogous to Cournot-Nash duopoly. Person A chooses the level of usage, Q_1 , to maximize his/her rent, given person B's choice of usage, Q_2 . The rent to person A will be

$$\text{Max}_{Q_1} PQ_1 = (a - bQ_1 - bQ_2)Q_1 \quad (3)$$

Similarly, person B chooses the level of usage, Q_2 , to maximize his/her rent, given person A's choice of usage, Q_1 . The rent to person B will be

$$\text{Max}_{Q_2} PQ_2 = (a - bQ_1 - bQ_2)Q_2 \quad (4)$$

The Nash equilibrium of the maximization problems (3) and (4) yields $Q_1 = Q_2 = a/3b$, and the total usage $Q = Q_1 + Q_2 = 2a/3b$, which is larger than the efficient usage ($a/2b$) in the collusion case, leading to the overuse of the resource, i.e., the "tragedy of commons."

Alternatively, if each person is assigned a right to exclude but no independent usage right, then he/she can exercise this excluding right by setting the price of his/her ticket independently from the practice of the other owner. Let P_1 denote the price of person A's ticket and P_2 denote person B's ticket. Users are required to secure both tickets by paying the total price $P = P_1 + P_2$ for each unit of usage of the facility. Using Eq. 1, the quantity demanded by the users, Q , will be determined by

$$P_1 + P_2 = a - bQ. \quad (5)$$

A Nash equilibrium can be obtained by formulating a game in which each owner tries to maximize

his/her rent by setting his/her ticket price. Person A chooses P_1 so as to

$$\max_{P_1} P_1 Q = P_1 (a - P_1 - P_2)/b \quad (6)$$

From the first-order condition of this maximization problem, we can express P_1 as a function of P_2 :

$$P_1(P_2) = a/2 - P_2/2 \quad (6a)$$

Similarly, we can express P_2 as a function of P_1 :

$$P_2(P_1) = a/2 - P_1/2 \quad (6b)$$

Solving the system of simultaneous equations gives a stable solution, $P_1^* = P_2^* = a/3$. The price a user pays is $P^* = P_1^* + P_2^* = 2a/3$. The corresponding total usage is $Q = (a - P^*)/b = a/3b$, which is smaller than the efficient usage in collusion case, leading to an underuse of the resource, i.e., the “tragedy of the anticommons.”

Although the above model only includes two owners for simplicity, it can be generalized to any number of owners, while the main conclusion remains unchanged: in the case of commons when each owner is endowed with usage right but without excluding right, there will be overuse of the resource; in the case of anticommons when each owner is endowed with excluding right but without independent usage right, there will be underuse of the resource. It can be further proved that the larger the number of owners is, the less efficient the usage of resources will be, either in the case of commons or in the case of anticommons.

The above model also demonstrates that the tragedy of the anticommons is just a mirror image of the tragedy of commons. The equilibrium in either the multiple-user model or the multiple-excluder model is structurally analogous to that familiar in Cournot-Nash duopoly-oligopoly settings of interfirm competition. Competition among users, on the one hand, or among excluders, on the other, tends to reduce the total rents to the owners, which represents the efficiency of usage in the commons/anticommons model setting.

References

- Bretsen SN, Hill PJ (2009) Water markets as a tragedy of the anticommons. *William Mary Environ Law Rev* 33(3):723–783
- Brewer J, Glennon R, Ker A, Libecap G (2008) 2006 presidential address water markets in the west: prices, trading, and contractual forms. *Econ Inq* 46(2):91–112
- Buchanan JM, Yoon YJ (2000) Symmetric tragedies: commons and anticommons. *J Law Econ* 43(1):1–13
- Carey JM, Sunding DL (2001) Emerging markets in water: a comparative institutional analysis of the Central Valley and Colorado-Big Thompson Projects. *Nat Resour J* 41(2):283–328
- Hardin G (1968) The tragedy of the commons. *Science* 162(3859):1243–1248
- Heller MA (1998) The tragedy of the anticommons: Property in the transition from Marx to markets. *Harv Law Rev* 111(3):621–688
- Heller M (2008) The gridlock economy: how too much ownership wrecks markets, stops innovation, and costs lives. Basic Books, New York
- Heller MA, Eisenberg RS (1998) Can patents deter innovation? The anticommons in biomedical research. *Science* 280(5364):698–701
- Hunter D (2003) Cyberspace as place and the tragedy of the digital anticommons. *Calif Law Rev* 91(2): 439–519
- Parisi F, Schulz N, Depoorter B (2004) Simultaneous and sequential anticommons. *Eur J Law Econ* 17(2):175–190
- Ying Q, Zhang G (2008) Fragmentation of licensing right, bargaining and the tragedy of the anti-commons. *Eur J Law Econ* 26(1):61–73

Anti-dumping

Arianna Ciabattini and Roberto Ippoliti
Department of Management, University of Turin,
Turin, Italy

Definition

Anti-dumping measures are represented by differential duties aimed at countering dumping, which is a form of unfair competition by foreign companies achieved through international price discrimination.

The procedure for defining anti-dumping duties is long and complex and takes place under the supervision of the WTO.

Anti-Dumping Under the WTO

Anti-dumping measures – taken by a government or an international organization – are aimed at counteract the dumping practices that occur when a good or service is sold on foreign markets at a lower price, in some cases, even at its production cost. Therefore, dumping is a practice of unfair competition and international price discrimination.

However, anti-dumping measures, which are defensive tools of the national economy, are also susceptible of an offensive use by the country that imposes them and, for this reason, are regulated within the World Trade Organization (WTO) by art. VI of the General Agreement on Tariffs and Trade 1994 (GATT 1994). The WTO Member States may therefore apply these measures only under the circumstances provided for in Article VI and pursuant to investigations initiated and conducted in accordance with the provisions of the Agreement on Implementation of Article VI of the GATT 1994.

In particular, for the purpose of this Agreement, a product is to be considered as being dumped if it is introduced into the commerce of another country at less than its normal value, i.e., the export price of the product exported from one country to another is less than the comparable price, in the ordinary course of trade, for the like product when destined for consumption in the exporting country. The Agreement on Implementation sets out further criteria for the determination of dumping in cases where it is more difficult to quantify the normal value of a product.

This practice is to be condemned, in force of the Article VI, if it causes or threatens material injury to an established industry in the territory of a contracting party or materially retards the establishment of a domestic industry, referring to the domestic producers as a whole of the like products. The determination of prejudice is based on positive evidence and involves an examination of the volume and the effect on prices of the dumped imports in the domestic market for like products, and the impact of these imports on domestic producers of such products.

The initiation of an investigation to determine the existence, degree, and effect of any alleged dumping takes place normally upon a written application filed by the national industry or on its behalf, which, for the purposes of its validity, is necessarily supported by evidence of:

- Dumping
- Injury within the meaning of article VI of GATT 1994 as interpreted by its Implementation Agreement
- A casual link between the dumped imports and the alleged injury

The same evidence is required from the authorities concerned when, in special circumstances, they decide to initiate an investigation without having received the aforementioned written application.

It is up to the authorities to examine the adequacy and accuracy of the supporting material attached to the application in order to justify the initiation of an investigation which must be completed within 1 year, and in any case not later than 18 months from its opening.

All parties involved in an anti-dumping investigation are notified of the information required by the authorities and have ample opportunity to present in writing all the evidence they consider relevant, as well as the power to acquire non-confidential information. Exporters or foreign producers receiving an anti-dumping investigation questionnaire have a time limit for the reply of at least 30 days, which may be extended. In order to verify the material received or to obtain further details, the authorities may carry out the necessary investigations in the territory of other Members, if they obtain the consent of the Government and of the companies concerned.

For the duration of the investigation, interested parties are allowed to defend their interests. To this end, the authorities organize, upon request, meetings with the opposing parties to start a possible contradictory.

In order to neutralize or prevent dumping, each Contracting Party may receive, on any dumped product, an anti-dumping duty not exceeding the dumping margin for that product.

As a general rule, the authorities should determine the dumping margin on a case-by-case basis for each exporter or producer of the product under investigation.

Provisional or definitive measures may be authorized if the investigation is successful. The former takes the form of a provisional duty or, preferably, a guarantee – deposit in cash or security – equal to the amount of the provisional anti-dumping duty, which must not exceed the provisionally estimated dumping margin.

The decision to introduce an anti-dumping duty in cases where all the conditions are met, as well as to impose an anti-dumping duty equal to or less than the dumping margin, is taken by the authorities of the importer Member. Once applied to any product, the duty shall be levied, for the amount appropriate to the case and without discrimination, on all the relevant imports considered to be dumped and causing injury, whatever their origin.

An anti-dumping duty remains in force for the period of time and to the extent necessary to neutralize the dumping which is causing the injury. Any definitive anti-dumping duty should be revoked no later than 5 years after its imposition.

The European Anti-Dumping System

Council Regulation (EC) No 1225/2009 of 30 November 2009, and subsequent amendments, concerning the defense against dumped imports by countries not members of the European Union, transposes into the European law the anti-dumping rules contained in the Agreement on Implementation of Article VI of the GATT 1994. Specifically, the legislation contains provisions concerning the calculation of dumping, the procedure for the opening and subsequent conduct of investigations, for the imposition of provisional and definitive measures, and for the duration and review of anti-dumping measures.

In relation to anti-dumping duties, the European Union can choose to adopt one or more of the three basic forms:

- Ad valorem duty, the most frequent duty form, the amount of which is established in proportion to the net price of the goods at the EU border.
- Specific duty, which refers to the physical structure of the goods (weight, length, capacity, volume) and is related to it as a fixed value, e.g., 100 euros per ton of product.
- Variable duty, calculated on the basis of representative prices (the levy is in this case a variable element that can adapt to fluctuations in the international price).

The investigation to determine the existence, degree, and effect of the alleged dumping practices is opened following a written complaint lodged by any natural or legal person, or any association not having legal personality, acting on behalf of the EU industry. The complaint is considered to have been presented by the EU industry or on its behalf if it is supported by EU producers, which together account for over 50% of European production. The complaint can be brought to the Commission or to a Member State which sends it to the Commission, and it must contain evidence of the existence of dumping, injury, and causal link between the allegedly dumped imports and the alleged injury. It is examined in the Advisory Committee, composed of representatives of each Member State, and chaired by a representative of the Commission. If, after consulting the Committee, the evidence is sufficient to justify the initiation of an investigation, the Commission shall commence it within 45 days of the date on which the complaint was lodged. The investigation ends normally within 15 months of its initiation, with the closure or imposition of a definitive anti-dumping measure.

If the existence of dumping and of material injury results from the definitive findings of the facts and the interests of the EU are relevant, the EU Council, acting on a proposal submitted by the Commission after hearing the Advisory Committee, shall impose a definitive anti-dumping duty. As in the case of provisional measures, the amount of the anti-dumping duty must not exceed the dumping margin and should be less than this margin, if a lower amount is sufficient to eliminate

the injury caused to the EU industry. The duty shall be established for the amount appropriate to each case and without discrimination on imports of dumped products. The regulation imposing the duty specifies the amount imposed on each supplier or, if this is not possible, to the supplier country concerned.

The amount of the duty depends on the established value of the dumping margin. This is given by the difference between the normal value and the export price of a product.

For the purpose of determining the normal value of a product and the dumping margin, it is necessary to distinguish between States whose economic system is assessed as being based on a market economy and countries that are not considered as such. The normal value of a product originating in a non-market economy country is identified by referring to a third country governed by a market economy, for example, comparing China to Brazil's production costs. Thanks to this procedure the EU has active anti-dumping duties for about 50 Chinese products.

Anyway, the Section 15 of the Chinese WTO Accession Protocol of China, which attributes it the non-market economy status, expired in December 2016 but EU refused to grant China market economy status. Thus, during its plenary session in Strasbourg on November 15, 2017, the EU Parliament passed a new anti-dumping rule. The ex-ante definition between market or non-market economy disappears: all countries are equal. And China is like the other members of the WTO. But a punctual mechanism is introduced to defend against competitors who distort markets (anyone, not just China). If EU companies or trade unions or other stakeholders have suspicions of possible distortions, they can report to the Commission, which initiates an investigation and draws up a report. If it is established that the distortions are really there, then the old method of the analogue country can be applied.

With two substantial differences compared to the previous system. The first is that it will be possible to be selective, distinguishing also between sectors and not just between countries. It will thus be possible to treat Chinese exports not subject to excessive distortions such as those of

any market economy. The second is that the principle of distortion has been considerably enlarged, including social and environmental dumping. It will be argued that there is a distortion if workers are not treated according to ILO criteria or if environmental conventions are not respected. These rules reduce the most unfair sources of unfair competition but open the door to greater randomness in the identification of distortion.

Cross-References

- [European Integration](#)
- [Preferential Tariffs](#)
- [State Aids and Subsidies](#)
- [TRIPS Agreement](#)
- [WTO: Procedural Rules](#)

References

- European Council (2009). Council Regulation (EC) No 1225/2009
- European Parliament (2017). Legislative resolution of 15 November 2017 on the proposal for a regulation amending Regulation (EU) 2016/1036 and Regulation (EU) 2016/1037
- World Trade Organization (1994). Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994
- World Trade Organization (1947). General Agreement on Tariffs and Trade (GATT 1947), Article VI
- World Trade Organization (2001). The WTO Accession Protocol of the People's Republic of China (2001), Section 15

Asylum Law

Marc Deschamps¹ and Jenny Helstroffer²

¹CRESE, Univ. Bourgogne Franche-Comté, Besançon, France

²BETA, Université de Lorraine, Nancy, France

Abstract

The purpose of this essay is to indicate that asylum law essentially derives from international commitments made in the aftermath of

the Second World War. We then present, by way of illustration, the structure and main mechanisms of the common European asylum system. Finally, we will present the analyses of some of the economic studies dealing specifically with the question of asylum law.

Synonyms

Refugee law

The Economic Analysis of Asylum Law

Immigration is the source of passionate controversies throughout the world among populations and the governments. The issue is tense on all continents, and it has substantial, though varying, effects on electoral results and policies. For example, the results of the Standard Eurobarometer 85 (2016) of May 2016 show that 48% of European Union citizens consider immigration to be the main European problem, thus ranking higher than terrorism (39%), the economic situation (19%), public finances (16%), and unemployment (15%). It is the most frequently cited concern in 20 Member States, with peaks in Estonia (73%) and Denmark (71%). However, at the national level, the issue of immigration ranks second to unemployment, and it is of minor importance when Europeans are asked about the problems with which they are confronted individually.

Three clarifications are necessary here. First, a distinction must be made between immigrants and asylum seekers. Indeed, the latter are by definition person who apply for refugee status on the grounds that he (or she) is at a significant risk to his safety or to his life in his country of origin. Asylum seekers are therefore to be distinguished from persons who migrate primarily for economic or family reasons. Secondly, according to United Nations High Commissioner for Refugees (UNHCR), more than 65 million people worldwide have been forced to flee their homes, including more than 21 million persons who have left their country of origin. Sixty-eight percent are hosted in Africa and the Middle East. Thirdly,

the reduction of legal channels of immigration since the 1970s leads to an increasing pressure on asylum mechanisms.

In order to better understand asylum law, we will present the international instruments on which it is based (1), describe the structure of the common European asylum system (2), and present the economic literature on asylum (3).

International Instruments Concerning Asylum Law

The population movements connected with the Russian Revolution, the collapse of the Austro-Hungarian and Ottoman empires, and the rise of totalitarianism at the beginning of the twentieth century led the League of Nations and the International Labor Organization to examine the question of millions of displaced or endangered people. One of the strongest acts was undoubtedly the introduction of a passport allowing stateless refugees to travel, on the initiative of Fridtjof Nansen, the first High Commissioner for Refugees of the League of Nations. However, the four main international instruments that currently define asylum law appeared only after the Second World War.

The Universal Declaration of Human Rights was adopted on December 10, 1948, by the 58 member states which then constituted the United Nations General Assembly. Article 13 provides that “every person has the right to leave any country.” Most importantly, article 14 states in its first paragraph that: “Everyone has the right to seek and to enjoy in other countries asylum from persecution.”

It is nevertheless with the Convention of 28 July 1951 relating to the status of refugees (the so-called Geneva Convention) that a true right to asylum first arises. Indeed, the Geneva Convention still presents the centerpiece of the legal definition of the right to asylum for signatory states, since article 1 defines as a refugee any person who “[...] owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country

of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country; or who, not having a nationality and being outside the country of his former habitual residence as a result of such events, is unable or, owing to such fear, is unwilling to return to it.” The remainder of the text defines the rights and duties of refugees as States, in particular article 31, under which States undertake not to apply penal sanctions against refugees entering or remaining on their territories without authorization, provided that they report to the authorities without delay and explain their irregular entry or presence.

The Protocol of 31 January 1967 on the Status of Refugees (the so-called New York Protocol) generalizes the scope of the Geneva Convention by lifting all time and space limitations, since initially it was applicable only to events occurring before 1 January 1951 either “in Europe” or “in Europe and elsewhere.”

Finally, in order complete the multifaceted picture of asylum law, we must add the United Nations Convention on the Rights of the Child of 20 November 1989. Its article 22 deals specifically with refugees below 18 years of age, whether accompanied or not, and with other provisions relating in particular to family reunification, deprivation of liberty, or criminal protection.

The Common European Asylum System

The question of asylum is an issue that concerns all countries. Here we present the current architecture set up by the European Union to deal with this issue.

According to data provided by Eurostat in April 2016, in 2015, the European Union with 28 members experienced almost twice as many asylum seekers compared with the figure recorded in 1992 with 15 members. According to UNHCR, there were nearly 15,000 deaths solely in the Mediterranean between 2011 and 2016. Faced with the scale of this tragedy and to avoid the destabilization of certain countries of entry (such as Greece or Italy), the European Union decided to implement a common European asylum system

(CEAS) at the Tampere summit in 1999. It is outlined in article 78 of the Treaty on the Functioning of the European Union.

After a preparatory period of introduction of minimum standards in European asylum law, the CEAS aims to achieve uniform speed, impartiality, quality, and protection for all those seeking protection in all Member Countries. The objective is to apply the following chronology: (1) the asylum seeker arrives on European territory and faces a harmonized procedure for the entire territory of the EU, (2) he (or she) is granted accommodation and means of subsistence during the processing of the application, (3) if he is at least 14 years old, his fingerprints are recorded and sent to the Eurodac database in order to determine which Member State is responsible for the application; (4) the asylum seeker is granted an interview to determine whether he is eligible for refugee status, subsidiary protection, or temporary protection; and (5) the procedure ends with either a refusal of his application (in which case the claimant is granted an appeal) and an order to leave the territory or the request is accepted and the person is granted a residence permit or citizenship, access to the labor market, and health care.

Legally, the CEAS is a set of legislative texts laying down common standards and procedures. Today it is essentially composed of two Regulations and three Directives: (a) Regulation 604/2013 (known as “Dublin III”) laying down the criteria and mechanisms for determining the Member State responsible for the examination of an application for international protection lodged in one of the Member States by a third-country national or stateless person, (b) Regulation 603/2013 creating Eurodac for the comparison of fingerprints of asylum seekers, (c) Directive 2011/95/EU (the so-called Qualification Directive) laying down the conditions for benefiting from protection; (d) Directive 2013/32/EU on the granting and withdrawal of international protection, and (e) Directive 2013/33/EU (the so-called Reception Directive) governing the reception of asylum seekers in EU countries.

These texts are complemented by Regulation 439/2010 establishing the European Asylum

Support Office, whose task is to strengthen cooperation between Member States and to help them in times of crises, as well as Regulation 516/2014 establishing the Asylum, Migration, and Integration Fund.

Asylum in the Prism of Economic Analysis

The treatment of asylum by economists is the subject of both theoretical and empirical and positive and normative analyses. Without providing a complete account, we will now present some of its salient aspects.

Explaining asylum migration flows Unlike labor migration flows, for which expected differentials between countries of origin and destination are major determinants, for explaining asylum migration flows, absolute conditions in the country of origin carry a much greater weight. Empirical studies (see Hatton (2011) for a review) such as Schmeidl (1997) show that push factors, and especially political violence, are particularly important predictors of the generation of refugee flows, unlike economic variables. While this is a feature that distinguishes refugee migration from economic migration, a certain element of choice of the destination is common to both.

Determinants of destination can be grouped into factors influencing the migration path, such as geography and migration costs, and destination country characteristics, such as GDP, refugee stocks, and relevant policies. The latter include visa policies, asylum application recognition rates (see, e.g., Toshkov 2014), the speed of decision on asylum applications, labor market access of asylum applicants and refugees, asylum procedures, acceptance rates of asylum applications, expulsion rates, and living conditions of asylum applicants. Different combinations of these migration determinants have been analyzed in theoretical models of destination country choices by Czaika (2009), Schaeffer (2010), and Djajić (2014). However, Keogh (2013) shows that destination country-specific characteristics explain over 70% of the overall variation in asylum applications.

Asylum lawmaking Economics scholars have also studied normative policy-making aspect of asylum. Thus, Helstroffer and Obidzinski (2010) develop a theoretical model to predict which actors benefit from moving asylum law from the national to the EU level and from the common European asylum system. They find that in a context of regulatory competition, host states do not benefit from the Europeanization of asylum law because of the loss of discretionary power. Refugees however can benefit from common minimum standards, though not from the latest step of the development, which aims at total harmonization.

Helstroffer and Obidzinski (2014) show that the asylum lawmaking procedure in the European Union, which is based on codecision of the Council and the European Union, favors the institution least favorable to change and gives the Commission crucial influence over the outcome of the lawmaking procedure. It is Kaldor-Hicks inefficient.

Bubb et al. (2011) propose a model in which the 1951 Geneva Convention is a means of resolving the problem of the provision of the global public good of refugee protection between states. They consider two possible policies to resolve the issue of filtering between refugees and economic migrants: reforms that make states less attractive for potential immigrants and monetary transfers. They show that countries with high incomes should subsidize those with low incomes in order to avoid the problem of spatial concentration of refugees in southern states.

Fernandez-Huertas and Rapoport (2015) propose to create a market for exchangeable asylum quotas (i.e., countries participate in the public good of international protection either by visas or monetary contributions) coupled with a matching mechanism linking the preferences of asylum seekers and those of host countries. The study compares its recommendations with the EUREMA (European Relocation from Malta) program initiated in 2009 by the European Commission. The authors conclude that their proposal would be a “perfect instrument” to reallocate asylum seekers in the EU.

Conclusion

The question of asylum has long been the subject of specific analyses by philosophers, legal scholars, political scholars, psychologists, geographers, sociologists, and anthropologists. Since the turn of the century, economists have started to contribute to the understanding of asylum migration flows through empirical study and models of migration options. They have further studied the legislators' responses to asylum flows, with the objective of identifying a system of hosting asylum seekers that is in the interest of both host countries and refugees.

In our view, there are two main avenues to be pursued: firstly, a better understanding of the observable and non-observable characteristics of asylum seekers and their choices (as Alvin Roth pointed out in 2015 Article *Politico*: “refugees are not widgets”) and, secondly, to find permanent and socially accepted mechanisms to guarantee the effective existence of the right to asylum. We are thus sure that in the years to come, economic studies will complement those of the other human sciences in order help find a solution to one of the most pressing humanitarian problems of our times.

Cross-References

- [Asylum Law](#)
- [Court of Justice of the European Union](#)
- [Immigration Law](#)

References

- Bubb R, Kremer M, Levine D (2011) The economics of international refugee law. *J Leg Stud* 40(2):367–404
- Czaika M (2009) The political economy of refugee migration and foreign aid. Palgrave Macmillan
- Djajić S (2014) Asylum seeking and irregular migration. *Int Rev Law Econ* 39:83–95
- Fernandez-Huertas M, Rapoport H (2015) Tradable refugee-admission quotas and EU asylum policy. *CESifo Econ Stud* 61(3–4):638–672
- Hatton T (2011) Seeking asylum: trends and policies in the EU. Centre for Economic Policy Research
- Helstroffer J, Obidzinski M (2010) Optimal discretion in asylum lawmaking. *Int Rev Law Econ* 30(1):86–97

- Helstroffer J, Obidzinski M (2014) Codecision procedure biases: the European legislation game. *Eur J Law Econ* 38(1):29–46
- Keogh G (2013) Modelling asylum migration pull-force factors in the EU-15. *Econ Soc Rev* 44(3):371–399
- Schaeffer P (2010) Refugees: on the economics of political migration. *Int Migr* 48(1):1–22
- Schmeidl S (1997) Exploring the causes of forced migration: a pooled time-series analysis, 1971–1990. *Soc Sci Q* 78(2):284–308
- Standard Eurobarometer 85 (2016) Public opinion in the European Union, May, Brussels, 221 p
- Toshkov DD (2014) The dynamic relationship between asylum applications and recognition rates in Europe (1987–2010). *Eur Union Polit* 15(2):192–214

Asymmetric Information in Litigation

Shay Lavie
The Buchmann Faculty of Law, Tel Aviv
University, Tel Aviv, Israel

Definition

Information asymmetries in litigation refer to situations in which one party is privately informed concerning an issue that is relevant to the outcome of the legal process. In particular, information asymmetries are widely conceived to be an obstacle to settlements. They are prevalent in the legal discourse and have been extensively analyzed in the last decades through game-theoretical models. This entry briefly discusses the theoretical models of asymmetric information in litigation, their underlying assumptions, real-world applicability, and possible avenues to bridge informational gaps.

Introduction

Litigation is costly. We would therefore expect the parties to seek a contractual solution that saves their joint litigation costs – a settlement. Most of the filed cases indeed settle, but some are not (Spier 2007, p. 268). The literature has offered a variety of reasons to explain settlement failures, among

which asymmetric information is considered a major obstacle to settlements.

Asymmetric information in litigation refers to situations in which one party is privately informed concerning an issue that is relevant to the outcome of the legal process. It can stem from many reasons (Spier, p. 272). Plaintiffs may have private information regarding their level of damages; defendants may have private information concerning their fault. Asymmetric information can have various other sources. Parties could be privately informed as to the quality of their lawyers and their capacity to withstand a long trial. Intuitively, under information asymmetries, the uninformed party could not evaluate its case, and trials are more likely.

Theoretical Perspectives

There are several common perspectives to theoretically analyze the effects of asymmetric information on settlements. In a now-classic paper, Bebchuk (1984) modeled asymmetric information as a game in which the uninformed litigant makes a single take-it-or-leave-it offer to the informed litigants. Informed parties decide whether to accept or reject that offer by comparing it to their expected utility from trial, i.e., the expected judgment and the trial costs they likely to incur should they reject the offer. In this so-called screening model, the informed parties whose case is relatively strong – beyond what the single offer represents – reject the offer and go to trial. Similarly, informed litigants whose case is relatively weak would accept the offer. In this sense, the offer creates a cutoff that “screens” different types of informed litigants. The uninformed litigant sets that cutoff according to its gain from the settlement offer and the risk that the offer would be rejected. The screening model can thus predict both the proportion of cases that go to trial and their average type – note that the cases that made it to trial under this model are those in which informed litigants are more confident regarding their case. This basic screening model was developed and extended by other papers (Spier 2007, pp. 273–275, provides a brief survey).

Reinganum and Wilde (1986) present another influential direction to analyze asymmetric information situations. In their model, it is the informed party who makes a take-it-or-leave-it offer to the uninformed, and that offer can “signal” to the uninformed the strength of the informed litigant’s case. Reinganum and Wilde’s signaling model results in an elegant, fully separating equilibrium. In this equilibrium, informed litigants propose a truthful offer, i.e., an offer corresponding to their expected judgment, and the uninformed mixes between accepting the offer and rejecting. In order to prevent weak informed types from masking as stronger ones, the rate of acceptance that the uninformed type employs in this model should depend on the offer. To demonstrate, where the defendant holds private information, the plaintiff should more frequently reject low offers. Knowing that the plaintiff tends to reject low offers, weak defendants would hesitate to mimic as strong ones, i.e., offer a settlement lower than their type. As before, the basic signaling model predicts both the proportion of cases that fail to settle and their merits, where the cases that made it to trial are not representative of the entire population.

There is now a considerable body of game-theoretical literature that discusses asymmetric information in litigation, and it is important to highlight some of the common assumptions that underlie these models (e.g., Daughety and Reinganum 2014a, pp. 84–86; Bone 1997, pp. 567–571). First, this literature assumes that aside from the private information, all relevant parameters are common knowledge. Moreover, while the uninformed litigant does not know the strength of its rival litigant, it does know the background distribution of the different potential types of its rival, i.e., the probabilities that the rival’s case is of a certain strength. This is a nontrivial description, which facilitates the theoretical analysis of asymmetric information. Second, the legal proceedings are assumed to resolve at least some of the informational gaps, for instance, through testimonies or pretrial discovery. However, this revelation process is costly. The foregoing literature could therefore be described as modeling the bargaining that

precedes costly revelation. Hence, the asymmetric information story does not fit situations in which the private information could be conveyed cheaply before trial, e.g., through credible disclosure by the informed or by the uninformed litigant's pre-filing investigation (Bone 1997). There can be a variety of other theoretical complications. For example, most of the papers depict simple situations in which one litigant is informed and the other is not; in reality, both parties can have private information regarding different issues (for instance, Daughety and Reinganum 1994).

Asymmetric Information in Real-World Litigation

Given this discussion, one wonders whether asymmetric information situations are prevalent in real-world settings. As in actuality most of the cases settle, rigorous empirical investigation is quite complicated. One strategy is to focus on the characteristics of cases that made it to trial. As noted above, asymmetric information models predict that the cases that fail to settle are a non-random sample of the entire population. Particularly, the predictions of asymmetric information models could be compared to competing theories of settlements and in particular to the influential models that assume that both parties diverge on their expectations regarding the outcome at trial (e.g., Priest and Klein 1984). Hylton and Lin (2012) survey empirical papers according to this logic and conclude that asymmetric information appears to be relevant in pretrial settlements, though competing theories are also confirmed by the literature. They also caution that "the empirical work so far has to be considered preliminary" (Hylton and Lin 2012, p. 505). There is also some experimental analysis that uses ultimatum games to test the screening and signaling models of asymmetric information in litigation. This literature has generated mixed results, particularly with regard to the signaling setting (where the informed party makes the offer) (Pecorino and van Boening 2016).

Beyond robust empirical findings, the notion of asymmetric information in litigation is quite prevalent in the current legal discourse, at least in the USA. Asymmetric information is sometimes referred to by law professors as "[t]he most important problem in dispute resolution" (Rhee 2009, p. 548). The academic interest in asymmetric information has risen following several decisions of the US Supreme Court that encourage judges to assume a greater gate-keeping role (*Bell Atl. Corp. v. Twombly*, 550 U.S. 570 (2007)). These precedents direct trial courts to screen out – at the outset of litigation and before proceeding to discovery – cases that do not present a sufficient factual threshold. These stricter rules allegedly deter meritorious filings, particularly in areas where the information about the merits of the case is thought to reside with the defendant – asymmetric information settings. Salient examples are medical malpractice and employment discrimination cases (Hubbard 2016; Bone 2009).

Bridging Informational Gaps

To the extent parties can somehow transmit information, the importance of asymmetric information diminishes. Given the notion that asymmetric information is common, and the recent legal trend to make filing cases harder, one may wonder how parties could bridge informational gaps. A straightforward way to convey information is voluntary disclosure; alternatively, formal, court-supervised discovery proceedings are aimed at forcing the informed to reveal its information (Shavell 1989; Farmer and Pecorino 2005). There are more subtle alternatives. Information can be conveyed indirectly, through the employment of various litigation tactics (Daughety and Reinganum 2014a, pp. 90–92). For example, recent papers have shown how litigants can use various litigation features such as filing for costly injunctions (Jeitschko and Kim 2012) or investing in observable pretrial preparation (Choné and Linnemer 2010) to credibly signal the value of their case. The use of intermediaries such as

attorneys (Leshem 2009) and litigation funders (Daughety and Reinganum 2014b; Avraham and Wickelgren 2014) can likewise play an indirect role in facilitating settlements under asymmetric information. More generally, recent papers attempt to show the extent to which parties can signal information through employing costly signals, e.g., filing fees (Hubbard 2017), or by committing to augment the judgment should the rival party win (Lavie and Tabbach 2017).

Conclusion

Asymmetric information – settings in which one of the parties enjoys better information – is seemingly a prevalent phenomenon in actual litigation. Game-theoretical models constitute an effective analytical tool to understand pretrial bargaining and settlements under informational asymmetries. In the last decades rich literature has developed the basic models, adding a considerable theoretical depth. However, some of the underlying assumptions of the theoretical models seem questionable under the current litigation landscape, and the empirical applicability of the theory to the field is a direction for future research.

Cross-References

- [Access to Justice](#)
- [Information Disclosure](#)
- [Legal Disputes](#)
- [Litigation Decision](#)
- [Signalling](#)
- [Third-Party Litigation Funding](#)

References

- Avraham R, Wickelgren A (2014) Third-party litigation funding—a signaling model. *DePaul Law Rev* 63:233–264
- Bebchuk LA (1984) Litigation and settlement under imperfect information. *RAND J Econ* 15:404–415
- Bone RG (1997) Modeling frivolous suits. *Univ Pennsylvania Law Rev* 145:519–605
- Bone RG (2009) Twombly, pleading rules, and the regulation of court access. *Iowa Law Rev* 94: 873–936
- Choné P, Linnemer L (2010) Optimal litigation strategies with observable case preparation. *Games Econ Behav* 70:271–288
- Daughety AF, Reinganum JF (1994) Settlement negotiations with two-sided asymmetric information: model duality, information distribution, and efficiency. *Int Rev Law Econ* 14:283–299
- Daughety AF, Reinganum JF (2014a) Revelation and suppression of private information in settlement-bargaining models. *Univ Chicago Law Rev* 81:83–108
- Daughety AF, Reinganum J (2014b) The effect of third-party funding of plaintiffs on settlement. *Am Econ Rev* 104:2552–2566
- Farmer A, Pecorino P (2005) Civil litigation with mandatory discovery and voluntary transmission of private information. *J Leg Stud* 34:137–159
- Hubbard WHJ (2016) A fresh look at plausibility pleading. *Univ Chicago Law Rev* 83:693–757
- Hubbard William HJ (2017) Costly signaling, pleading, and settlement. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No. 805; U of Chicago, Public Law Working Paper No. 622. Available at SSRN: <https://ssrn.com/abstract=2947302>
- Hylton KN, Lin H (2012) Trial selection theory and evidence. In: Sanchirico CW (ed) *Encyclopedia of law and economics—procedural law and economics*, vol 8, 2nd edn. Edward Elgar, Cheltenham/Northampton, pp 487–506
- Jeitschko TD, Kim B-C (2012) Signaling, learning, and screening prior to trial: informational implications of preliminary injunctions. *J Law Econ Org* 29: 1085–1113
- Lavie, Shay, Avraham Tabbach (2017) Judgment contingent settlements https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3000218
- Leshem S (2009) Contingent fees, signaling and settlement authority. *Rev Law Econ* 5:435–460
- Pecorino, Paul, van Boening, Mark (2016) An empirical analysis of the signaling and screening models of litigation. Available at SSRN: <https://ssrn.com/abstract=1988244>
- Priest GL, Klein B (1984) The selection of disputes for litigation. *J Leg Stud* 13:1–55. 1984
- Reinganum J, Wilde L (1986) Settlement, litigation, and the allocation of litigation costs. *RAND J Econ* 17:557–566
- Rhee RJ (2009) Toward procedural optionality: private ordering of public adjudication. *N Y Univ Law Rev* 84:514–571
- Shavell S (1989) Sharing of information prior to settlement or litigation. *RAND J Econ* 20:183–195
- Spier KE (2007) Litigation. In: Mitchell Polinsky A, Shavell S (eds) *The handbook of law and economics*. Amsterdam-Oxford, North-Holland, pp 259–342

Audit Committees

Patrick Velte

Leuphana University, Lüneburg, Germany

Abstract

A profound international competition between corporate governance and corporations constitutions systems has been going on since the middle of the last century (La Porta et al. 2002, p. 1147). A basic categorization has been made with regard to the ratio between internal and external corporate governance as well as to the management and supervising structure of publicly owned firms (one-tier and two-tier system). Amongst others, a partial convergence of both constitutional models indicates a high acceptance of audit committees in both systems of corporation's constitutions. However, the committee's competences are different in the one- and two-tier system as well as the main motives of their implementation. Within the two-tier system, the audit committee has been implemented to support and relieve the supervisory board in preparing various tasks. In addition the committee is expected to strengthen corporate governance as a consequence of the high number of supervisory board members. Moreover, the appointment of financial experts as audit committee members is to counteract the lack of respective knowledge in the supervisory board. In contrast, the one-tier system is by trend forcing a stronger personal separation between executive and nonexecutive directors in the board. In addition, major importance is placed on the independence of the committee members in the one-tier system which is usually symptomatic for the separation of functions within the two-tier system. As with the example of audit committees, it becomes clear that both models try to use the advantages of the respective constitutional systems. However, a general superiority of one system cannot be concluded.

The aim of the present analysis is to provide an overview of empirical survey results with

regard to the acceptance of audit committees on the capital market and the influence of audit committees on corporate governance. Major attention is paid to a statistically proven relation between certain corporate governance variables and the implementation of audit committees, especially with regard to the independence and financial expertise of its members. The German stock corporation law will be used as an example to demonstrate the importance of audit committees within the two-tier system. Similarly, the US-American capital market with its particular regulations of the stock exchange commission will be used for the one-tier system.

Definition

Audit committees are part of the board of directors (one-tier system) or the supervisory board (two-tier system). They play a major role in modern corporate governance to supervise the management, the risk management (system), and the external auditor. The following analysis gives an overview of empirical survey results with regard to the acceptance of audit committees on the capital market and the influence of audit committees on corporate governance. To compare the different status in the one-tier and the two-tier system, audit committees in German- and US-American-listed companies will be presented.

Normative Analysis

Germany (Two-Tier System)

Implementation

The discussion regarding the implementation of audit committees to enhance corporate governance grew more intense with the "control and transparency law" in 1998. Amongst others, the empirical survey of Coenenberg et al. (1997) supports this relatively young movement deriving from a scientific economic source. The majority of the management board members of the 100 top German corporations in question were not aware

of the necessity to implement audit committees in 1995. In contrast, the survey of Quick et al. (2008) was able to prove an implementation quota of 100% for the DAX30 and 86% for the MDA-X. Hence, the present survey results suggest that the majority of the listed stock corporations in the German prime standard have implemented audit committees to strengthen corporate governance.

In contrast to the USA, the German stock corporation law and commercial law have not yet stipulated a general, legally binding obligation for the implementation of audit committees. In fact, the decision to implement audit committees is subject to the autonomy of the supervisory board in terms of § 107 III 1 AktG. This voting right has already been part of the stock corporation law of 1937 and was reinforced by further reform act. The national legislator on purpose did not include the obligation to implement audit committees in order to guarantee highest flexibility with regard to the corporation's management. However, the demand for due diligence of the supervisory board accounting for an appropriate organization of its activities will lead to the implementation of audit committees with rising number of board members. Without audit committees, the necessary intensity of the supervision is no longer ensured. Consequently, the corporations' voting right to implement audit committees becomes redundant with rising number of supervisory board members. Accordingly, this fact gives reasons for the recommendation in the German Corporate Governance Code (GCGC). According to this recommendation, the implementation of qualified audit committees should depend on the specific circumstances of the company as well as its numbers of members. Since its introduction, the GCGC explicitly advises the implementation of audit committees. In case the supervisory board is composed of only three to six members, the prevailing opinion allows for an abandonment of the implementation of audit committees. In this case, no explanation according to § 161 AktG is required, since such small supervisory board usually would not relate to the implementation of an audit committee.

The audit committee has been explicitly mentioned for the first time within the context of the

commercial and stock corporation law since 2009. However, a general obligation to implement audit committees is still missing. In principle, only capital market-oriented stock corporations in terms of § 264d HGB that do not have a supervisory board with the respective job specifications are obligated to implement audit committees with at least one independent financial expert. Since all tasks of the audit committee may also be fulfilled by the plenum of the supervisory board, all stock corporations that are legally forced to implement supervisory boards still hold a voting right concerning the implementation of audit committees. Hence, the national legislator relies on the empirically proven high quota in complying with the GCGC.

Job Specification

The matter of independence is implemented in the German stock corporation law in § 105 I 1 AktG. Thus, a member of the audit committee is not allowed to be an active management board member or permanent deputy, authorized officer, or a general agent authorized for the entire corporations management at the same time. In addition and according to the prohibition of crosswise intersection, a member of the audit committee is not allowed to be a legal representative of a dependent company or of another corporation that engages a management board member of the corporation in question in the supervisory board. These regulations are common practice in the German two-tier system. Therefore, the audit committee needs to evolve from the supervisory board, and its members are not allowed to fulfil any managerial functions. In accordance with § 264d HGB, capital market-oriented stock corporations need to appoint at least one independent member in the supervisory board or audit committee. However, this is the only article with regard to the term "independence" so far. In fact, the recommendation of the EU commission of the 15th of February 2005 can be classified as a general guidance. A cooling off period of 2 years for former management board members to become supervisory board members of listed stock corporations is advised in 2009. An exception is granted, if shareholders holding more than

25% of the voting rights of the corporation are in favor of the nomination.

In addition to the stock corporation law standards, the GCGC recommends that supervisory board and hence audit committees should be composed of an adequate number of independent members. Thus, a member is independent if he has no commercial or personal relation to the corporation or its management that accounts for a conflict of interests. Furthermore, the GCGC advises a nomination of no more than two former management board members for the supervisory board. Moreover, the GCGC suggests that the present supervisory board chairman should not take the chair of the audit committee. However, the chairman of the audit committee should be independent. The cooling off period of two years for former management members to become audit committees chairman should be strictly adhered to. A missing compliance with the before-mentioned code suggestions will not account for a justification with regard to the conformity declaration, since the compliance statement only relates to recommendations.

In addition to the independency, the job specification of the audit committee emphasizes on the financial expertise of its members. In terms of § 100 I AktG, no specialist knowledge is mentioned explicitly. However, a minimum level of common, economic, organizational, and judicial knowledge, necessary for understanding and appropriately judging on all regular business transactions unassisted, is demanded (BGH 1982, p. 991). Nonetheless, financial expertise is not mentioned explicitly. At least one member of the audit committee is expected to have the necessary expert knowledge with regard to accounting or auditing (financial expert). Yet, no comment is made on whether and in how far the audit committee chair is to be involved in this part.

Similarly, the GCGC only recommends that the audit committee should be composed of members that are able to fulfil all tasks with the required knowledge, skills, and professional experience at all times. Though, the GCGC provides a detailed job description of the audit committee's chairman. According to this, the audit committees chairman is expected to have special knowledge and

experience with regard to the application of accounting standards and internal control procedures. Consequently, the GCGC expects the chairman to be a financial expert, whereas the national legislator only demands for compliance with the legal minimum requirements.

USA (One-Tier System)

Implementation

The implementation of audit committees on the US-American capital market was first recommended in 1939/1949 by the Securities and Exchange Commission (SEC) and the New York Stock Exchange (NYSE). Since corporations did not put this recommendation into effect in the following years, the American Institute of Certified Public Accountants (AICPA) (1967) renewed and enhanced the recommendations of the SEC. Within this context, the composition of audit committee members and their tasks were discussed for the first time. A liability law case in 1968 led to a vote for an obligatory disclosure in the proxy statement with regard to the implementation of audit committee and its members by the SEC (1974). In addition to the name of the members, the disclosure of the number of meetings and their main tasks and responsibilities became obligatory on the 1st of July 1978. Since that time, it became mandatory for all listed corporations at the NYSE to implement an independent audit committee. This was stipulated by the SEC (1978). The American Stock Exchange (AMEX) followed in 1980 and the National Association of Securities Dealers Automated Quotations (NASDAQ) in 2001. In 1987, the results of the National Commission on Fraudulent Financial Reporting became public, also emphasizing the importance of audit committees regarding the corporation's supervision. Within this context, the national commission recommended the implementation of audit committees for all publicly owned firms. The "Blue Ribbon Report" went along with this after a couple of years in 1999. The Sarbanes Oxley Act stipulated an implicit obligation for the implementation of audit committees as permanent committees of the board of directors for all corporations listed at a US stock

exchange. In addition, the job specification of the audit committee's members was described in detail. Opposed to German stock corporation law, the corporations in question do not have an option with regard to the implementation of audit committees.

Job Specification

According to the Sarbanes Oxley Act, all members of the audit committee have to be financially and personally independent of the corporation's management (Section 301). The term independent is applicable only if no direct or indirect corporate or affiliate payment is collected by an audit committee member. The regulations of the Sarbanes Oxley Act are of extraterritorial nature. Hence, the rules of financial independence would only hardly be applicable in countries with internal employee participation (e.g., German corporations that are secondary listed at a US-American stock exchange). The codetermination of the supervisory board would be dependent in terms of their salary. In order to preserve the extraterritorial effect, the SEC is expecting only managing employees to comply with the rules of financial independence (see Altmeyen 2004, p. 401).

Depending on the stock exchange listing, supplementary regulations of the NYSE, respective the NASDAQ, may apply in addition to the ones of the SEC. According to Section 303 of the Sarbanes Oxley Act, a listing at the NYSE requires the independence of all audit committee members. Thus, an audit committee member is independent if he is not an employee of the (affiliated) corporation currently or has not been for the past three years. In addition his direct relatives are not part of the management and have not been for the past three years (NYSE 2004). With regard to the audit committee member's independence, the NASDAQ demands for an enhancement of the greater SEC criteria, thus demanding that an audit committee member has not participated in the preparation of the annual financial statements as a governing body within the last three years. The Sarbanes Oxley Act does not provide for such cooling off periods after termination of employment. However, as already described above, the German stock corporation

law generally arranges for a cooling off period of two years for former management board members to become supervisory board members.

In addition to the requirements of independence, the Sarbanes Oxley Act is demanding for at least one financial expert within the audit committee. Initially, the SEC was interested in stipulating that this person ought to be an expert in terms of accounting. However, in the end they refrained from doing so. In addition to accounting, it is hence acceptable if the expert has knowledge of other finance areas. An exception to this rule may apply if it has been briefly described why no financial expert was appointed as an audit committee member. In general, this is not often the case in order to maintain a good reputation (see Altmeyen 2004, p. 397). The requirement to appoint at least one financial expert is consistent with the amendments of the German stock corporation law. Though in contrast to the German legislator, the SEC is specifying the financial expert qualification in detail. Thus the financial expert is expected to have good knowledge with regard to the preparation of annual financial statement and accounting standards. In addition, he must have the relevant skills to generally judge on the application of accounting policies with regard to estimation, amortization, and the setting up of accruals. Furthermore, he needs to be experienced in the preparation, assessment, analysis, and evaluation of financial statements which are comparable in scope and complexity to the registered corporation's financial statement. Moreover, he is expected to be experienced in actively supervising people that are assigned to the previously described tasks (Section 401 and 407 of the Sarbanes Oxley Act). Such requirements correspond to the job specifications of accountants, finance directors, accounting directors, or similar profession. The Sarbanes Oxley Act does not comment on the qualification of other audit committee members.

In case a corporation is listed at the NYSE, at least one member of the audit committee needs to be experienced in finance and accounting management (NYSE 2004, Section 307). This is consistent with the minimum requirement of one financial expert according to the Sarbanes Oxley

Act. Furthermore, all members need to prove basic knowledge in finance and accounting or are required to be financially literate after a reasonable time. Hence, the NYSE requirements are more demanding than the ones of the Sarbanes Oxley Act with regard to the professional qualification of the audit committee members.

In case a corporation is listed at the NASDAQ, all audit committee members are expected to understand and comprehend the respective corporation's financial statements at the time of their nomination. The regulation with regard to the financial expertise is comparable to the one of the NYSE. In accordance with the regulations of the NYSE, at least one audit committee member is to be experienced in finance and accounting (financial expert). Thus, a professional qualification with regard to accounting or any other comparable experience or basic background knowledge is expected (NASDAQ 2006, Section 4350).

Empirical Relevance of Audit Committees

Capital Costs and Market Reactions

Since only few multivariate empirical studies concerning the impact of audit committees on corporate governance are available for the German capital market (see Velte 2009, 2011; Velte and Stiglbauer 2011), US-American studies have been used primarily. The following explanation provides an overview of current research. According to Ashbaugh et al. (2004), the number of independent audit committee members is related to lower costs of capital. Anderson et al. (2003) empirically proved that audit committees with independent members imply lower interest on debt. In contrast, the results of Bhagat and Black (1999) suggest a lower corporate performance in case the majority of the audit committee members are independent. Similarly, this holds true for the analysis of Klein (1998). Likewise, no statistical significance exists regarding the number of nonexecutive directors and the enhancement of corporate performance.

In addition, the study of DeFond et al. (2005) was addressing the question whether the existence of an accounting expert or a member with any other financial expertise had an influence on the amount of accumulated abnormal return on investment. The results of this study provided a statistical significant positive evidence for an accounting expert. The studies of Wild (1994, 1996) found a significant positive increase of accumulated abnormal accruals, e.g., stock price fluctuation on the statement results.

The empirical results suggest that the implementation of independent and financially literate audit committees provides and increases confidence on the capital market. Hence, the demonstrated attempts of the standard setter regarding the job specifications of audit committee members (independence and financial expertise) are legitimated from an economic point of view for the US-American one-tier system.

Earnings Management and External Management Reporting

An offensive earnings management is sanctioned by the capital market with regard to balance sheet analysis. Hence, a conservative performance is approved. The earnings management performance is measured by means of abnormal accruals. By supervising managing directors, the audit committee is due to provide incentives for the reduction of earnings management.

According to Ebrahim (2007), a significant negative correlation exists between the number of independent audit committee members and the accounting policy, measured by means of abnormal accruals. Xie et al. (2001) analyzed the financial expertise of audit committee members. They were able to prove a significant evidence for a negative influence of investment banking members and nonexecutive directors on the amount of corporations accounting policy, measured by means of discretionary (disproportional) accruals.

Bédard et al. (2004) verified a significant negative influence on the accounting policy, in case at least one audit committee member had the

respective financial expertise. A corresponding relation applies to audit committees with solely nonexecutive directors without substantial corporate integration, provided that the corporate addressees have detailed knowledge of the audit committee's job specification. According to the research of Yang and Krishnan (2005), a significant positive relation exists between the share property of the audit committee members and the amount of nondiscretionary operative accruals. Further studies of Klein (2002) provide evidence for a significant negative correlation between the audit committee's independence and accounting policy in case the audit committee not solely but by majority consists of nonexecutive directors. The respective relation is measured by means of the absolute value of adjusted abnormal accruals.

Other areas of research seek to address the impact of audit committees on the occurrence of subsequent accounting adjustment. Reactive adjustment leads to negative market reactions. From a capital market point of view, they are caused by (intentional) misinterpretations of the corporate management. According to the empirical results of Abbott et al. (2004), the frequency of occurrence of subsequent adjustment of the annual financial statement may be reduced significantly by audit committees solely consisting of independent members and/or the existence of at least one financial expert.

Furthermore, accounting policy is directly influencing quality and quantity of the external management reporting. Hence, by pooling financial expertise, the audit committee fulfils an advisory function to the managing directors. The joint effort is to provide the capital market with the best available management reporting. The survey of Karamanou and Vafeas (2005) proves a significant positive correlation between financial expertise in the audit committee and the frequency, e.g., quality of the management's performance forecast. The results differentiate in how far the corporation responds to negative forecasts ("bad news") and how well they are documented. In addition, attention has been paid to the conformity of corporation information with the analyst's opinion.

Management Fraud

In addition to the impact on accounting policy, empirical corporate governance research is addressing possible consequences of audit committees on the existence and prevention of management fraud. Here, the occurrence of fraud is associated with an intentional erratic behavior of the management and results from information asymmetries between the corporation's management and the capital market. The continuous supervision of the management by the audit committee seeks to increase the likelihood of revealing fraud. In addition, it is likely that the implementation of audit committees may impede the occurrence of accounting fraud preemptively and avoid falsification of the balance sheet by means of due diligence.

In case the submitted financial statement documents are rejected by the SEC in the context of enforcement, negative publicity and damage to the corporation's reputation will be the consequences. According to Abbott et al. (2000), audit committees without continuous employees, holding a meeting for at least twice a year, might be able to alleviate the rejection of the SEC. A corresponding significant negative influence can be verified for audit committees without employees or managing directors having substantial relations to the corporation or its management. These findings are consistent with the research of Krishnan (2005). Hence, an independent and financially literate audit committee reduces the risk of internal control-system failure. However, the corporation is obliged to report on the weakness in case of a change of the auditor. The survey of McMullen (1996) reveals a significant negative correlation between the existence of audit committees and the sanctions of the SEC. According to Beasley et al. (2000), the likelihood of management fraud diminishes with the implementation of audit committees that solely consist of independent members. The sole existence of audit committees leads to a corresponding significant negative influence. The research of Uzun et al. (2004) corresponds with the mentioned empirical findings. Thus, the occurrence of fraud is negatively correlated with the existence of audit committees and, respectively, positively

correlated with audit committees consisting of dependent, nonexecutive directors. These results are supplemented by the research of McMullen and Raghunandan (1996). By trend, corporations with no financial statement fraud have audit committees solely consisting of non-managing directors, i.e., independent audit committees nominating at least one financial expert (e.g., auditor).

External Audit

Amongst others, US-American surveys emphasize on the relation between audit committees and external audit. In addition to the supervision of management and accounting, this activity aims at supervising the external auditor. By continuous monitoring of the auditor's qualification, the audit committee is able to enhance the quality of corporate governance. Amongst others, the relation between compensation of audit and non-audit activities provides a basis for judging on the independence of the external auditor. According to the prevailing opinion, an increase in compensation of audit (non-audit) activities leads to an increase (decrease) in the auditor's independence *ceteris paribus*. By trend, non-audit activities such as consulting promote the annual auditor's dependence on the management. In addition, they imply the risk of financial side transfers, leading to an inferior audit quality. Hence, the auditor might be willing to grant a concession with regard to the certification of the financial statement he/she might not be granting in case he/she had no consulting mandate.

Carcello et al. (2002) provided evidence for a significant positive relation between audit committees solely or by majority consisting of independent members and the amount of compensation for audit activities of the auditor. According to Abbott et al. (2003a), a completely independent audit committee with respective financial expertise has a positive influence on audit fee. Another survey of Abbott et al. (2003b) concludes that audit committees with solely independent members holding a meeting at least four times a year might reduce the ratio for the compensation of the non-audit activities, since they might endanger auditor independence. Consequently, this implies a significant

positive relation between the independence of audit committee members and auditor independence. However, the results of Vafeas and Waagelein (2007) are opposed to the aforementioned findings. Their results suggest a significant positive relation between the requirement of appointing at least one managing director or person being a member of an audit committee of another Fortune 500 corporation into the audit committee and the amount of the audit fee.

Auditor independence serves as a substitute for the audit quality. Within an international framework, it is measured not only by means of the auditor's fee but of the size of the audit company. According to the basic description of the audit theory of DeAngelo (1981), auditor independence and hence audit quality increase with the appointment of international awarded and top-selling audit firms in comparison with other audit and trust companies. Empirical surveys have been addressing possible relations between the implementation of audit committees and the nomination of the annual auditor. If an independent audit committee is responsible for the nomination of the auditor and thus might generate an adequate audit quality in favor of the shareholders, counterproductive intervention of the management is less likely.

The empirical survey of Eichenseher and Shields (1985) already verifies that corporations tend to implement audit committees in case a new auditor needs to be appointed and one of the eight top-selling audit companies is involved. Additional empirically proven relations between audit committees and the external audit refer to the independence of the audit committee members and the likelihood of a cancellation of the auditor's contract. According to Lee et al. (2004), a significant negative relation exists between a solely independent audit committee and the cancellation, e.g., resignation of the audit mandate. The research of Knapp (1987) suggests a significant positive influence of the existence of audit committees on the appointment of one of the eight top-selling companies, the economic situation of the corporation in question, and the likelihood of the board supporting the annual

auditor in case a conflict between auditor and management arises.

The majority of the US-American empirical research could verify a positive influence of audit committees on the quality of external annual audit resulting from the normative approach of the legislator. Until the end of the 1990s of the twentieth century, empirical research was emphasizing only on the existence of audit committees. Later, with the introduction of the Sarbanes Oxley Act, the job specification of the audit committee became more important in terms of empirical research. Attention needs to be paid to the trend that only a cumulative existence of independence and financial expertise leads to significant positive impacts on the amount of the audit fee. The surveys often comply with the normative status quo of the Sarbanes Oxley Act, e.g., all members of the audit committee are independent and at least one member is a financial expert.

Conclusion

Audit committees are of great importance in order to strengthen corporate governance within the US-American one-tier system and the German two-tier system. The comparative normative analysis suggests that audit committees are representative for the convergence of the one- and two-tier system. With regard to the one-tier system, the independent audit committee serves as a monitoring instrument for the managing directors of the board of directors. With regard to the two-tier system, the audit committee is responsible for preparing the plenum's decision. And with the nomination of at least one financial expert, it is ought to counteract the increase of professionalism within the supervisory board. The ideas of the European Commission regarding the job specification of audit committees have been realized in Germany. As a result, independence and financial expertise are of equal importance. This is due to the fact that the EU member states use one- and two-tier systems, therefore demanding the equality of both requirements.

Overall, the requirements for the implementation and job specification of audit committees are more restrictive in the US-American one-tier system. They ought to impede a potential self-assessment of the board of directors. An objective supervision of financial accounting and executive directors is not feasible with dependent audit committee members. Hence, the subject of the member's independence is of major importance within the one-tier system. In contrast, the two-tier system is characterized by a vast separation between managing and supervising tasks. As a result, the requirements for audit committee members are described in detail and more restrictive in the USA. However, the independence of audit committee members might be impaired as well in the two-tier system. The requirements of the German law (at least one independent member in the audit committee) might not be sufficient if a member accepts an additional position in the supervisory board of another corporation of the same industry. This would lead to an increase in risk of conflicts of interests of audit committee members. Though, with the implementation of audit committees, the German two-tier system aims at a professional execution of the supervisory board's tasks by a purposive preparation of the plenum's decision.

The normative concretion has been analyzed along with empirical findings of the international corporate governance research concerning audit committees. Yet, the present empirical results of capital market surveys are primarily based on the US-American one-tier system. With regard to the rising importance of audit committees in the two-tier system, further studies are needed. Emphasis should be placed on the question whether and in how far the implementation of audit committees, including respective job specification, has an actual influence on the improvement of corporate governance. With regard to the one-tier system, empirical results suggest a correlation between the implementation and job specification of audit committees and several corporate governance indicators. Many surveys conclude a significant positive correlation between the nomination of financial experts and independent members in the audit committee and the aforementioned corporate governance variables.

Hence, further studies should address the question whether and in how far the improvement of corporate performance within the one-tier system by the appointment of independent and financially literate audit committee members can be adopted to the German two-tier system. Yet, it needs to be considered that the competencies of the German audit committee cannot be compared to the US-American as a result of the separation between the corporation's management and supervision. By trend, the majority of the respective studies suggest that the US-American capital market has more confidence in corporations with independent and financially literate audit committee members. Thus, the certification of an increase in corporate governance quality might become more realistic. Again, this fact should lead to an increase in research on audit committees within the German two-tier system.

Cross-References

- [Auditing](#)
- [Hedge Fund Activism in Corporate Governance](#)
- [Institutional Review Board](#)
- [Risk Management, Optimal](#)

References

- Abbott, Park, Parker (2000) The effects of audit committee activity and independence on corporate fraud. *Manag Finance* 26:55–67
- Abbott et al (2003a) The association between audit committee characteristics and audit fees. *Auditing* 22:17–32
- Abbott et al (2003b) An empirical investigation of audit fees, nonaudit fees, and audit committees. *CAR* 20:215–234
- Abbott, Parker, Peters (2004) Audit committee characteristics and restatements. *Auditing* 23:69–87
- AICPA (1967) Executive committee's statement. *JoA* 163:10–11
- Altmeppen (2004) Der Prüfungsausschuss. *ZGR* 33:390–415
- Anderson RC, Mansi SA, Reeb DM (2003) Board characteristics, accounting report integrity, and the cost of debts. Working paper, Washington, DC
- Ashbaugh, Collins, LaFond (2004) Corporate governance and the cost of equity capital. The University of Wisconsin, Madison
- Beasley et al (2000) Fraudulent financial reporting. Consideration of industry traits and corporate governance mechanisms. *Account Horiz* 14:441–454
- Bédard, Chtourou, Courteau (2004) The effect of audit committee expertise, independence and activity on aggressive earnings management. *Auditing* 23:13–35
- BGH (1982) Bundesgerichtshof II ZR 27/82 vom 15.11.1982. *NJW* 36:991–992
- Bhagat, Black (1999) The uncertain relationship between board composition and firm performance. *Bus Lawyer* 54:921–963
- Carcello, Hermanson, Neal (2002) Disclosures in audit committee charters and reports. *Account Horiz* 16:291–304
- Coenenberg, Reinhardt, Schmitz (1997) Audit committees. *DB* 50:989–997
- DeAngelo (1981) Size and audit quality. *JoAE* 3:183–199
- DeFond, Hann, Hu (2005) Does the market value financial expertise on audit committees of boards of directors? *JoAR* 43:153–193
- Ebrahim (2007) Earnings management and board activity. An additional evidence. *Rev Account Finance* 6:42–58
- Eichenseher, Shields (1985) Corporate director liability and monitoring preferences. *J Account Public Policy* 4:13–31
- Karamanou, Vafeas (2005) The association between corporate boards, audit committees, and management earnings forecasts. An empirical analysis. *JoAR* 43:453–486
- Klein (1998) Firm performance and board committee structure. *JoLE* 41:275–303
- Klein (2002) Audit committee, board of director characteristics and earnings management. *JoAE* 33:375–400
- Knapp (1987) An empirical study of audit committee support for auditors involved in technical disputes with client management. *Account Rev* 62:578–588
- Krishnan (2005) Audit committee quality and internal control. An empirical analysis. *Account Rev* 80:649–675
- La Porta et al (2002) Investor protection and corporate valuation. *JoF* 57:1147–1170
- Lee, Mande, Ortman (2004) The effect of audit committee and board of director independence on auditor resignation. *Auditing* 23:131–146
- McMullen (1996) Audit committee performance. An investigation of the consequences associated with audit committees. *Auditing* 15:87–103
- McMullen, Raghunandan (1996) Enhancing audit committee effectiveness. *J Account* 179:79–81
- NASDAQ (2006) Section 4350 qualitative listing requirements for Nasdaq issuers except for limited partnerships. NASDAQ, New York
- National Commission on Fraudulent Financial Reporting (1987) Report of the National Commission on Fraudulent financial reporting, Washington, DC
- NYSE (2004) Listed company manual. Section 303A.00 Corporate Governance Standards (Last Modified 3 Nov 2004), New York
- Quick, Hoeller, Koprivica (2008) Prüfungsausschüsse in deutschen Aktiengesellschaften. *ZCG* 3:25–35
- SEC (1974) SEC accounting release no 165, 20 Dec 1974

- SEC (1978) SEC accounting release no 13-346, 9 Mar 1978
- SEC (2002) Proposed rule disclosure required by 404, 406 and 407 of the Sarbanes Oxley Act of 2002, Washington, DC
- SEC (2003) SEC accounting release no 33-8820, Washington, DC
- Uzun, Szewczyk, Varma (2004) Board composition and corporate fraud. *Financ Anal J* 60:33–43
- Vafeas, Waagelein (2007) The association between audit committees, compensation incentives and corporate audit fees. *Rev Quant Finance Account* 28:241–255
- Velte (2009) Die Implementierung von Prüfungsausschüssen/Audit Committees des Aufsichtsrats/Board of Directors mit unabhängigen und finanzkompetenten Mitgliedern. *JfB* 59:123–174
- Velte (2011) The link between audit committees and corporate governance quality. *Corp Ownersh Control* 8:5–13
- Velte, Stiglbauer (2011) Impact of audit committees with independent financial experts on accounting quality. *Probl Perspect Manag* 9:17–33
- Wild (1994) Managerial accountability to shareholders: audit committees and the explanatory power of earnings for returns. *Br Account Rev* 26:353–374
- Wild (1996) The audit committee and earnings quality. *J Account Audit Finance* 11:247–276
- Xie, Davidson, DaDalt (2001) Earnings management and corporate governance. The roles of the board and the audit committee. *J Corp Finance* 9:295–316
- Yang, Krishnan (2005) Audit committees and quarterly earnings management. *Int J Audit* 9:201–219

More broadly, auditing is a mechanism for quality assurance: An auditor is a third-party agent that acts as a certification intermediary and that examines, corrects, and verifies the financial accounts of a business in order to make them more credible to various stakeholders, such as investors.

Audit Regulation and Profession

Auditing is regulated by laws, by global and local guidelines published by professional bodies, and also by professional ethics and practice. Audit legislation varies across countries, but typically, auditing of firms listed in stock markets (public firms) is strictly regulated, whereas that of non-listed (private) firms is either not required or is clearly less regulated. The International Standards on Auditing (ISA) issued by International Federation of Accountants (IFAC) gives guidelines for auditing profession globally.

The audit industry has consolidated over the years, and there are currently four global audit firms. These so-called Big 4 audit firms are the PricewaterhouseCoopers (PWC), Ernst & Young (EY), KPMG, and Deloitte Touche Tohmatsu (Deloitte). Other audit firms are smaller and act more locally.

Accounting and audit failures, including those of Enron, WorldCom, and Tyco International, shaped accounting industry and raised serious questions about the auditors' independence of their clients. One consequence was the adoption of Sarbanes-Oxley Act of 2002 (SOX) in the US, which includes rules aimed to improve audit quality and auditor independence. The issue has evolved during the financial crisis, and consequently, many countries have either suggested or already taken regulatory initiatives to (further) improve auditor independence and audit quality. These initiatives include the separation of mandatory auditing services from consulting services, the restriction of the market share of a given auditor in a country, and mandatory audit rotation, which requires that an audit firm can audit the same client firm only over a limited period of time.

Auditing

Ari Hyytinen¹ and Juha-Pekka Kallunki²

¹Jyväskylä School of Business and Economics, University of Jyväskylä, Jyväskylä, Finland

²Oulu Business School, University of Oulu, Oulu, Finland

Definition

In its commonly cited definition, the American Accounting Association defines (financial) auditing as “A systematic process of objectively obtaining and evaluating evidence regarding assertions about economic actions and events to ascertain the degree of correspondence between those assertions and established criteria and communicating the results to interested users.”

Audit Research

Identifying the determinants of audit fees, which auditors charge from their customers, has been probably the most intensively explored area in audit research. Since the seminal paper by Simunic (1980), many researchers have shown that audit fees are related to client characteristics, such as size, complexity, and risk. For instance, firms with greater financial leverage, lower liquidity, or greater reported losses pay larger audit fees because the perceived audit risk and the required amount of audit work are greater in these firms.

Another frequently cited finding is that clients are willing to pay a premium for the Big 4 audit firms. Some studies also report a fee premium for certain audit offices and audit firms with industry-specific expertise (e.g., Francis et al. 2005).

Much research efforts have also been devoted to examining auditor independence. High audit fees and non-audit services provided by auditors may particularly impair auditor independence. However, many academics and practitioners remind that an incumbent auditor's deep knowledge about its client may actually improve audit quality. Research evidence on the area has not been entirely conclusive. Some recent papers find no evidence that auditors compromise their independence (e.g., Hope and Langli 2010).

Recently, audit research has focused, e.g., on how auditors' personal traits, such as their attitudes toward risk-taking, affect audit outcomes (Amir et al. 2014) and on the effects and possible design of reforms that aim at improving the workings of the auditing industry (e.g., Ronen 2010).

Third-party auditing of the compliance of firms is also common in areas other than financial accounting, such as in the enforcement of product and safety standards and environmental regulation and standards (see, e.g., Dranove and Jin 2010; Duflo et al. 2013).

Cross-References

► [Information Disclosure](#)

References

- Amir E, Kallunki JP, Nilsson H (2014) The association between individual audit Partners' risk preferences and the composition of their client portfolios. *Rev Account Stud* 19:1
- Dranove D, Jin GZ (2010) Quality disclosure and certification: theory and practice. *J Econ Lit* 48(4):935–963
- Duflo E, Greenstone M, Pande R, Ryan N (2013) Truth-telling by third-party auditors and the response of polluting firms: experimental evidence from India. *Q J Econ* 128(4):1499–1545
- Francis JR, Reichelt K, Wang D (2005) The pricing of national and city-specific reputations for industry expertise in the US audit market. *Account Rev* 80:113–136
- Hope O-K, Langli JC (2010) Auditor independence in a private firm and low litigation risk setting. *Account Rev* 85(2):573–605
- Ronen J (2010) Corporate audits and how to fix them. *J Econ Perspect* 24(2):189–210
- Simunic DA (1980) The pricing of audit services: theory and evidence. *J Account Res* 18:161–190

Austrian Economics

- [Austrian Perspectives in Law and Economics](#)
- [Austrian School of Economics](#)
-

Austrian Perspectives in Law and Economics

Shruti Rajagopalan¹ and Mario J. Rizzo²

¹Department of Economics, State University of New York, Purchase College, Anderson, NY, USA

²Department of Economics, New York University, New York, NY, USA

Abstract

This encyclopedia discusses some of the important representative ideas in the Austrian tradition such as spontaneous orders, individuals coping with decentralized knowledge and uncertainty, coordination, and entrepreneurship. The encyclopedia highlights the essential and distinctive feature of the Austrian school of law and economics is its emphasis on both

economic and legal *processes*. The Austrian emphasis on *processes* can be applied to both these branches of law and economics: individual behavior within institutions, as well as individual behavior, leading to the emergence of institutions.

Synonyms

Austrian Economics

Introduction

The Austrian tradition is distinct in emphasizing the role of uncertainty and ignorance of the individual in decision-making. Austrian scholars emphasize that the knowledge in society is fragmented and dispersed across individuals. Therefore, the main problem faced in society is one of coordination and social cooperation. The essential and distinctive feature of the Austrian school of law and economics is its emphasis on both economic and legal *processes*. The Austrian emphasis on *processes* can be applied to both these branches of law and economics: individual behavior within institutions, as well as individual behavior, leading to the emergence of institutions. This encyclopedia discusses some of the important representative ideas in the Austrian tradition such as spontaneous orders, individuals coping with decentralized knowledge and uncertainty, and coordination.

Origin of Institutions

The Austrian law and economics is most closely associated with Nobel Laureate Friedrich Hayek and his work on the origins of legal institutions. Hayek described evolved law as “conceived at first as something existing independently of human will” and distinguished it from legislation, which was “the deliberate making of law” by few individuals (Hayek 2011 [1960], pp. 118–119; and Hayek 1973, pp. 72–73). Hayek’s analysis is the direct outgrowth of the earliest Austrian

insights as well as those of the Scottish Enlightenment of the eighteenth century. Carl Menger, the founder of the Austrian approach, stated that social scientist must explain “how can it be that institutions which serve the common welfare and are extremely significant for its development come into being without a common will directed toward establishing them” (Menger 1963 [1883], p. 146). Menger’s question links the Austrian approach to Scottish enlightenment scholars, who explained the spontaneous orders as “the result of human action, but not the execution of any human design” (Ferguson 1782[1767], p. III, S.2). This line of enquiry has continued in the Austrian tradition where modern scholars like Mario Rizzo and Gerald O’Driscoll ask, “How can individuals acting in the world of everyday life unintentionally produce existing institutions?” (O’Driscoll and Rizzo 1996, p. 20).

Menger argued that the emergence of money is one such example of spontaneous development of institutions. To solve the problem of the double co-incidence of wants, individuals find more highly valued commodities to exchange and therefore, add an exchange value to the use value of these goods, increasing the demand. As more individuals participate in such exchange, they converge to one or two generally accepted media of exchange, which we call money (Menger 1892).

In the same spirit of Menger’s explanation for the emergence of money, Ludwig von Mises, one of the most prominent scholars in the Austrian tradition, attempted to explain the emergence of legal rules. Mises argued that property law originally arose from recognition of simple possession and contract law from primitive acts of exchange within localized areas. While the former may have had as its primary motive the avoidance of violence and the creation of peaceful conditions, the latter was almost bound to arise under conditions of de facto property in order to pursue the gains from exchange. But ultimately the world created by these early efforts produced institutions that could be viewed as “a settlement, an end to strife, an avoidance of strife” and thus “their result, their function” is to produce peace within a community (Mises 1981 [1922], p. 34).

Hayek applied spontaneous order analysis, not just to specific legal institutions, but to the entire legal tradition of common law. Hayek described it as “deeply entrenched tradition of a common law that was not conceived as the product of anyone’s will but rather as a barrier to all power” (Hayek 1973, p. 84). Scholars have written about the role of litigants, judges, lawyers, etc., in the emergence of common law rules. Some have described the efficiency of common law as a result of private interests of litigants to resolve the dispute. On the supply side, Zywicki (2003) describes the common law system in the Middle Ages as polycentric law-making and attributes the emergence of efficient rules to courts and judges competing for litigants and fees in overlapping jurisdictions. Since the evolution of legal rules and institutions is a continual process, institutional entrepreneurs have an important role in finding opportunities to resolve conflicts and form more efficient, context-specific rules. This may inadvertently give rise, not only to the development of the specific rule, but the entire legal tradition.

Time and Ignorance

Perhaps the most important contribution of Austrian economics, as exemplified especially in the work of F.A. Hayek, is the understanding that individual behavior and social cooperation takes place in the face of decentralized knowledge. However, individuals have limited knowledge, and social knowledge is dispersed or decentralized. Furthermore, this knowledge may not be costless to acquire, or even exist in the form required for decision-making.

In his essay *The Use of Knowledge in Society*, Hayek notes “economic problem of society is thus not merely a problem of how to allocate ‘given’ resource . . . It is rather a problem of how to secure the best use of resources known to any of the members of society, for ends whose relative importance only those individuals know” (1945, pp. 519–520). So, for Hayek, the function of the law is to provide dispersed economic decision-makers the “additional knowledge” or,

more exactly, surrogates for the knowledge necessary to rationally plan (Ibid, p. 521).

The problem of decentralized knowledge and uncertainty is at the very core of the Austrian approach to economics, especially the Austrian approach to law and economics. Karen Vaughn, while describing the overall Austrian approach, wrote, it is “impossible to think of Austrian economics as anything but the economics of time and ignorance” (Vaughn 1994, p. 134). Rizzo clarifies that it is the economics of individuals *coping* with real time and radical ignorance (O’Driscoll and Rizzo 1996, p. xiii). If individuals were not continuously faced with uncertainty and ignorance, a system of legal rules would be quite different. If one takes seriously the fact that all knowledge is decentralized, institutions must have certain characteristics to solve the problem of dispersed knowledge in society.

The knowledge problem may take a static or dynamic form. The static knowledge problem is the utilization of the current stock of dispersed factual knowledge so that individuals can coordinate their actions and plans at a particular time. The dynamic knowledge problem is the growth of new knowledge, not currently in the system, so that improved forms of coordination can occur through time (Kirzner 1973, 1992; Leoni 1991 [1961]). Given static and dynamic knowledge problems, legal institutions are important in the coordination of different individuals at a point in time, and over a period of time.

Ludwig Lachmann argued that legal institutions act as “signposts,” because these institutions as they help individuals overcome problems of exchange by enabling individuals to form expectations and coordinate plans. And the most important characteristic of these legal institutions is their stability. He argues that incremental changes of rules, especially in their application to particular new conflicts, must not change the predictable nature of legal rules. “If institutions are to serve us as firm points of orientation their position in the social firmament must be fixed. Signposts must not be shifted” (Lachmann 1971, p. 50). However, this does not mean that a specific legal rule cannot or should not change. It is that the system of rules must be predictable. Lachmann’s unchanging

signposts are about the stability of the system, rather than the stability of each particular rule. This echoes Hayek's argument that laws "are intended to be merely instrumental in the pursuit of people's various individual ends. . . . They could almost be described as a kind of instrument of production, helping people to predict the behavior of those with whom they must collaborate, rather than as efforts toward the satisfaction of particular needs" (Hayek 1944, pp. 72–73).

In a world filled with ever-changing and complex interactions, what is required is a simple system of rules to guide individual behavior. Epstein (1995) argues that *simple* legal rules can act as signposts in a complex world that is uncertain and dynamic. As systems go from hierarchical orders to spontaneous orders, the degree of coordination required, and the inherent complexity in mutual compatibility of plans, magnifies. Once it is appreciated that in referring to legal rules we are referring to inputs into *individual* decision-making, it becomes evident that the more decentralized and complex a system is the more critical it is that rules are simple.

Coordination and Optimality

The emphasis on ignorance, decentralized knowledge, and uncertainty leads us to the Austrian emphasis on coordination in society. It is important to understand the difference between coordination and the neoclassical concept of optimality. In the Austrian approach, the focus is on coordination, and not on optimality.

The fundamental meaning of coordination is simply the mutual compatibility of plans. This requires two things. First, each individual must base his plans on the correct expectation of what other individuals intend to do. Second, all individuals base their expectations on the same set of external events (Rizzo 1990, p. 17). In this basic meaning, the existing dissemination of knowledge has led to a state of affairs where each party is able to implement his plans. All offers to buy are accepted by sellers. All offers to sell are accepted by buyers. This is to be distinguished from the *process* of coordination whereby through

trial and error learning and entrepreneurial discovery agents are able to make their plans compatible or more nearly compatible with those of others.

Coordination is analytically different, though not incompatible, with the concept of optimality. Pareto optimality implies that individuals exhaust all the potential gains from trade. This is a special case of coordination. However, there can be coordination, or the execution of mutually compatible plans, which do not exhaust all potential gains from trade. "...these plans are mutually compatible and that there is consequently a conceivable set of external events, which will allow all people to carry out their plans and not cause any disappointments" (Hayek 1937, p. 39). A state of mutually compatible plans "represents in one sense a position of equilibrium, it is however clear that it is not an equilibrium in the special sense in which equilibrium is regarded as a sort of optimum position" (Hayek 1937, p. 51). Everyone within a system may have mutually compatible plans, and yet there may be better trading opportunities out there so that at least some parties can improve their positions by alternative trades. Thus, if there is a sense in which the mutual compatibility of plans is an optimum, it is only a local optimum, that is, between the direct parties to an exchange.

In a fully or perfectly coordinated state of affairs, each individual correctly takes into account: (1) the actions being taken by everyone else in the set and (2) the actions which the others might take, if one's own actions were to be different (Kirzner 2000, p. 136). The latter ensures that no buyer transacts at a price higher than that which a potential seller would offer. And that no seller transacts at a price lower than that which a potential buyer would offer. In this sense, the Austrian idea of coordination is compatible with the neoclassical concept of Pareto optimality. If each individual fully takes account of the actions (and potential actions) of every other individual, all courses of action, which might be preferred by any one participant without hurting anyone else, must already have been successfully pursued. In this sense, Pareto-optimality corresponds to perfect coordination (Kirzner 2000, p. 144).

The importance of law to basic and perfect coordination is indirect. Legal rules obviously

cannot affect a state of affairs in which plans are mutually compatible and all arbitrage opportunities are eliminated. Nevertheless, they can, by facilitating exchange and protecting or ensuring the right to entrepreneurial (arbitrage) profit, make the discovery processes that move the system *toward* mutual compatibility and full coordination more likely to be unleashed. Lachmann emphasizes the aspect of institutions that aid the formation of expectations and argues that institutions “enable each of us to rely on the actions of thousands of anonymous others about whose individual purposes and plans we can know nothing. They are nodal points of society, coordinating the actions of millions whom they relieve of the need to acquire and digest detailed knowledge about others and form detailed expectations about their future action. But even what knowledge of society they do provide in highly condensed form may not all be relevant to the achievement of our immediate purposes” (1971, p. 50).

More generally in the field of law and public policy, simply to assume that the lawmakers, paternalist, or central planner each has the relevant knowledge to bring out his stated goals is to assume the knowledge problem away (e.g., Rizzo and Whitman 2009, p. 905). The task is actually the opposite, to *solve* the problem of decentralized knowledge. When lawmakers act on the basis of a pretense of knowledge to which they have no access, they *increase* uncertainty relative to attainment of individuals’ goals.

Legal Order

Economic activity takes place within the framework of a “given” legal order. However, some explanation is required to produce clarity about the meaning of such “givenness.” Something can be given in the objective sense, in which the legal rules are given to the omniscient observing economist. This is a conceptual expedient for the creation of narrowly specified and limited models. More important is the subjective sense, in which they are given to the individuals whose actions we are trying to explain (Hayek 1937, p. 39).

Givenness in this second sense means that the framework, at least insofar as it affects the plans of the individual, is predictable.

However, it is impossible for any system of legal rules to be completely defined, specified, unambiguous, and hence perfectly predictable either in theory or in application. If ignorance and genuine uncertainty is taken into account, then it is problematic to assume a completely specified set of legal rules. While legal institutions may help individuals cope with ignorance in the market, these institutions are themselves subject to the knowledge problem. Hayek emphasized the knowledge problem not only in the context of the market, but extended it to other orders. The language of rules and legal decisions is always characterized by some ineradicable degree of uncertainty or vagueness, and therefore even if it were conceptually possible to define all rules clearly, it would be prohibitively costly (Whitman 2002, p. 6).

Whitman argues that it is inaccurate to see law as a process of one-way causation where a given set of exogenous legal rules resolves conflicts (2002, p. 3). In this area, an important aspect of the Austrian approach to law and economics is to endogenize the system of legal rules. Especially in a legal system in which judges make law by establishing, modifying, overturning, and reaffirming precedents, the actions of participants play a pivotal role in determining the direction of the law. Even when there is no new legal rule, or a novel application of an old legal rule, the law still changes as a result.

If the legal rules are and constantly evolving, then what do we mean by the “givenness” of rules to the agents in the system? If the rules are constantly challenged and modified, then how do they provide any kind of guidance for human action? And more importantly, how does a constantly evolving system of rules act as a constraint for individual behavior? Within any legal system, there is a tension between the need to produce certainty of the laws with the need for the law to evolve and be relevant to new situations. This is particularly the case in common law. The question is often posed as a tradeoff between certainty and flexibility of the law.

In the first place, at the moment of choice, a certain framework of rules is given. Today's market transactions must be executed within the framework of rights as given today, but that framework is itself the unintended result of the past actions of many individuals. These rules of the game are the "relics" of successful plans of earlier generations that have "gradually crystallized" into institutions (Lachmann 1971, pp. 68–69). Second, legal rules are not being changed in entirety, but the change is marginal. "A change in the law can be marginal in the sense that it is perceived as deviating only slightly from precedent" (Rizzo 1980a, p. 651). Third, Rizzo further argues that certainty of the law and its flexibility are not incompatible and that the "law endures by changing" (1999, p. 499). The law must have a certain plasticity to survive through economic changes. A rigid or static framework would break apart. These considerations imply that the "system" of rules is relatively stable, while marginal changes to specific rules adapt to the new or changing circumstances. This is the idea of the decomposability of the system of rules.

The most important factor that enhances the predictability of law, even as it adapts and changes to new circumstances, is the nature of the process involved. To see this we must distinguish between two forms of coherence in the law. Rizzo (1999) differentiates logical coherence of the law, from the praxeological coherence, or the coordination, which arises from the law. For Rizzo, and the Austrian approach more generally, it is praxeological coherence, or coordination in society, which is at the forefront of analysis. The logical consistency of laws is neither necessary nor sufficient for such coordination.

Another related reason for the emphasis on praxeological coherence of rules is the recognition that there is no one single correct set of legal rules. If the moral intuitions of individuals are not completely consistent or if they have gaps, then there may be more than one right answer in a particular case. All of these may be in the range of expectations of the agents. Presumably, this does not unduly disturb the order of actions as long as the acceptable range is within ordinary limits.

The process of rule-evolution or generation in a common law system is based on trial and error (Hayek 2011 [1960], pp. 122–125). Therefore, at any given point in time some rules or application of rules will simply be wrong. In other words, they will be ripe for revision as the process continues. For Hayek the primary focus is on the overall system. The system is or should be the primary object of normative evaluation. Its mistakes are in a sense simply part of the process.

Entrepreneurship

An important and recurring theme in Austrian economics is the role of entrepreneurial alertness in seizing profit opportunities and thereby enhancing the level of coordination in the market. Krecké argues that the Austrian concept of entrepreneurship is, in principle, applicable to legal decision-making. Decisions on which course to follow in a given case, and on which sources to rely, can be supposed to involve entrepreneurial judgments (2002, p. 8). Legal entrepreneurs, like their counterparts in the market, are alert to the "flaws, gaps and ambiguities in the law" (Krecké 2002, p. 10). Whitman (2002) also extends the idea of entrepreneurship to the role played by lawyers and litigants. He examines how legal entrepreneurs discover and exploit opportunities to change legal rules – either the creation of new rules or the re-interpretation of existing ones to benefit themselves and their clients. Harper (2013) believes that the entrepreneurial approach lays the groundwork for explaining the open-ended and evolving nature of the legal process – it shows how the structure of property rights can undergo continuous endogenous change as a result of entrepreneurial actions within the legal system itself.

The most important differentiating factor separating the entrepreneurship of the market process from legal entrepreneurship is the absence of the discipline of monetary profit and loss in the latter case. Although money may change hands in the process of legal entrepreneurship, its outputs may not be valued according to market prices, especially when there is a public-goods quality to the

rule at issue. Whether effective feedback mechanisms exist in the contexts is therefore an open question. Martin argues that, in such structures, the feedback mechanism in politics is not as tight as feedback in the market mechanism, and therefore, ideology plays a greater role in such decision-making (Martin 2010).

Legal entrepreneurship can be coordinating and yet also increase uncertainty and conflicts in society. It all depends on the kind of legal order in operation and the mechanism by which it is generated and maintained. Rubin (1977) and Priest (1977) originally analyzed how the openly competitive legal process tends to promote economic efficiency.

Conclusion

Austrian scholars extend the themes of ignorance, uncertainty, and fragmented knowledge, to the legal order; and this has important implications for their approach to law and economics. First, the legal rules cannot be assumed to be exogenously given; they must be evolved and discovered through a process. Second, legal systems can become important signposts enhancing expectational certainty, even if they are constantly evolving. And third, entrepreneurship is no longer restricted to within a given set of rules; entrepreneurs also operate in legal and political spheres attempting to create, change, and evolve rules. Finally, through these forces, legal institutions evolve spontaneously without a central mastermind.

Cross-References

- [Austrian School of Economics](#)
- [Common Law System](#)
- [Customary Law](#)
- [Efficiency](#)
- [Hayek, Friedrich August von](#)
- [Liberty](#)
- [Mises, Ludwig von](#)
- [Rule of Law and Economic Performance](#)
- [Smith, Adam](#)

References

- Epstein RA (1995) Simple rules for a complex world. Harvard University Press, Cambridge
- Ferguson A (1782 [1767]) An Essay on the History of Civil Society. 5th edition. London: T. Cadell
- Harper DA (2013) Property rights, entrepreneurship and coordination. *J Econ Behav Organ* 88:62–77
- Hayek FA (1937) Economics and knowledge. *Economica* 4(13):33–54
- Hayek FA (1944) The road to sefdom. Chicago University Press, Chicago
- Hayek FA (1945) The use of knowledge in society. *Am Econ Rev* 35:519–530
- Hayek FA (1973) Law legislation and liberty: rules and order (Vol. I, Law Legislation and Liberty). Chicago University Press, Chicago
- Hayek FA (2011 [1960]) Constitution of liberty: the definitive edition. Chicago University Press, Chicago
- Kirzner IM (1973) Competition and entrepreneurship. Chicago University Press, Chicago
- Kirzner IM (1992) The meaning of market process: essays in the development of modern Austrian economics. Routledge Press, London
- Kirzner IM (2000) The driving force of the market. Routledge, New York
- Krecké E (2002) The role of entrepreneurship in shaping legal evolution. *J des Econ et des Etudes Humaines* 12(2):1–16
- Lachmann LM (1971) The legacy of max weber. The Glendessary Press, Berkeley
- Leoni B (1991 [1961]) Freedom and the law. D. Van Nostrand Company Inc., Princeton
- Martin A (2010) Emergent politics and the power of ideas. *Stud Emergent Order* 3:212–245
- Menger C (1892) On the origin of money. *Econ J* 2(6):239–255
- Menger C (1963 [1883]) Problems of economics and sociology. Univeristy of Illinois Press, Urbana
- Mises L (1981[1922]) Socialism: an economic and sociological analysis. Liberty Fund, Indianapolis
- O'Driscoll GP, Rizzo MJ (1996) The economics of time and ignorance. Routledge Press, New York
- Priest GL (1977) The common law process and the selection of efficient rules. *J Leg Stud* 6(1):65–77
- Rizzo MJ (1980) The mirage of efficiency. *Hofstra Law Rev* 8(3):641–658
- Rizzo MJ (1990) Hayek's four tendencies toward equilibrium. *Cult Dyn* 3(1):12–31
- Rizzo MJ (1999) Which kind of legal order? Logical coherence and praxeological coherence. *J des Econ et des Etudes Humaines* 9(4):497–510
- Rizzo MJ, Whitman DG (2009) The knowledge problem of new paternalism. *Bingham Young University Law Review Issue* 4:905–968
- Rubin PH (1977) Why is the common law efficient? *J Leg Stud* 6(1):51–63
- Vaughn KI (1994) Austrian economics in America: the migration of a tradition. Cambridge University Press, New York

- Whitman DG (2002) Legal entrepreneurship and institutional change. *J des Econ et des Etudes Humaines* 12(2):1–11
- Zywicki TJ (2003) The rise and fall of efficiency in the common law: a supply side analysis. *Northwest Univ Law Rev* 97(4):1551–1633

Austrian School of Economics

Michael Litschka

Department of Media Economics, University of Applied Sciences St. Pölten, St. Pölten, Austria

Abstract

This entry describes the role of Austrian Economics as a branch of economics which has enriched economic theory with important concepts and theoretical alternatives. After focusing on main representatives and their ideas (specifically uncertainty, entrepreneurship, evolution, insufficient knowledge, and “rules”) it stresses some differences to mainstream neoclassical economics and possible applications to the discipline of Law and Economics. It argues that some discussions within Austrian Economics (e.g. social design of rules vs. evolutions of rules, the role of markets and institutions, and a specific understanding of efficiency and rationality) have a direct bearing on this discipline and should therefore be analyzed in more detail.

Synonym

[Austrian economics](#)

Definition

A branch of economics focusing on evolution and institutions, which deals with states of disequilibrium, time, uncertainty, and incomplete knowledge and is represented by Austrian and international scholars.

Introduction

The Austrian school of economics is a branch of economics founded by an Austrian economist Carl Menger and his major work “*Grundsätze der Volkswirtschaftslehre* (sic)” (1871). While mainly known for his contribution to marginal analysis in economic methodology, Menger put forward many other ideas his pupils in Austria and followers all over the world (but mainly in the USA) took over and developed further. Today it seems not clear whether “Austrian” economics and its major theories have been incorporated into the neoclassical mainstream or make up a theoretical stream of its own, like other heterodox approaches (e.g., Post-Keynesianism or new institutional economics). However it can be said that Austrian ideas have changed economic thinking in a specific way and have contributed to a better understanding of issues of uncertainty, causal relationships in economics, disequilibrium, the role of entrepreneurs, evolution, and informational issues in economic life. Moreover, Law and Economics as a discipline has profited from some of the insights of Austrian economics over time.

Representatives and Main Ideas

According to Leube (1995), we can distinguish at least seven generations of scholars in this tradition, beginning with Carl Menger and his pupils Eugen Böhm-Bawerk and Friedrich Wieser and later on comprising popular names like Ludwig Mises, Joseph Schumpeter, Friedrich Hayek, Gottfried Haberler, Fritz Machlup, Oskar Morgenstern, Israel Kirzner, Murray Rothbard, and Ludwig Lachmann. They all shared a methodological understanding of economic science as being based on “individualism,” i.e., all actions and intentions can be traced back to individual decisions, and “subjectivism,” i.e., all knowledge in economics must come from the subjective interpretation of the environment by individuals. “Human action” (also the title of an important contribution by Ludwig Mises) should stand in the center of any explanation of the social world.

It is important for Austrian economics (AE) that human actions are always determined by insufficient knowledge and the course of time, which is contrary to most of the early neoclassical economics (NE), where complete knowledge plays an important part in theory building (see also section “Differences to and Similarities with Mainstream Economics” in this entry).

In the famous argument about the right methodology for economics (the so-called German debate on methodology), i.e., whether the historical-inductive (collecting single data and historical specifications) or the causal-deductive (constructing general laws by logically deducing from basic axioms) method suits the discipline better, the Austrians favored the last one. In a famous passage (Menger 1871, pp. 25ff.), Menger explained the evolution of money through a process of market forces, where in the end money arises as institutionalized and government approved means of exchange, but without the will of a social planner beforehand. This focus on institutions arising as spontaneous orders in unhampered market processes continued throughout the history of AE, culminating in the work of Friedrich Hayek (1945, 1952) who explained the role of dispersed knowledge within society and the importance of the relative price system as guiding line for individual rational decisions.

The “subjectivist” view embraced by AE, i.e., that correct explanations of human behavior can only be had by analyzing the subjective sense an individual attaches to his or her action, has had implications for the understanding of costs, marginal values, expectations, cause-effect relationships over time, and the meaning of entrepreneurs. As such, this kind of subjectivism is a more general concept than, e.g., utility maximization by rational individuals in NE. This also explains the important tasks of entrepreneurs in the market as explained by Schumpeter, Kirzner, Menger, and the like. The “watchful” entrepreneur seizes opportunities given by the discovery of states of disequilibrium and fills the gap within such states. Subjectivism in this understanding also has some bearing for an understanding of modern rational choice theories which themselves are important for Law and Economics as a discipline. As Law

and Economics demands a fairly active role of judges in using economic reasoning to assess legal problems and the consequences of the arising rules, it seems clear that subjectivist analysis (like the evaluation of damages according to marginal cost analysis) is part of that task.

Lastly, the concept of spontaneous orders put forward by Hayek (1973) is a natural consequence of the above said. The system of rules a judge or lawyers have to accept and stick to was brought about by evolution: Rules prevail because they made a group successful; they were not adopted because of that group knowing their effects. Legal positivism (in the sense of deriving all law from the will of a law-making authority) is therefore a constructivist and unacceptable theory for AE; it is important that the common law judge (to name an example) infers general rules from precedents to apply them to new cases and thereby serves the legal order (the customs and rules evolution has developed). Law coordinates the actions of individuals according to general rules, and via this route and the spontaneously arising institutions, it also contributes to the efficiency of groups.

Differences to and Similarities with Mainstream Economics

There are mainstream neoclassical economists who believe that AE and NE share many premises, and therefore AE has been incorporated into NE (e.g., Stigler 1941). Both apply simplified logical models abstracting from complex actual events, both rely on methodological individualism, and both assume rational behavior by individuals. However, whereas NE stresses the predictive power of their theories, AE focuses on the explanatory power in their approaches. Also, when knowledge and rationality are concerned, NE concentrates on the conscious part of decisions, whereas AE deals with the meaning of tacit knowledge in such decisions.

Remarkable differences concern time and uncertainty. In AE, preferences and knowledge can change before a goal is achieved; in NE, knowledge is assumed to be perfect, at least in some defining approaches of the discipline (e.g.,

Gary Becker 1976). As regards the meaning of markets, the main difference is that Austrians concentrate on market processes and the role of the entrepreneur in the creative destruction of equilibria (see, e.g., Joseph Schumpeter's work), while neoclassicals come up with constrained optimization models to prove the states of equilibrium in markets. Applying these insights to Law and Economics, NE poses a stable framework of customary practices and legal rules, whereas AE sees continuous change and arising "spontaneous orders" (Friedrich Hayek) which develop over time and through the rational decisions by all individuals, in a so-called evolution of rules. As such an evolution is by definition an unintentional process, Austrian economists doubt the possibility of applying efficiency criteria (like Pareto or Kaldor-Hicks criteria) to legal rules, at least if they have quantitative implications. They would rather deem efficiency a criterion already implied in evolutionary processes. To give an example, Law and Economics scholar Posner and Austrian economist Hayek did not agree on the role of judges within the legal system, but share an understanding of the law having a built-in correctness (Posner 2005). While Hayek would see this correctness manifest in a spontaneous order, Posner would equal correctness with efficiency and use economic criteria for that task.

Contribution to Law and Economics

How has Law and Economics applied some of the "Austrian" insights mentioned above? If we take as example the methodological tenets of marginalism and subjectivism, one application is provided by the theory of efficient breach of contract. Roughly stated, a person A sells a good to person B but then finds another person C who is willing to pay more for this good. An efficient allocation of resources would require this good to be transferred to C (when there are equal wealth restraints), so A should breach the contract and pay B damages, as long as those damages are not too large (e.g., Posner 1998, p. 133). Other examples would be the protection of property rights as an incentive to produce goods and information,

the task of judges to subjectively calculate the efficiency of rules (see above), issues of deregulation, and the fact that courts (and the law) have to mimic market processes whenever the market fails to optimally allocate resources.

While these and other points could be made in favor of a positive influence AE may have exercised on Law and Economics (see Litschka and Grechenig (2010), for a more detailed account), there are also "Austrian" criticisms directed toward a neoclassical understanding of Law and Economics: There are doubts, e.g., as to whether the legislator or court can have the necessary information to postulate an efficiency solution, a claim that Posner or Becker would readily make. Rather, law should provide for space for spontaneous orders which result from human action, but not human design.

Considering the problem of incomplete information and the problem of social design of rules and institutions, how could an "Austrian" version of Law and Economics look like? Generally speaking, it would focus on those institutions best suited to promote decentralized decision making, especially on customs, social norms, and legal rules devised by evolution. The government's task is, among other things, to specify clear and transparent property rights and keep markets free from state influence. Judges and legislators should then fill the gaps in existing rules which have evolved out of social economic activity. To enable people to make their own subjective evaluations of state of affairs is the normative criterion economic policy should adhere to (and not sticking to the usual Pareto criteria of NE) in order to internalize external effects the right way.

Conclusion

As Litschka and Grechenig (2010) argue, it was essential for an economic approach to be accepted in legal theory to show that the law as it should be is already the law properly understood. Hayek's theory of legal evolution constituted one important basis for such a transformation of law, finding its way to, e.g., Posner's theory on the efficiency of

legal rules. Inefficient rules endangering the social welfare of the litigating parties would be challenged by the parties and thereby less likely to survive. Transformations of such a kind allowed Law and Economics to develop as a discipline and enter the daily work of lawyers and judges. To support such normative claims, a positive theory of legal institutions and their evolution, states of disequilibrium, time issues, uncertainty, and informational asymmetries under incomplete knowledge was necessary. This is exactly what (among other important inputs from, e.g., new institutional economics) the Austrian school of economics provided.

Cross-References

- [Becker, Gary S.](#)
- [Choice Under Risk and Uncertainty](#)
- [Coase and Property Rights](#)
- [Economic Analysis of Law](#)
- [Efficiency](#)
- [Hayek, Friedrich August von](#)
- [Institutional Economics](#)
- [Mises, Ludwig von](#)
- [Posner, Richard](#)
- [Rothbard, Murray](#)
- [Schmoller, Gustav von](#)

References

- Becker G (1976) The economic approach to human behaviour. The University of Chicago Press, Chicago
- Hayek F (1945) The use of knowledge in society. *Am Econ Rev* 35(4):519–530
- Hayek F (1952) Individualismus und wirtschaftliche Ordnung. Eugen Rentsch Verlag, Zürich
- Hayek F (1973) Law, legislation, and liberty. A new statement of the liberal principles of justice and political economy, vol 1, Rules and order. Routledge, London
- Leube K L (1995) Die österreichische Schule der Nationalökonomie. Texte. Band 1: von Menger bis Mises. Manz, Wien
- Litschka M, Grechenig C (2010) Law by human intent or evolution? Some remarks on the Austrian school of economics' role in the development of law and economics. *Eur J Law Econ* 29:57–79
- Menger C (1871) Grundsätze der Volkswirtschaftslehre. Braumüller, Wien
- Posner R (1998) Economic analysis of law, 5th edn. Aspen, New York

- Posner R (2005) Hayek, law, and cognition. *NYU J Law Libert* 1:147–165
- Stigler G (1941) Production and distribution theories. Macmillan, New York

Avoidance

Jacob Nussim
Faculty of Law, Bar-Ilan University,
Ramat-Gan, Israel

Abstract

The term “Avoidance” avoids clear definition in the academic literature. It generally describes illegitimate reactions to legal rules that wholly or partially thwart their legal effect. Self-interested, utility-maximizing individuals are obviously expected to try avoiding legal restrictions and limitations to their behavior. Indeed, examples of avoidance in reality are abundant. Research on avoidance was initiated by the economics of crimes literature in the 1970's and developed since in two additional spinoffs of that literature – environmental economics and economics of tax evasion. These developments are organized and reviewed here. The focus is on the relationship between avoidance and deterrence.

Introduction

Avoidance is not a term of art and no consistent definition exists. It refers to individual reactions to legal rules that circumvent the legal consequences of these rules to some extent. But it does not encompass any type of reaction. Avoidance typically refers only to noncompliant or illegitimate reactions rather than to behavioral changes that are in compliance with the law or considered legitimate reactions to legal rules.

Drawing the line between avoidance behavior and non-avoidance reactions is a challenging task. Individual behavior may face expected legal outcomes that consequently affect individual behavior. Earning income subjects taxpayers to income

taxes, which in turn may induce individuals to reduce their effort to earn income or to underreport their earned income. Driving activity is subject to various safety rules, which in turn make individuals drive more safely, drive less, or try to avoid the traffic police. Regulating the amount of air, water, and soil pollution may limit production and pollution, or it may induce production and pollution processes that are more difficult to observe and monitor. Individuals react to expected positive or negative legal consequences. Whether or not legal rules are regulatory, i.e., intended to change behavior, they trigger individual reactions. In particular, individuals change their behavior to reduce utility-decreasing legal consequences and to take advantage of utility-increasing ones. But not all these reactions are considered avoidance. If the behavioral reaction is intended by the social planner, it is clearly socially desirable. If the reaction is not intended or is not consistent with the goals of legal rules, it is commonly considered socially undesirable and represents a social cost, although it is not necessarily perceived illegitimate or noncompliant. For example, in response to income taxation, individuals may work less, change their tax status (e.g., from a corporation to a partnership), move investments to other countries, or register their business offshore. None of these responses are intended by income taxation; all of them are (locally) welfare reducing, but not all are commonly considered illegitimate. Another example is regulation of air pollution, which may induce polluters to move their operations elsewhere, pollute water rather than the air, misreport the extent of their pollution, change their product lines, litigate with the environmental protection agency, or capture regulators. Again, although none of these responses are intended by pollution regulation, not all of them are commonly considered avoidance.

Economists shy away from this type of discussion and assume certain individual reactions are avoidance. They typically neglect the definitional difficulty and treat avoidance as distinct, assuming that a clear legal dividing line exists between allowed and disallowed responses to legal rules, in other words, that a legal test can be applied. Legal scholars attempt to draw a line between

legitimate and illegitimate responses to law, but are yet to offer a satisfactory theory. For example, economists and legal scholars who analyze tax systems adopt various definitions, but provide no justification for their choices.

Ignoring the definitional problem, the literature on avoidance has developed along three largely parallel paths: the economics of crime, environmental economics, and economics of taxation. The last one is concerned with revenue raising and therefore engages in both efficiency and distributive analyses, whereas the former two are regulatory in nature and focus on efficiency. Economists and legal scholars do not necessarily distinguish between enforcing offenses such as theft, murder, speeding, or insider trading and regulatory misconduct such as pollution. All are considered externalities by economists and crimes by legal scholars. But whereas most crimes are controlled through quantity regulation, various other control instruments are considered and used in controlling pollution (see also Shavell 2011). Therefore, the analyses of crime enforcement and of environmental enforcement have also developed in parallel. In the present chapter, I discuss these strands in the literature, focusing on the more recent development of costly avoidance in the economics of crime literature.

The economics literature provides conventional examples of avoidance activities. In the tax literature, misreporting the tax base (tax evasion) is the leading example of costless avoidance. Other tax planning examples, such as revising the tax entity, changing tax location, and non-arm's-length transactions (e.g., transfer pricing), are abundant in the tax avoidance literature (see, e.g., ___tax avoidance papers___). The environmental economics literature also uses the misreporting example in its models of costless avoidance, as well as other examples of costly avoidance, such as changing the mode of operation upon inspection or making surveillance difficult by purchasing private land around the polluting facility (see, e.g., Heyes 1994). In the economics of crime literature, we find examples such as covering incriminating evidence, lying, fleeing the scene of the crime, moving to other countries, committing perjury, destroying

evidence, falsifying financial books or other documentations, concealing assets, hoarding cash, moving money to offshore banks, litigating, and using sophisticated expert advice (e.g., legal, accounting, financial, and medical).

To these examples, we should add the largely independent political economics (of regulation) literature, which generally begins with Stigler (1971). This body of literature is grounded in observations of various activities undertaken by private entities who try to avoid the effects of regulatory rules by influencing politicians, administrators, or the judiciary. (The political economics literature is voluminous and not reviewed here. See, e.g., Persson and Tabellini (2000), Grossman and Helpman (2001), and Besley (2004).) Another general example that may be considered avoidance is substituting away from legally controlled to uncontrolled or lightly controlled behavior (see also Nussim and Tabbach 2008b below). For example, controlling air pollution may encourage replacing air-polluting production activities with water- or soil-polluting processes; punishing the shredding of document for the purpose of obstructing justice may induce new retention policies with more frequent shredding or limited documentation (Chase 2003); and punishing producers who do not reveal safety information to consumers may also dissuade them from collecting such information (Kronman 1978; Farrell 1986).

Economic analysis of avoidance can be organized along several recurring features, emphasized in the following discussion. First, positive versus normative analysis involves analyzing the response of individuals to legal rules against designing an optimal public mechanism to cope with avoidance behavior. Typically, positive models are more refined and may better conform to reality, whereas normative models, for reasons of tractability, are more general. Second, avoidance activities may exhibit various characteristics. They may be costless or costly; avoidance measures may be observable or not and hence punishable and controllable or not; the costs of controlling avoidance vary (e.g., ex ante vs. ex post control); avoidance measures may affect the probability of punishment, its magnitude, the

effectiveness of public enforcement effort, etc.; and certain avoidance activities may be more suitable for price controls, while others to quantity or other control instruments. Some of these features are investigated in the literature.

This entry proceeds as follows. The first section “[Crime and Avoidance](#)” – sets the stage historically and traces the development of the economic research of avoidance in the tax and environmental literatures. The second section “[Tax Avoidance and Tax Evasion](#)” – briefly introduces the distinction between avoidance and evasion in the tax literature and fits it into the literature on avoidance. The following sections “[Costless Avoidance](#)” and “[Costly Avoidance](#)” – discuss the move from costless to costly avoidance, focusing on public and environmental economics. The next section “[Economics of Crime and Costly Avoidance](#)”, goes into the recent developments in the economics of crime literature on costly avoidance. Section “[Avoidance and Self-Reporting](#)” takes the logical step of connecting avoidance with self-reporting, and the last section on “[Avoidance in Private Law](#)” extends the observation and analysis of avoidance into private law.

Crime and Avoidance

The economics of crime began largely with Becker’s (1968) seminal work, in which he used economic tools to explain criminal behavior, and welfare economics to design crime enforcement schemes. Generally, incentives to engage in crime are affected by the relative market prices of legal and illegal activities. Crime generates externalities (harm) and hence represents a market failure that results in inefficiency. Therefore, controlling criminal behavior by central planning may be desirable. Crime control relies on several instruments, such as punishment and enforcement effort, changing the opportunity costs of crime, self-reporting subsidies, education, and more (Ehrlich 1973, 1975).

Avoidance was introduced into economic analysis in the 1970s, mainly as part of the economic analysis of crime, environmental

economics, and economics of tax evasion. (Avoidance was also discussed in other types of economic analysis, but not extensively. The emerging political economy literature of the 1970s also dealt, among others, with similar issues.) Although all three types of literature originated in Becker's (1968) deterrence-based framework of the economics of crime, they developed independently, probably because of their different focus of interests. The economics of crime literature is centered on deterrence of externality-producing behavior, whereas the focus of tax and environmental studies is different.

Tax and public economics scholars are interested mostly in optimal taxation and tax-related questions, in particular, inefficiency and distribution. Hence, the economic analysis of tax evasion developed around questions of labor supply (Pencavel 1979; Cowell 1981; Sandmo 1981; Weiss 1976), occupational choice (Pestieau and Posse 1991; Kolm and Larsen 2004) and empirical evidence by Parker 2003; Feldman and Slemrod 2007), tax rates and tax schedules (Srinivasan 1973). See also empirical evidence by Clotfelter (1983), Feinstein (1991), Alm et al. (1993), and Joulfaian and Rider (1996) and experimental findings by Friedland et al. (1978), Alm et al. (1992), and Baldry (1987), and redistribution (see, e.g., Christian 1994). Later, normative tax analysis, i.e., optimal taxation, naturally paid attention to both efficiency and distribution (and even complexity), rather than only to efficiency, as is the case under economics of crime studies (Chander and Wilde 1998; Boadway and Sato 2000).

The treatment of avoidance in environmental economics, although based on a deterrence rationale, was typically related to the choice of control instruments. Pollution can be controlled by several legal instruments, in particular, taxes (i.e., prices), standards (i.e., quantities), or tradable permits. The environmental economics literature examined the effects of avoidance activities, particularly on the choice of control instruments (Downing and Watson 1974; Harford 1978, 1987; Lee 1984; Linder and McBride 1984; Khambu 1990).

This entry focuses on the effect of avoidance on deterrence and hence on the economics of

crime literature. I begin by briefly reviewing the tax and environmental literatures and then delve into the economics of crime, situating it in relation to the general development and analysis of avoidance.

Tax Avoidance and Tax Evasion

The tax literature has come up with two arguably distinct terms for illegitimate behavior: "tax avoidance" and "tax evasion." These have become terms of art among tax and public economics scholars. Both are differentiated from taxpayers' legitimate behavior or "real responses," but the dividing line between them is blurred, and therefore the definition of "tax avoidance" remains to be settled. Unlike real responses to changes in relative prices due to taxes, tax avoidance embodies "a variety of tax planning, renaming, and retiming activities whose goal is to directly reduce tax liability without consuming a different basket of goods" (Slemrod 2001, p. 119). Other economists use other tentative definitions (Slemrod and Yitzhaki 2002, p. 1428; Piketty and Saez 2013, p. 417; Mayshar 1991, p. 78; and Sandmo 2005, p. 645), whereas legal scholars and legal reality still differ (Gunn 1978; Isenbergh 1982; Bankman 2000; Weisbach 2002).

The dividing line between tax avoidance and tax evasion is also unclear. Whereas tax avoidance is considered legal, although illegitimate, tax evasion is illegal. Whether or not legality can function as a clear separating criterion (Weisbach 2002), economists fail to see the normative underpinning and hence meaningfulness of such a distinction (Cross and Shaw 1981; Slemrod and Yitzhaki 2002). Other economists define the difference between avoidance and evasion differently (Cowell 1990b).

Both tax evasion and tax avoidance are considered avoidance activities by the economics of crime literature, which can cause some confusion. Both conform to the definition of avoidance adopted in this entry: tax avoidance and tax evasion represent attempts to avoid the legal consequences of behavior. They seem to differ by whether they are punishable (tax evasion) or not

(tax avoidance) (see Nussim and Tabbach 2008b, in the crime enforcement context, and Neck et al. 2012 in the tax context). The punishable/non-punishable divide is related to the tax-originated legal/illegal divide, but the two are not congruent. In any case, the feasibility of sanctioning avoidance activities is important, and it is discussed below.

Another potential difference between tax evasion and tax avoidance is the cost of such behaviors. This difference does not necessarily manifest in reality, and economic modeling is also inconsistent in making such a distinction. Tax evasion is commonly characterized as underreporting of the tax base (e.g., income for revenue-based taxes or emissions for Pigouvian taxes). Underreporting involves either no direct costs or negligible costs. Tax avoidance, by contrast, is commonly described as requiring costly planning, advice, means, or actions. But this distinction is not a sharp one. In practice, cheating (evasion) may also require substantial investment of resources.

In sum, rather than making a decisive distinction between avoidance and evasion, legal reality and economic modeling have taught us that in thinking about avoidance and its impact on behavior and enforcement of legal rules, we should pay attention to the feasibility of sanctioning avoidance and to its private costs.

Costless Avoidance

Costless avoidance is probably best associated with reporting issues. The economics of crime is founded on the built-in assumption of non-reporting: criminals do not self-report their crimes, and therefore enforcement effort is required. The non-reporting of crimes appears to be a costless avoidance measure. Although notice that non-reporting can generate risk-bearing costs, more effective enforcement, or emotional burdens (Polinsky and Shavell 1979; Mayshar 1991). But the literature mostly ignored these costs of non-reporting. Non-reporting was adopted into enforcement-based economic studies of taxation and environmental control and was

naturally extended into a continuous measure of avoidance in the form of partial reporting, which is more suitable for these issues (See also Bebachuk and Kaplow 1993 for heterogeneous costless avoidance).

Allingham and Sandmo (1972) and Srinivasan (1973) offered positive models of tax evasion – i.e., non-reporting or concealment of taxable income – which prompted a rich literature on tax evasion (See Surveys by Cowell 1990b; Andreoni et al. 1998; Slemrod and Yitzhaki 2002; Sandmo 2005). Notice that unlike non-reporting of crimes or environmental misbehavior, non-reporting in tax issues constitutes the criminal behavior.) The tax evasion literature is a spin-off of the positive analysis of crime. It is formulated as a portfolio model in which a taxpayer makes allocation choices, the returns to which are partially uncertain (Schmidt and Witte 1984). In these models, a taxpayer must allocate income between legal reporting and illegal evasion or allocate effort between observable-taxable and non-observable-nontaxable income-producing behavior, and at times must also allocate effort between taxable income-producing behavior and nontaxable leisure production.

At its inception, the tax evasion literature focused on the choice of legal and illegal activity in the form of reporting or misreporting taxable income. In light of the costless evasion of taxes – i.e., the costless misreporting of income – it examined the effects of various designs of ex post punishment (or tax rates) and enforcement effort on taxpayers' behavior (e.g., income production, occupational choice, misreporting), assuming different types of risk preferences (Pencavel 1979; Cowell 1981; Sandmo 1981; Weiss 1976; Yitzhaki 1974; Pestieau and Posen 1991), or various enforcement responses by the tax authority (Reinganum and Wilde 1985, 1986). Only later was the tax enforcement literature incorporated into normative models of optimal tax theory. First, the normative analyses assumed away the costs of avoidance, i.e., tax evasion (Sandmo 1981; Cremer and Gahvari 1996), but subsequently the costs of avoiding punishment were accounted for, i.e., tax avoidance (Usher 1986; Mayshar 1991).

The environmental regulation literature also used crime enforcement theory and incorporated avoidance activities. Similarly to the tax evasion literature, environmental studies first considered avoidance in its costless form, as misreporting of pollution levels. Unlike the tax literature, however, the environmental literature focused on comparing environmental control instruments, in particular Pigouvian taxes, quantity standards, and tradable permits, and examined the effect of costless avoidance activities on the choice of control instruments (Downing and Watson 1974; Harford 1978, 1987; Viscusi and Zeckhauser 1979; Jones 1989; Garvie and Keeler 1994).

Costly Avoidance

Assuming that avoidance consumes real resources conforms better to observations of behavior and therefore allows for a much richer set of circumstances. The potential use and effect of avoidance has already been mentioned by Isaac Ehrlich (1972, 1973), who further developed Becker's theory. Diligent readers have noted not only that Beccaria (1764) and Bentham (1789) had laid the groundwork for Becker's (1968) neoclassical economic treatment of crime, but that Beccaria (1764) also acknowledged the likelihood of reactive avoidance behavior (Sanchirico 2006, p. 1350 n. 64). The first serious investigation of Ehrlich's observation is in Malik's (1990) influential study, discussed in greater detail below.

Anticipating, in a way, Malik's (1990) contribution, both the tax avoidance and environmental economics literatures addressed costly avoidance in the mid-1980s. In environmental economics, Lee (1984) studied the positive effects and normative implications of costly avoidance on the design of affluent taxes. Linder and McBride (1984) examined the expected effects of costly avoidance on the choice of control instrument under an agency-hierarchy framework. These studies were followed by further positive analyses in the environmental literature. Khambu (1989) showed that increased regulatory standards may reduce compliance because of engagement in costly avoidance and that larger fines or more

stringent enforcement may not improve compliance. He also compared taxes and quantity standards given regulated entities invest in costly avoidance (Khambu 1990). Shaffer (1990) analyzed nonlinear penalties in enforcement of standards in the face of costly avoidance by firms. Nowell and Shogren (1994) examined the effect of costly avoidance on the enforcement of illegal dumping of hazardous waste. Heyes (1994) investigated the difference between thoroughness of inspection and probability of inspection under costly avoidance activity by polluting firms. Huang (1996) modeled costly avoidance and examined its positive effect on the control of externalities, using quantity standards and emission charges.

The tax literature introduced costly avoidance into both positive and normative analyses. Cowell (1990a) studied tax reporting behavior with risky evasion and costly avoidance. Slemrod (2001) incorporated costly tax avoidance into a positive model of labor supply under a linear wage tax, and examined how the availability of tax avoidance activity as a substitute for income-producing effort affects the extent of both tax avoidance and labor. Usher (1986) applied costly avoidance and evasion to a normative model and focused on their effect on the marginal costs of funds, and hence on investment in public goods. Mayshar (1991) demonstrated the normative consequences of costly tax avoidance in a cost-benefit framework in terms of its effect on the marginal costs of public funds.

Kaplow (1990) and Cremer and Gahvari (1993) analyzed the effect of costly tax evasion and enforcement using optimal commodity tax setups. Both studies focused on optimal taxation rather than the economics of crime enforcement, and hence the effect of evasion on enforcement was not emphasized. They showed, *inter alia* how costly evasion affects the optimal mix of taxes and enforcement effort. Their analyses, however, implicitly reveal a Malik-like effect of costly evasion on enforcement. Kaplow (1990) modeled enforcement of evasion with no specific penalties, because in his model taxpayers face no enforcement uncertainty: they know in advance whether they are going to be audited and pay taxes or not

audited and evade them. (This feature in Kaplow's model limits its analogy with crime enforcement. To put it in the terms of crime enforcement and accordingly embed uncertain enforcement in the model, taxpayers either evade taxes successfully and do not pay taxes, or they pay the same tax rate whether they do not evade or evade unsuccessfully.) Larger investment in enforcement reduces evasion, whereas lower private marginal costs of evasion increase its rate. Kaplow showed that a higher tax rate increases wasteful investment in evasion and therefore should be lowered. (Note that Kaplow's model ignores the social value of tax revenue (i.e., production of public goods); it parallels the common assumption in the economic analysis of crime enforcement.) He additionally showed that enforcement effort may also be limited by potential inducement of additional evasion.

Cremer and Gahvari (1993) modeled uncertain enforcement in a crime-like fashion, but they collapsed the expected penalty into an expected tax rate and therefore did not differentiate between their effects. They also collapsed the costs of evasion into product prices and exogenously limited the penalty rate, assuming away Becker's (1968) main result to begin with. As a result, we cannot draw clear results about the effect of penalties on evasion from their model. But they generally showed that a higher tax rate, which may stand for a higher penalty, increases evasion and, hence, should be limited. Cremer and Gahvari (1994) are similar, but in an optimal income tax framework. Costly evasion is socially wasteful and as such affects optimal tax rates and enforcement. Again, the penalty for evasion was treated as constant.

The normative analysis of tax avoidance is either rather abstract or quickly becomes quite complicated. The reason is that unlike the economics of crime literature, which focuses on efficiency only, optimal taxation theory also necessarily involves social preferences for the distribution of utility (Slemrod 1994; Cremer and Gahvari 1994). Indeed, enforcement of tax evasion and avoidance may affect distributive outcomes (Reinganum and Wilde 1985; Slemrod and Yitzhaki 1987; Border and Sobel 1987;

Cremer and Gahvari 1996; Chander and Wilde 1998).

Economics of Crime and Costly Avoidance

Isaac Ehrlich (1972, 1973), one of the founders of the economics of crime literature, identified the potential effect of crime control on individual incentive to avoid punishment, and hence on deterrence: "an increase in [punishment]... will generally increase an offender's incentive to spend resources on self-protection... This may decrease the probability of his being apprehended and punished which in turn may at least partly offset the deterrent effect of [punishment]" (Ehrlich 1972, p. 266). (Using the terminology of Ehrlich and Becker (1972), avoidance measures that reduce the probability of punishment are called "self-protection," and measures that diminish the magnitude of punishment are called "self-insurance.") But his observations did not attract the attention of economics of crime scholars and went largely unnoticed in the literature until the 1990s, beginning with Malik's (1990) important work.

Malik (1990) incorporated the costs of avoiding criminal punishment into the basic normative model of enforcement. In his model, sanctioning crime not only deters criminal activity but also induces individuals who engage in crime to invest real resources in reducing the probability of punishment. Therefore, to the extent that individuals engage in crime, their expected investment in avoidance, which is generally considered socially wasteful, should be weighed in the design of enforcement policy, that is, the punishment and enforcement effort. In particular, given avoidance, Becker's prescription of maximum fines is not necessarily socially desirable. Indeed, certain normative analyses of crime and avoidance after Malik also emphasized the required conditions for Becker's prescription under avoidance (Langlais 2008, 2009; Innes 2001). Thus, incorporating avoidance into crime enforcement policy expands its functions. Crime enforcement does not only aim at deterring crime, but is also

responsible for minimizing socially wasteful avoidance. Although more lenient fines are socially costly in terms of (under-) deterrence, they also reduce the incentive to engage in avoidance, which saves socially valuable resources.

After Malik, the economics of crime literature expanded and revised Malik's analysis in several ways, incorporating different modeling or assumptions. Avoidance is analyzed using normative models similarly to Malik (Langlais 2008, 2009; Stanley 1995a; Innes 2001; Tabbach 2009) or by adopting positive frameworks (Nussim and Tabbach 2008a, b, 2009; Friehe 2011; Baumann and Friehe 2013). The latter typically allow for a more accurate description of reality at the expense of having no rigorous normative conclusions. Various characteristics of avoidance have been studied. For example, Malik (1990), Langlais (2008), and Friehe (2011) assumed the existence of a self-protection measure of avoidance (or partial self-protection in Langlais (2009)), which is non-punishable. Nussim and Tabbach (2009) allowed for both self-protection and self-insurance measures of avoidance, which is again non-punishable. Stanley (1995b), Nussim and Tabbach (2008a), and Sanchirico (2006) assumed that avoidance is observable and punishable, whereas Nussim and Tabbach (2008b) and Baumann and Friehe (2013) assumed it is only partially observable. The implications of avoidance for various substitutable public tools have also been examined (Nussim and Tabbach 2009). Naturally, various assumptions reflect reality in different manners and imply different behavioral effects and predictions.

Nussim and Tabbach (2005, 2009) incorporated avoidance activities into the common positive modeling of crime, where individuals choose an allocation of time, effort, or wealth between legal and illegal activities (Ehrlich 1973), and showed that harsher punishment of crimes may encourage rather than deter criminal activity. (See also Stanley (1995b), who shows that under a specific structure of avoidance with increasing returns to scale, increasing the punishment of crime may induce more criminal acts.) The important observation inferred from their model is that crime and avoidance are complements in the sense

that more crime increases the marginal benefit of avoidance and thus triggers more investment in avoidance; similarly, larger investment in avoidance reduces the marginal costs of criminal activity and thus further encourages crime. Therefore, although harsher punishment of offenses directly deters potential offenders, it also fosters investment in avoidance, which in turn, indirectly, encourages criminal activity. The final outcome of these two opposing forces may be of more, rather than less crime owing to the harsher punishment. The authors showed that this counterintuitive outcome holds true under various assumptions concerning avoidance and enforcement technology as well as the offender's personal characteristics. Although Nussim and Tabbach (2009) extended the basic positive, well-known model of crime, their counterintuitive results may have confused some scholars. For example, Sanchirico (2012), using a common normative economic model of crime, unfortunately reached erroneous conclusions. Note further that Nussim and Tabbach's (2005, 2009) theoretical analysis can provide an explanation for the mixed empirical results on the deterrent effect of punishment (Nagin 2013; Chalfin and McCrary 2017). Clearly, other explanations may apply as well, such as risk-preferring offenders (Ehrlich 1973), choice of leisure (Schmidt and Witte 1984), or marginal deterrence effects (Stigler 1970).

Friehe (2011) extended Nussim and Tabbach (2009) to include forfeiture of illegal gains. He showed that given avoidance activity by offenders, forfeiting illegal gains, similarly to increasing punishment, may increase crime rather than deter it. The same mechanism and rationale as in Nussim and Tabbach (2009) apply.

The complementarity between crime and avoidance is missing in Malik (1990) because of the way in which his model is constructed, and therefore no indirect effects of enforcement are considered. For example, under his model, deterrence improves with increased punishment. Relatedly, less than maximal punishment and under-deterrence are optimal. Nussim and Tabbach (2009) noted that if complementarity is acknowledged, social optimality may actually entail harsher punishments and over-deterrence.

The reason is that although harsher punishment directly encourages avoidance, it also directly reduces crime, which due to complementarity also discourages avoidance. If the indirect effect of punishment on avoidance is sufficiently strong, over-deterrence is optimal.

The possibility of optimal over-deterrence due to avoidance is also suggested in a normative framework by Langlais (2008) (see also Langlais 2009). Langlais reconsiders Malik's framework by assuming that avoidance and enforcement effort are complementary, whereas Malik assumed that they are independent. This assumption has several consequences. In particular, contrary to what Malik argues, optimal enforcement may produce over-deterrence. Langlais also shows that punishment and enforcement effort may become complements, rather than substitutes, because of avoidance. The reason is that larger punishment induces avoidance, which in turn, given complementarity, induces enforcement effort. (Note that although Langlais' complementarity assumption is not impossible in reality, it is far from being straightforward. It requires that any additional public investment in enforcement increases the marginal effect of an investment in avoidance on the probability of punishment. Unfortunately, Langlais provides no real-life examples to support this assumption. See also a short discussion in Nussim and Tabbach (2005, 2009).)

Nussim and Tabbach (2005, 2009) also showed that investment in avoidance provides a relative advantage to other policies over punishment in deterring crime; these policies are enforcement effort and subsidizing legal alternatives. Enforcement effort deters crime without affecting the complementarity between crime and avoidance, and as such also reduces avoidance. Moreover, certain enforcement techniques may reduce the marginal effectiveness of the avoidance effort, further reducing avoidance, which in turn further reduces crime. For example, investing in the quality of inspections or audits, rather than in their quantity, may reduce the marginal effectiveness of avoidance activities. Following Nussim and Tabbach (2005, 2009), Sanchirico (2006) reiterated the point of the

relative advantage of enforcement effort and provided a legal viewpoint and valuable examples of legal procedures. Clearly, legal procedures are not the only relevant set of enforcement activities.

Another policy tool is subsidizing legal alternatives to crime, such as work subsidies, job training, education, and vocational programs. Subsidies to legal work increase the opportunity cost of crime, reducing crime, but have no direct influence on avoidance. Because crime and avoidance are complements, wasteful avoidance is indirectly reduced by such subsidies. Therefore, Nussim and Tabbach concluded that both enforcement effort and subsidies are socially costly, but they can deter crime and save on socially wasteful avoidance activities. Punishment generally requires negligible costs, but it may have a high social cost expressed in crime rate and wasteful avoidance.

The empirical literature on deterrence generally supports the conclusions of these avoidance-driven results, such as enforcement effort and incentives for legal alternatives are better deterrent tools than punishment. Empirical studies of crime deterrence show not only that the effect of punishment on crime is unclear or insignificant but that both enforcement effort and legal alternatives generate consistent deterrent effects.

Langlais (2009) constructed a "self-protection" model under which avoidance activities by offenders, upon apprehension, may only reduce potential forfeiture of benefits gained by crime, but do not affect fines. Based on this assumption, avoidance is clearly independent of fines, and therefore fines are not constrained by potential wasteful avoidance, as in Malik's model. Becker's result of maximum fines is thus restored. Note that Langlais abstracted from offenders wealth and therefore ignored the increase in the offenders' wealth and in the potential fines due to benefits gained from crime. Taking the offenders' wealth into account in this manner would change his results. Put differently, forfeiture of illegal gains is a part of punishment; therefore avoidance of forfeiture should not be different in general from avoidance of fines. See also Friehe (2011). Admittedly, the economic literature typically ignores benefits gained from

crime when considering the offenders' wealth, in particular because benefits added to the offenders' utility (they are consumed immediately) rather than to their wealth. This assumption is usually innocuous (Polinsky and Shavell 2007). But the case is different when benefits can be forfeited and only partially so, as in Langlais's model.

Malik implicitly assumed that avoidance is non-observable and therefore cannot be controlled directly. Several studies, however, examine the effects of punishing or regulating observable avoidance activities. Stanley (1995a) assumed that avoidance activities are costlessly observable and therefore can be directly punished. He showed that it is both possible and socially desirable to completely eliminate avoidance by an appropriate sanction. Note that Stanley (1995a) assumed that offenders are identical and that their choice is binary (either engage or not engage in crime), with no intensive margin of decision.

Nussim and Tabbach (2008a) extended the positive model of crime and avoidance to allow for either ex ante or ex post punishment of avoidance. Ex ante punishment (or regulation) of avoidance can take the form of a tax levied on avoidance activity or of restrictions on its use (e.g., prohibiting its consumption or requiring a license). Ex ante punishment of avoidance is independent of its use in criminal activity or its enforcement. For example, taxing radar detectors is independent of their use in speeding or of any punishment imposed if speeding is detected. Ex post punishment of avoidance is imposed upon detection of avoidance activity, such as perjury, and can be either related or unrelated to enforcement of a "principal" crime. Hence, punishments for detected crimes and avoidance activities can be either independent or interdependent. Such and other descriptions of reality have been investigated by Nussim and Tabbach. They showed first that ex ante punishment or regulation of avoidance is superior to ex post punishment because it necessarily deters both avoidance and crime (through their complementarity, as shown in Nussim and Tabbach (2009)). Furthermore, they showed that punishing avoidance ex post is not trivial. Unless carefully designed, sanctioning avoidance can actually encourage investment in it

because, although harsher punishment discourages it by increasing its price, it also increases the return on investment in avoidance if its punishment can be avoided. Moreover, crime deterrence may also be impaired by sanctioning avoidance, in particular, due to complementarity between avoidance and crime. Thus, Nussim and Tabbach delineated the critical features in the design of socially beneficial ex post punishment of avoidance.

In another study, Nussim and Tabbach (2008b) reconsidered the effect of sanctioning avoidance when certain avoidance activities are non-punishable. For example, although financial advice and legal litigation are generally used for legitimate purposes, they may also be used for avoidance purposes in a non-distinguishable manner and therefore are not punishable in practice. The authors showed that if offenders can engage in non-punishable avoidance activities, the control of punishable avoidance is not necessarily desirable. Based on the reasonable assumption that various avoidance activities are substitutes, (although Complementarity is not ruled out), ex ante punishment (i.e., taxation) of avoidance may improve deterrence and even reduce investment in non-punishable avoidance (due to its complementarity with crime). But it may also dilute deterrence and encourage investment in substitutable avoidance. Nussim and Tabbach carefully investigated the scenarios in which these different outcomes can be expected due to ex ante or ex post control of avoidance.

Baumann and Friehe (2013) examined measures that can provide legal consumption utility or facilitate avoidance (e.g., legal and illegal use of DVD writers). Unlike Nussim and Tabbach (2008b), who assumed that these measures are non-punishable, Baumann and Friehe allowed for their control either ex ante or ex post. They reached similar results to Nussim and Tabbach (2008a) on the difference between ex ante and ex post control of avoidance, but showed further that it may also be socially desirable to control/punish such measures. The intuition is that although taxing or regulating these measures distorts the individuals' legal behavior, it may also deter crime. Taxing (subsidizing) measures

that are complementary (substitutable) to crime improves deterrence.

In an informal discussion, Sanchirico (2006) noted that punishing avoidance is not different from punishing the “principal” offense in Malik’s framework. In other words, punishing avoidance may induce investment in additional avoidance activities in order to avoid the detection and punishment of avoidance. Given multiple hierarchical avoidance activities, Sanchirico showed that punishing avoidance – and similarly, the principal crime – is not trivial. Although analytically correct, the application in reality of “avoidance of avoidance of avoidance” (and so on) is questionable.

Lastly, whereas the economics of crime literature assumes that avoidance is socially costly and therefore is always undesirable, Tabbach (2009) showed that avoidance may actually be desirable, and, counterintuitively, it should be encouraged in certain circumstances. It is not that avoidance is not socially wasteful – it is. But it may be less socially wasteful than the social costs of punishment. Tabbach showed that if punishment is costly to society (e.g., imprisonment), avoiding punishment saves the costs of imposing punishment, but may still function as a socially desirable deterrent, because avoidance activities impose costs on offenders. In other words, costly avoidance may serve as a substitute for costly punishment; both are costly to offenders and therefore act as a deterrent, but the latter is also costly to society.

Avoidance and Self-Reporting

Strongly related to crime avoidance is the literature on self-reporting. Where law enforcement is required, eliciting self-reporting can be socially valuable for several reasons. First, self-reporting saves on enforcement costs for a given level of deterrence. With self-reporting, the government can monitor fewer individuals for the same probability of detection (Malik 1993; Kaplow and Shavell 1994). Second, self-reporting transforms uncertain punishment into a certain sanction and thus saves on risk-bearing costs (Kaplow and

Shavell 1994). Third, self-reporting can improve social welfare where remediation is required after an offense. Self-reporting, then, induces higher incidence of remediation for the same level of deterrence (Innes 1999). Fourth, under heterogeneous probabilities of apprehension, self-reporting can improve deterrence (Innes 2000).

Innes (2001) studied the effects of incorporating self-reporting into Malik’s model of avoidance and showed that with costless self-reporting, avoidance can be eliminated, restoring Becker’s prescription for maximum punishment. Intuitively, raising punishment to the maximum increases investment in wasteful avoidance, as stressed by Malik, but offering an alternative equal punishment to self-reporters generates self-reporting and therefore no avoidance.

Although not directly discussed by Innes, it appears that there is a strong inherent connection between avoidance and self-reporting. (Stanley (1995b) also made the connection between self-reporting and avoidance, although he did not pursue it rigorously.) Self-reporting indicates a lack of avoidance. Additionally, whereas avoidance is commonly modeled as reducing the probability of punishment, self-reporting increases this probability. Only that self-reporting is typically costless, and it is modeled as a corner outcome in which the probability of punishment equals one, and enforcement costs are zero.

Avoidance in Private Law

There is no reason to restrict the study of avoidance to public law (i.e., taxation and regulation). The rules of private law can be avoided in much the same way as those of public law, the difference being that the interaction is between private entities rather than opposite the government. (Innes (2001) also discussed the application of avoidance and self-reporting to tort law.)

Friehe (2009) incorporated avoidance activities into a unilateral care tort model and showed that injurers’ potential investment in avoidance activities makes the negligence regime superior

to strict liability. The intuition is that under an optimal negligence regime, the injurer pays no damages, and therefore he invests either in optimal precautions or in avoidance. Given a low effectiveness of avoidance, the injurer prefers optimal behavior. Under an optimal strict liability regime, the injurer must pay compensation and therefore may invest in both low-effectiveness avoidance measures and lower-than-first-best precautions. (Friehe's (2009) intuition is reminiscent of Buchanan and Tullock (1975).) Friehe showed further that even for high-effectiveness avoidance, a second-best design of a negligence regime can outperform strict liability. The intuition is similar: there is always a care standard that makes precautions cheaper than the consequences of investing in avoidance.

Friehe (2010) further investigated the effects of avoidance on various aspects of an optimal negligence regime. He showed that punitive damages should also account for their expected effect on avoidance, in a manner similar to that of punishment in Malik's crime enforcement model. He further showed that uncertain care standards reduce the incentive to avoid and thus may be superior to certainty. Lastly, Friehe examined the effect of avoidance on the choice of compensation under negligence: full compensation for harm or only harm due to negligence (Grady 1983; Kahan 1989). He showed that full compensation induces avoidance less frequently, but when it does, a full compensation regime produces a larger investment in avoidance and lower levels of care.

Conclusion

Avoidance is ubiquitous in reality. Indeed, we should expect utility maximizing individuals to circumvent legal rules for their own benefit. Economic analyses of avoidance attempt to explain various kinds of avoidance behavior, to predict avoidance reactions, and to suggest normative responses by a social planner. But the avoidance literature in public law and, in particular in private law, is still rather scarce, and further research is indispensable.

References

- Allingham MG, Sandmo A (1972) Income tax evasion: a theoretical analysis. *J Public Econ* 1(6):323–338
- Alm J, Jackson BR, McKee M (1992) Estimating the determinants of taxpayer compliance with experimental data. *Natl Tax J* 45:107–114
- Alm J, Bahl R, Murray MN (1993) Audit selection and income tax underreporting in the tax compliance game. *J Dev Econ* 42(1):1–33
- Andreoni J, Erard B, Feinstein J (1998) Tax compliance. *J Econ Lit* 36:818–860
- Baldry JC (1987) Income tax evasion and the tax schedule: some experimental results. *Public Finance* 42(3): 357–383
- Bankman J (2000) The economic substance doctrine. *Calif Law Rev* 74:5, 13
- Baumann F, Friehe T (2013) Regulating harmless activity to fight crime. *J Econ* 113(1):1–17
- Bebchuk LA, Kaplow L (1993) Optimal sanctions and differences in individuals likelihood of avoiding detection. *Int Rev Law Econ* 13:217–224
- Beccaria C (1764) On crimes and punishments. In: Manzoni A (ed) The column of infamy prefaced by Cesare Beccaria's of crimes and punishments (trans: H. Paolucci, 1963). Prentice Hall, Oxford University Press, 1963
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Bentham J (1789/1948) An introduction to the principles of morals and legislation (JH Burns, HLA Hart eds., 1970). (Oxford: Blackwell, 1948 ed.), Athlone Press, pp 158–159
- Besley T (2004) Principled agents? Motivation and incentives in government. The Lindahl Lectures (Oxford University Press, 2005)
- Boadway R, Sato M (2000) The optimality of punishing only the innocent. *Int Tax Public Financ* 7:641–664
- Border KC, Sobel J (1987) Samurai accountant: a theory of auditing and plunder. *Rev Econ Stud* 54(4): 525–540
- Buchanan JM, Tullock G (1975) Polluters' profits and political response: direct controls versus taxes. *Am Econ Rev* 65:139–147
- Chalfin A, McCrary J (2017) Criminal deterrence: a review of the literature. *J Econ Lit* 55(1):5–48
- Chander P, Wilde LL (1998) A general characterization of optimal income tax enforcement. *Rev Econ Stud* 65:165–183
- Chase CR (2003) To shred or not to shred: document retention policies and federal obstruction of justice statutes. *Fordham J Corp Financ Law* 8:721
- Christian CW (1994) Voluntary compliance with the individual income tax: results from the 1988 TCMP study. In: The IRS Research Bulletin, 1993/1994, Publication 1500 (Rev. 9–94). Internal Revenue Service, Washington, DC, pp 35–42
- Clotfelter CT (1983) Tax evasion and tax rates: an analysis of individual returns. *Rev Econ Stat* 65(3):363–373

- Cowell FA (1981) Taxation and labour supply with risky activities. *Economica* 48(192):365–379
- Cowell FA (1990a) Tax sheltering and the cost of evasion. *Oxf Econ Pap* 42:231–243
- Cowell FA (1990b) Cheating The Government: The Economics of Evasion
- Cremer H, Gahvari F (1993) Tax evasion and optimal commodity taxation. *J Public Econ* 50(2):261–275
- Cremer H, Gahvari F (1994) Tax evasion, concealment and the optimal linear income tax. *Scand J Econ* 96(2):219–239
- Cremer H, Gahvari F (1996) Tax evasion and the optimum general income tax. *J Polit Econ* 60(2):235–249
- Cross RB, Shaw GK (1981) The evasion-avoidance choice: a suggested approach. *Natl Tax J* 34:489–491
- Downing PB, Watson WD Jr (1974) The economics of enforcing air pollution controls. *J Environ Econ Manag* 1(3):219–236
- Ehrlich I (1972) The deterrent effect of criminal law enforcement. *J Leg Stud* 1:259–276
- Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. *J Polit Econ* 81:521–565
- Ehrlich I (1975) On the relation between education and crime. In: Juster FT (ed) *Education, income and human behavior*. McGraw-Hill, New York, pp 313–337
- Ehrlich I, Becker GS (1972) Market insurance, self-insurance, and self-protection. *J Polit Econ* 80(4):623–648
- Farrell J (1986) Voluntary disclosure: robustness of the unraveling result, and comments on its importance. In: Grieson RE (ed) *Antitrust and regulation*. Lexington Books, Lexington, pp 91–103
- Feinstein JS (1991) An econometric analysis of income tax evasion and its detection. *Rand J Econ* 22(1):14–35
- Feldman NE, Slemrod J (2007) Estimating tax non-compliance with evidence from unaudited tax returns. *Econ J* 117(518):327–352
- Friedland N, Maital S, Rutenberg A (1978) A simulation study of income tax evasion. *J Public Econ* 10(1):107–116
- Friehe T (2009) Precaution v. avoidance: a comparison of liability rules. *Econ Lett* 105:214–216
- Friehe T (2010) On avoidance activities after accidents. *Rev Law Econ* 6:3
- Friehe T (2011) A note on the deterrence effects of the forfeiture of illegal gains. *Rev Law Econ* 7(1):118–124
- Garvie D, Keeler A (1994) Incomplete enforcement with endogenous regulatory choice. *J Public Econ* 55(1):141–162
- Grady MF (1983) A new positive economic theory of negligence. *Yale Law J* 92:799–829
- Grossman GM, Helpman E (2001) *Special interest politics*. MIT Press, Cambridge, MA
- Gunn A (1978) Tax avoidance. *Mich Law Rev* 76:733–767
- Harford JD (1978) Firm behavior under imperfectly enforceable pollution standards and taxes. *J Environ Econ Manag* 5:26–43
- Harford JD (1987) Self-reporting of pollution and the firms behavior under imperfectly enforceable regulations. *J Environ Econ Manag* 14(3):293–303
- Heyes AG (1994) Environmental enforcement when ‘inspectability’ is endogenous: a model with overshooting properties. *Environ Resour Econ* 4(5):479–494
- Huang C-H (1996) Effectiveness of environmental regulations under imperfect enforcement and the firm’s avoidance behavior. *Environ Resour Econ* 8(2):183–204
- Innes R (1999) Remediation and self-reporting in optimal law enforcement. *J Public Econ* 72:379–393
- Innes R (2000) Self-reporting in optimal law enforcement when violators have heterogeneous probabilities of apprehension. *J Leg Stud* XXIX:287–300
- Innes R (2001) Violator avoidance activities and self-reporting in optimal law enforcement. *J Law Econ Org* 17:239–256
- Izenbergh J (1982) Musings on form and substance in taxation. *Univ Chic Law Rev* 49:859–884
- Jones CA (1989) Standard setting with incomplete enforcement revisited. *J Policy Anal Manage* 8:72–87
- Joulfaian D, Rider M (1996) Tax evasion in the presence of negative income tax rates. *Natl Tax J* 49(4):553–570
- Kahan M (1989) Causation and incentives to take care under the negligence rule. *J Leg Stud* 18:427–447
- Kaplow L (1990) Optimal taxation with costly enforcement and evasion. *J Public Econ* 43(2):221–236
- Kaplow L, Shavell S (1994) Optimal law enforcement with selfreporting of behavior. *J Polit Econ* 102(3):583–605
- Khambu J (1989) Regulatory standards, compliance and enforcement. *J Regul Econ* 1(2):103–114
- Khambu J (1990) Direct controls and incentives systems of regulation. *J Environ Econ Manag* 18(2):72–85
- Kolm A-S, Larsen B (2004) Does tax evasion affect unemployment and educational choice? SSRN papers
- Kronman A (1978) Mistake, disclosure, information, and the law of contracts. *J Leg Stud* 7(1):1–34
- Langlais E (2008) Detection avoidance and deterrence: some paradoxical arithmetic. *J Public Econ Theory* 10:371–382
- Langlais E (2009) On the ambiguous effects of repression. *Ann Econ Stat* 93/94:349–361
- Lee DR (1984) Monitoring and budget maximisation in pollution control. *Econ Inq* 18:565–575
- Linder SH, McBride ME (1984) Enforcement costs and regulatory reform: the agency and firm response. *J Environ Econ Manag* 11(2):327–346
- Malik AS (1990) Avoidance, screening and regulatory enforcement. *Rand J Econ* 21(3):341–353
- Malik AS (1993) Self-reporting and the design of policies for regulating stochastic pollution. *J Environ Econ Manag* 24:241–257
- Mayshar J (1991) Taxation with costly administration, The Scandinavian Journal of Economics 93:75–88
- Nagin DS (2013) Deterrence: a review of the evidence by a criminologist for economists. *Annu Rev Econ* 5(1):83–105

- Neck R, Wachter JU, Schneider F (2012) Tax avoidance versus tax evasion: on some determinants of the shadow economy. *Int Tax Public Financ* 19(1):104–117
- Nowell C, Shogren JF (1994) Challenging the enforcement of environmental regulation. *J Regul Econ* 6:265–282
- Nussim J, Tabbach AD (2005) Avoidance and deterrence, (October 20, 2005). Available at SSRN: <http://ssrn.com/abstract=927216>
- Nussim J, Tabbach A (2008a) Controlling avoidance: ex-ante regulation v. ex-post punishment. *Rev Law Econ* 4:Article 4
- Nussim J, Tabbach A (2008b) (Non)regulable avoidance and the perils of punishment. *Eur J Law Econ* 25: 191–208
- Nussim J, Tabbach A (2009) Deterrence and avoidance. *Int Rev Law Econ* 29:314–323
- Parker SC (2003) Does tax evasion affect occupational choice? *Oxf Bull Econ Stat* 65:379–394
- Pencavel JH (1979) A note on income tax evasion, labor supply, and nonlinear tax schedules. *J Public Econ* 12(1):115–124
- Persson T, Tabellini G (2000) Political economics: explaining economic policy. MIT Press, Cambridge
- Pestieau P, Posse UM (1991) Tax evasion and occupational choice. *J Public Econ* 45(1):107–125
- Piketty T, Saez E (2013) Optimal labor income taxation. In: *Handbook of public economics*, vol 5. (Elsevier) pp 391–474
- Polinsky AM, Shavell S (1979) The optimal tradeoff between the probability and magnitude of fines. *Am Econ Rev* 69:880–891
- Polinsky AM, Shavell S (2007) The theory of public enforcement of law. In: *Handbook of law and economics*, vol 1. (Elsevier), pp 403–454
- Reinganum JF, Wilde LL (1985) Income tax compliance in a principal–agent framework. *J Public Econ* 26(1):1–18
- Reinganum JF, Wilde LL (1986) Equilibrium verification and reporting policies in a model of tax compliance. *Int Econ Rev* 27(3):739–760
- Sanchirico CW (2006) Detection avoidance. *N Y Univ Law Rev* 81:1331–1399
- Sanchirico CW (2012) Detection avoidance and enforcement theory. *Proced Law Econ* 8:145
- Sandmo A (1981) Income tax evasion, labor supply and the equity–efficiency tradeoff. *J Public Econ* 16:265–288
- Sandmo A (2005) The theory of tax evasion: a retrospective view. *Natl Tax J* 58:643–663
- Schmidt P, Witte AD (1984) An economic analysis of crime and justice. Academic, Orlando, pp 142–216
- Shaffer S (1990) Regulatory compliance with non-linear penalties. *J Regul Econ* 2(1):99–103
- Shavell S (2011) Corrective taxation versus liability as a solution to the problem of harmful externalities. *J Law Econ* 54(4):S249–S266
- Slemrod J (1994) Fixing the leak in Okun’s bucket: optimal tax progressivity when avoidance can be controlled. *J Public Econ* 55(1):41–51
- Slemrod J (2001) A general model of the behavioral response to taxation, *International Tax and Public Finance* 8:119–128
- Slemrod J, Yitzhaki S (1987) The optimal size of a tax collection agency. *Scand J Econ* 89(2):183–192
- Slemrod J, Yitzhaki S (2002) Tax avoidance, evasion, and administration. In: *Handbook of public economics*, vol 3. (Elsevier), pp 1423–1470
- Srinivasan TN (1973) Tax evasion: a model. *J Public Econ* 2:339–346
- Stanley TJ (1995a) Radar detectors, fixed and variable costs of crime. Unpublished working paper
- Stanley TJ (1995b) Optimal penalties for concealment of crime. Unpublished working paper
- Stigler GJ (1970) The optimum enforcement of laws. *J Polit Econ* 78:526–536
- Stigler GJ (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 2:3–21
- Tabbach A (2009) The social desirability of punishment avoidance. *J Law Econ Org* 26:265–289
- Usher D (1986) Tax evasion and the marginal cost of public funds. *Econ Inq* 24:563–586
- Viscusi WK, Zeckhauser R (1979) Optimal standards with incomplete enforcement. *Public Policy* 27: 437–456
- Weisbach DA (2002) An economic analysis of anti-tax-avoidance doctrines. *Am Law Econ Rev* 4(1): 88–115
- Weiss L (1976) The desirability of cheating incentives and randomness in the optimal in-come tax. *J Polit Econ* 84(6):1343–1352
- Yitzhaki S (1974) A note on ‘income tax evasion: a theoretical analysis’. *J Public Econ* 3(2):201–202

B

Backhaus Jürgen: A Pioneer of Law and Economics in Europe

Alain Marciano^{1,2} and Giovanni Battista Ramello^{3,4}

¹MRE and University of Montpellier, Montpellier, France

²Faculté d'Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France

³DiGSPES, University of Eastern Piedmont, Alessandria, Italy

⁴IEL, Torino, Italy

Abstract

Jürgen Backhaus is one of the founders of law and economics in Europe. He is a specialist in public finance and history of economic thought. His role in law and economics was important not only as a scholar who published (and edited) tens of books and articles but also as an entrepreneur.

Biographical Sketches

In July 2015, Jürgen Backhaus retired from his position as Professor at the University of Erfurt, after a long and rich career, he started in 1970 at the University of Constance as an undergraduate student, and that brought him all around the

world. He was Professor of Public Finance at the University of Maastricht from 1986 to 2001, and, from 2001 to 2015, he held the Krupp Foundation Chair in Public Finance and Fiscal Sociology at the University of Erfurt. Jürgen Backhaus published (and edited) tens of books and articles in the fields of public choice, fiscal sociology, and law and economics. He was, and still is, a scholar and a man of great immense culture. But even within academia, Backhaus is not only a scholar. He also played the role of a cultural entrepreneur. He launched, in 1994, the *European Journal of Law and Economics*, edited an important reference book, the *Elgar Companion to Law and Economics*, and then had the idea of this *Encyclopedia of Law and Economics* as an ambitious project gathering together around the word large part of the law and economics community and including ourselves as editors. He organized for decades one of the first and long-lasting European workshops in law and economics (first at the University of Maastricht and then at the University of Erfurt) and an interdisciplinary workshop in Heilbronn (Marciano and Ramello 2016).

The Contribution

In one sentence Jürgen Backhaus played a seminal role in law and economics and especially in its development in Europe. In particular, he

was fundamental in reconnecting the Chicagoan tradition of law and economics to its European roots and in turn the rational choice approach to political economy. Those roots were then somewhat lost. Jürgen Backhaus is one of those scholars putting back in evidence the legacy of the discipline to the important debates that took place in Europe during the eighteenth and the nineteenth century, and that not only provided the ground for growing law and economics as a new discipline but also have been fundamental in shaping the physiognomy of modern Western societies (Ramello 2016). In other words, Jürgen Backhaus has the merit of having brought again in evidence the link between Europe and the USA.

That claim was clearly put forward in the aims of the *European Journal of Law and Economics* he tried in particular to shed light on the “European Community law and the comparative analysis of legal structures and legal problem solutions in member states of the European Community [...] and] the new European market economies” while preserving a “pragmatic position informed by the history of law and economics as a scholarly discipline and the current challenges this discipline faces.” (Backhaus and Stephen 1994). That is to say, his scholarship adopted a novel political economy approach to law and economics, and in this respect his contribution far exceeds the simple diffusion of law and economics in Europe. In other terms, his commitment was not only aimed at fostering the consciousness of the European roots of the discipline but equally at nurturing a European approach to it.

The reference to the continental European economic thought and its connection with law and economics has always been central to Backhaus. He always tried to match the current development of the discipline to what the great thinkers already discussed (Marciano and Ramello 2014). This was evidenced in the *European Journal of Law and Economics* but also in one of the other major publishing adventure of Backhaus, the *Elgar Companion to Law and Economics* (1999,

2005). A crossroads for favoring the blending between the American and the European scholarship of law and economics thanks to the contributions of members of the two communities, a significant part is devoted to introduce the work of important and somewhat neglected scholars like Cesare Beccaria, Friedrich List, Gustav von Schmoller, and Rudolf von Jhering, among others (Marciano and Ramello 2016). To Backhaus, going back to these scholars and to their works was necessary to enrich our understanding of the same policy questions that are still in debate nowadays.

This does not imply, however, that Backhaus ignored that the political context and the culture do matter. Much to the contrary. He was convinced that individuals do not behave in the same way in every environment. Behaviors are not changed when the conditions and culture change. Human action is contextualized. As a consequence, the same institutions – or legal rules – cannot be used in different places. Institutional arrangements are unique and linked to particular historical contexts (see, for instance, Boettke et al. 2013). Or, in other words, “different institutions are appropriate in different circumstances” (Djankov et al. 2003, 619). This was a leitmotiv in Jürgen Backhaus’s thinking (Josselin et al. 2016). This led him to develop an approach that is quite specific in law and economics, not heterodox but not mainstream either. Likewise his cultural background, his form of law and economics actually evidences a way of going back to political economy. His way of envisaging the dialogue between “law” and “economics” follows what James Buchanan Ronald Coase or Guido Calabresi did rather than that of Richard Posner. Not because the latter did not link their economic analysis of law to old thinkers – in that case, Bentham and even Beccaria – but because he tended to think that institutional arrangements can be duplicated and transplanted regardless of the context. Jürgen Backhaus’s work teaches us the reverse (l. 139) (see, for instance, Backhaus 2017).

Legacy

It is difficult to evaluate what will be Backhaus's legacy. He has undoubtedly influenced, directly and indirectly, a lot of scholars in law and economics, public choice, and public finance. What would remain after him is the importance of a form of *European* law and economics and the need to go back to old scholars and to anchor law and economics in their work to perpetuate a certain tradition.

Cross-References

- [Economic Analysis of Law](#)
- [Law and Economics](#)

References

- Backhaus J, (1999) *The Elgar Companion to Law and Economics*, Edward Elgar Publishing, Cheltenham/Northampton (second revised edition 2005)
- Backhaus JG (2017) Lawyers' economics versus economic analysis of law: a critique of professor Posner's "economic" approach to law by reference to a case concerning damages for loss of earning capacity. *Eur J Law Econ* 43:517–534. (first publication, 1978, *Munich Social Science Review*)
- Backhaus JG, Stephen FH (1994) The purpose of the *European journal of law and economics* and the intentions of its editors. *Eur J Law Econ* 1:5–7
- Boettke PJ, Coyne C, and Leeson P (2013) Comparative historical political economy. *J Ins Econ* 9(3):285–301.
- Djankov S, Glaeser E, La porta R, Lopez-de-silanes F, and Shleifer A, (2003) The new comparative economics, *Journal of Comparative Economics*, 31:595–619.
- Josselin JM, Marciano A, Ramello GB (2016) The law, the economy, the polity Jürgen Backhaus, a thinker outside the box. In: Marciano A, Ramello GB (eds) *Law and economics in Europe and the U.S. the legacy of Jürgen Backhaus*. Springer
- Marciano A, Ramello GB (2014) Consent, choice and Guido Calabresi's heterodox economic analysis of law. *Law Contemp Probl* 77:97–116
- Marciano A, Ramello G B (2016) Introduction. In: Marciano A, Ramello GB (eds) *Law and economics in Europe and the U.S. the legacy of Jürgen Backhaus*. Springer, 7–9.
- Ramello GB (2016) The past, present and future of comparative law an economic. In: Eisenberg T, Ramello GB (eds) *Comparative law and economics*. Edward Elgar Publishing, Cheltenham/Northampton

Banks

Danny Cassimon¹, Mark A. Dijkstra² and Peter-Jan Engelen^{2,3}

¹Institute of Development Policy and Management (IOB), University of Antwerp, Antwerp, Belgium

²Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

³Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

Definition

Banks are financial intermediaries between economic agents with surplus funds and those with a shortage of funds and convert short-term deposits into long-term loans.

Introduction

Banks are institutions that intermediate between agents (households, firms, the government) with surplus funds and those with a shortage of funds. As such, they manage to provide for safe and liquid savings opportunities as well as long-term credit facilities, without the need for those two parties to interact and negotiate a financial contract directly. In this way, they decrease search and other transaction costs, solve potential mismatches between demand and supply in terms of maturity and degree of liquidity, and aim at solving risk and asymmetric information problems. Ultimately, the purpose of financial intermediation by banks is to make finance available for investment, leading to economic growth and financial inclusion, i.e., access to financial services for all.

In the following, we first briefly describe the traditional model of (deposit-taking) banking, including the traditionally associated risks. Section "[Bank Regulation](#)" then discusses how the traditional banking model has changed in the lead up to the global financial

crisis, while section “[The 2007–2009 Global Financial Crisis](#)” refers to financial regulation put in place as a response to the financial crisis. Section “[Investment Banks, Shadow Banks and Fintech](#)” discusses non-traditional forms of banking while section “[Conclusion](#)” concludes.

Activities and Risks

A traditional deposit-taking bank is primarily deposit funded (from households and firms) and its assets are primarily made up of credits extended to the same sectors, using an originate-and-hold approach, meaning that these credits are typically held until maturity. For liquidity purposes, banks also hold cash and liquid instruments (reserves with the Central Bank (► [Central Bank](#)), government bonds). Commercial banks create money by issuing deposits and making loans, while only keeping a fraction of actual cash reserves on their balance sheet. Table 1 shows the composition of the balance sheet of JP Morgan Chase in 2017, the United States’ largest bank in that year. Note that like most banks, JP Morgan Chase is mainly funded by deposits and its biggest asset class is loans.

Banks’ activities can roughly be divided in brokerage (bringing together providers and users of capital) and qualitative asset transformation (Bhattacharya and Thakor 1993). These activities are typically combined, so that a bank takes deposits which are short-term, liquid, safe instruments, and transforms them into long-term, illiquid, and risky loans. While an individual

risk-averse depositor may not want to commit their savings in a long-term risky investment, a bank can pool a large number of deposits together and make long-term risky investments that earn a higher return. The bank will still be able to fulfill the deposit contract as long as its long-term investments pay off in regular intervals, freeing up liquidity for depositors when they want to take their money out of the bank (Diamond and Dybvig 1983; Gorton and Pennacchi 1990). Furthermore, banks are typically assumed to be better at overcoming information asymmetries than individual investors (► [Transaction Costs](#)), leading to a better allocation of credit by banks compared to individual investors (Holmström and Tirole 1997). Banks have a bigger incentive to actively monitor a loan contract compared to individual investors who only contribute a small part to a loan. Moreover, banks typically have a longer-running relationship with the borrower, leading to a possible information advantage as well as additional incentives to monitor over an investor that invests as a single-shot deal (Boot and Thakor 2002).

The transformation of maturity, liquidity, and risk may allow for better allocation of credit but also comes with risks, most notably the risk of a bank run (Mishkin and Eakins 2015). When depositors lose their trust in the ability of a bank to fulfill the deposit contract, they may withdraw their deposits. If enough depositors withdraw their money at once, the bank may be forced to liquidate its long-term contracts at a low price in so-called fire sales. This lowers the value of the bank’s assets, leading to additional withdrawals from depositors. As such, even a normally solvent

Banks, Table 1 Asset and liability composition JP Morgan Chase 2017 (percentage total assets)

Assets		Liabilities and equity	
Cash and due from banks	1	Deposits	57
Deposits with other banks	16	Wholesale repurchase borrowing	6
Loans (e.g., mortgages)	37	Long term debt	11
Wholesale repurchase lending	8	Trading liabilities	5
Securities	14	Other liabilities	11
Trading assets	15		
Other assets	9	Equity (incl. Retained earnings)	10
Total assets	100	Total liabilities and equity	100

bank may become bankrupt through a bank run as a self-fulfilling prophecy (Diamond and Dybvig 1983). Bank runs are highly disruptive and are at the forefront of financial crises, which can lead to severe recessions, such as the Great Depression and the Global Financial Crisis of 2007–2009 (Laeven and Valencia 2018).

Bank Regulation

Because of their role in operating payment systems and credit provision, bank operations are tightly regulated. Although exact regulation differs from country to country, most countries insure deposits, have equity and liquidity adequacy requirements, and put limits on combining the activities that banks can undertake.

Deposit insurance measures disincentivize bank runs by insuring deposits up to a maximum amount: Currently \$250,000 in the United States and €100,000 in EU member states (European Commission 2015). With deposit insurance depositors lose the incentive to both run on a bank and to monitor the bank, which in turn may create moral hazard: Banks then both hold fewer liquid assets as well as undertake more risky activities (Bhattacharya et al. 1998).

Countries also typically place minimum requirements on bank equity (capital requirements) and liquidity (liquidity requirements). On capital requirements, the Basel Committee on Banking Regulation provides guidelines, which also the United States and EU member states follow. Compared to Basel 1 and 2, the current iteration of Basel, Basel 3, has increased capital requirements and limits the use of internal risk models, which were a main feature of Basel 2. Basel 3 proposes an unweighted ratio of equity to assets of 3%, together with a ratio of equity to risk-weighted assets up to 13% (depending on the definition of equity, state of the economy and type of bank). Since bankruptcy can only occur when a company's assets are worth less than its debt liabilities, minimum equity requirements automatically protect against bank insolvency by raising equity and thus lowering debt. By including a ratio that defines

minimum equity in function of risk-weighted assets, higher average asset risk needs to be matched by more equity. Furthermore, more equity may prevent excessive risk taking resulting from debt overhang in banks (Admati et al. 2013).

Besides equity requirements, Basel 3 also proposes liquidity requirements: The Liquidity Coverage Ratio requires banks to hold enough liquid assets (such as cash and overnight deposits at the central bank) to cover a 30 day net cash outflow.

Different countries also implement restrictions on the portfolio composition of banks. In the United States, commercial banks were forbidden from undertaking investment banking activities and activities over state lines under the 1933 Glass-Steagall Act. EU member states had no such regulation and partially because of this, many banking sectors in the EU feature so-called universal banks which combine commercial and investment banking activities, while the US historically has had more specialized banks. This Glass-Steagall Act was partially repealed in 1999 by the Gramm-Leach-Bliley Act, which allowed US banks to both bank across state lines as well as undertake commercial and investment banking activities together.

Interest rate ceilings upheld through usury law are sometimes argued as an effective way to limit risk-taking by banks (Posner 1995). Usury laws were common among US states in the nineteenth century (Benmelech and Moskowitz 2010) but have since been replaced by other forms of regulation (such as equity requirements discussed above).

Indirectly, banks are also affected by competition law. Competing visions arise between competition-fragility and competition-stability (Beck 2008). The competition-fragility view notes that competition in banking may lead to excessive risk-taking to outcompete rivals, while the competition-stability view notes that market power may lead to higher interest rates with accompanying higher risk, while simultaneously creating banks that may enjoy implicit governmental subsidies from becoming too-big-to-fail. As such, it is unclear how financial stability is fully impacted by antitrust law.

Besides explicit regulation of banks, large interconnected banks can enjoy implicit governmental guarantees. Because of the adverse effects of a bank run, governments tend to be inclined to bail out banks when they would otherwise default, especially if they are big or highly connected to other banks, leading to banks that are too-big-to-fail or too-connected-to-fail. Like deposit insurance, bailouts disincentivize bank runs but create moral hazard. Because of deposit insurance and too-big-to-fail bailouts, depositors and other creditors have little incentive to monitor banks, because they will be compensated regardless of any risk the bank takes. Meanwhile, shareholders enjoy relatively high returns that come with risk during good economic times, while the government bears the risk of a downswing, leading to an appetite for risk-taking by the bank.

The 2007–2009 Global Financial Crisis

In the leadup to the 2007–2009 Global Financial Crisis, banks had over time strayed away from an originate-and-hold model towards an originate-and-distribute model through a process called securitization. In the process of securitization, banks repackage loans (e.g., mortgages) and sell them off to various other investors. Typically, these repackaged loans would then be sliced into tranches which differed by risk. The safest (super-senior) tranche would be paid first, while the junior tranche was paid out last. This junior tranche would normally be held by the issuing bank, giving the bank an incentive to monitor the

securitized loans as a form of insurance to the outside investors. In practice, the risk involved in these junior tranches would be insured through credit default swaps, an insurance against default. In theory, securitization and tranching of securities meant that risk would be borne by the parties that were most able to bear the risk. In practice however, the securitized products were rated too highly by credit rating agencies (► [Credit Rating Agencies](#)), while risks were underestimated and banks lost an incentive to monitor loans because they had insured their losses through credit default swaps (Brunnermeier 2009).

The trigger for the start of the crisis in 2007 was an increase in subprime mortgage defaults in the United States (Van Essen et al. 2013). These defaults directly led to a fall in the value of securitized loan products, which in turn rippled through an interconnected financial system. Northern Rock, a UK-based retail bank faced a bank run in September of 2007 and was eventually nationalized to prevent default. In March 2008, investment bank Bear Stearns was taken over by JP Morgan Chase to prevent bankruptcy. The crisis hit its peak in September 2008 with the fall of investment bank Lehman Brothers. In October 2008, the US Treasury announced a \$700 billion bailout plan, with similar bailout plans all over Europe. The financial crisis had a large impact on the banking sector. Table 2 shows a geographical shift of the biggest banks away from Europe towards Asia. Besides this geographical shift, the banking sector has been fundamentally altered by additional financial sector regulation (Hull 2018).

Banks, Table 2 Assets of the world's largest banks in 2007 and 2017

Biggest banks (2007)	Country	Assets (\$ bln)	Biggest banks (2017)	Country	Assets (\$ bln)
Royal Bank of Scotland	UK	3783	ICBC	CN	4009
Deutsche Bank	DE	2954	CCBC	CN	3400
BNP Paribas	FR	2477	ABC	CN	3236
Barclays	UK	2443	Bank of China	CN	2992
Crédit Agricole	FR	2068	Mitsubishi UFJ	JP	2785
UBS	CH	2007	JPMorgan Chase	US	2534
Société Générale	FR	1567	HSBC	UK	2522
ABN AMRO	NL	1499	BNP Paribas	FR	2357
ING Bank	NL	1453	Bank of America	US	2281
BoT-Mitsubishi UFJ	JP	1363	Crédit Agricole	FR	2117

Post-Crisis Regulation

In the wake of the financial crisis, regulation has focused on increased capital (i.e., equity) and liquidity requirements, the formulation of resolution schemes, and the reform of bank business models. In the USA, financial regulation in the wake of the financial crisis falls under the Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank), while in the EU financial regulation falls under the Bank Recovery and Resolution Directive (BRRD) and the Capital Requirements Directives (CRD IV) – Capital Requirements Regulation (CRR). Both follow recommendations made in Basel 3 by featuring higher capital (equity) requirements than before the crisis and feature explicit liquidity requirements, where Basel 2 had relied on supervisory review for liquidity coverage. Both the BRRD and Dodd-Frank also feature living wills, in which financial institutions have to submit a plan to the regulator on the orderly dissimulation of the bank in the case it defaults.

A unique feature of Dodd-Frank is the *Volcker rule*, which explicitly forbids deposit-taking banks from engaging in proprietary trading. Although there is some discussion on the feasibility of distinguishing proprietary trading from trading with the intent of hedging (Elliot and Rauch 2014), the rule attempts to limit risk-taking in banking and as such features elements that are similar to the 1933 Glass-Steagall Act. Similar proposals on bans on proprietary trading have been made in the EU (European Commission 2014) but have not been put into effect as of yet. As a special case, UK banks are subject to ringfencing: Large UK banks must separate their retail banking and investment activities. The activities can still be carried out under the name of a single bank, but the retail and investment bank divisions can only interact with one another at arm's length.

State Aid by national governments is not allowed within the European Union to prevent member states from favoring domestic companies. During the financial crisis, multiple EU

member states were forced to aid domestic banks in order to keep them from collapsing and causing severe recessions. In order to promote financial stability, competition policy (such as the policy on state aid) can be softened during times of financial crisis (Hasan and Marinc 2016). Since the financial crisis, the European Commission has started working on implementing a banking union within the EU. European banks now face a Single Supervisory Mechanism (SSM) and a Single Resolution Mechanism (SRM) with the intent of weakening the links between banks and the sovereign governments of the member states. Proposals on a European Deposit Insurance Scheme (EDIS) are a part of the banking union roadmap but have not been implemented as of yet.

Investment Banks, Shadow Banks, and Fintech

Delineating the borders of what constitutes a traditional bank is not always obvious as some financial intermediaries can act as complements or substitutes. *Investment banks* are private companies that cannot take deposits and that engage in finance related activities. Typical investment bank activities include raising funding, underwriting, trading in assets and derivatives, and assisting companies in issuing securities and supporting mergers and acquisitions. Traditionally, large investment banks such as Goldman Sachs followed a partnership model until the end of the twentieth century when many large investment banks went public. *Shadow banks* (► [Shadow Banks](#)) are financial intermediaries that provide bank-like services but are not classified as banks and as such fall outside of banking regulation (Claessens et al. 2012). Depending on the exact definition, shadow banks include investment banks, pension funds, money market funds, insurance companies, fintech companies and many more. Because shadow banks do not have a banking license, they are partially exempt from the regulation that banks face. However, they are also less likely to be bailed out, and they may be subject to other regulation, such as rules for insurance companies (Solvency regulation). *Fintech*

companies are somewhat of a catch-all term for nontraditional players that enter the market for financial intermediation with an emphasis on new technology, most notably data-analysis. Examples of fintech companies include Paypal in payment processing, Greensky in loan provision and Betterment in financial management. The biggest current fintech company with assets of \$150 billion in 2017 is China-based Ant Financial (former Alipay) which provides payment processing as well as wealth management and loan financing. It is unclear yet how traditional banks and fintech companies will interact (Boot 2017), although fintech may be instrumental in increasing financial inclusion in markets in which banks are relatively inactive, for example, through mobile banking.

Conclusion

Banks transform liquid short-term, risk-free deposits into long-term risky loans, which may optimize credit allocation but also opens up the possibility of bank runs. Deposit insurance schemes and governmental guarantees mitigate the possibility of bank runs but create moral hazard and risk-taking incentives in turn. Because excessive risk-taking by banks may lead to financial crises, tight regulation of banks is necessitated, even if regulation may never fully prevent another financial crisis.

Cross-References

- [Central Bank](#)
- [Credit: Rating Agencies](#)
- [Financial Regulation](#)
- [Shadow Banking](#)
- [Transaction Costs](#)

References

- Admati AR, DeMarzo P, Hellwig M, Pfleiderer P (2013) Fallacies, irrelevant facts, and myths in capital regulation: why bank equity is not socially expensive. Stanford University Working Paper Series no. 161
- Beck T (2008) Bank competition and financial stability: friends or foes? World Bank Policy Research Working Paper no. 4656
- Benmelech E, Moskowitz TJ (2010) The political economy of financial regulation: evidence from U.S. state usury laws in the 19th century. *J Financ* 65(3):1029–1073
- Bhattacharya S, Thakor AV (1993) Contemporary banking theory. *J Financ Intermed* 3(1):2–50
- Bhattacharya S, Boot AWA, Thakor AV (1998) The economics of bank regulation. *J Money Credit Bank* 30(4):745–770
- Boot AWA (2017) The future of banking: from scale & scope economies to fintech. *Eur Econ* 2(2):77–95
- Boot AWA, Thakor AV (2002) Can relationship banking survive competition? *J Financ* 55(2): 679–713
- Brunnermeier M (2009) Deciphering the liquidity and credit crunch 2007–2008. *J Econ Perspect* 23(1):77–100
- Claessens S, Pozsar Z, Ratnovsky L, Singh M (2012) Shadow banking: economics and policy. IMF Staff Discussion Note 12/12
- Diamond D, Dybvig P (1983) Bank runs, deposit insurance, and liquidity. *J Polit Econ* 91(3): 401–419
- Elliot DJ, Rauch C (2014) Lessons from the Implementation of the Volcker rule for banking structural reform in the European Union, Goethe University White Paper Series no. 13
- European Commission (2014) Proposal for the regulation of the European Parliament and of the Council on structural measures improving the resilience of EU credit institutions. Proposal 52014PC0043
- European Commission (2015) Towards a European Deposit Insurance Scheme. Five Presidents' Report Series Issue 01/2015
- Gorton G, Pennacchi G (1990) Financial intermediaries and liquidity creation. *J Financ* 45(1):49–71
- Hasan I, Marinc M (2016) Should competition policy in banking be amended during crises? Lessons from the EU. *Eur J Law Econ* 42(2):295–324
- Holmström B, Tirole J (1997) Financial intermediation, loanable funds, and the real sector. *Q J Econ* 106(1):1–39
- Hull JC (2018) Risk management and financial institutions, 5th edn. Wiley, Hoboken
- Laeven L, Valencia F (2018) Systemic Banking Crises Revisited. IMF Working Paper No. 18/206
- Mishkin FS, Eakins SG (2015) Financial markets and institutions, 8th edn. Pearson, Boston
- Posner EA (1995) Contract law in the welfare state: a defense of the unconscionability doctrine, usury laws, and related limitations on the freedom to contract. *J Leg Stud* 24(2):283–319
- Van Essen M, Engelen PJ, Carney M (2013) Does 'good' corporate governance help in a crisis? The impact of country- and firm-level governance mechanisms in the European financial crisis. *Corp Gov* 21:201–224

Becker, Gary S.

Jean-Baptiste Fleury

THEMA, University of Cergy-Pontoise, Cergy-Pontoise, France

Abstract

Gary Becker was an American economist who was awarded the Nobel Memorial Prize in 1992 for having expanded the scope of economics to such “outlandish” topics as crime, racial discrimination, education, addiction, fertility choices, and marriage and the family. A member of the Chicago school of economics, he was especially influential in the development of the economic analysis of law.

Biography

Gary S. Becker (December 2, 1930–May 3, 2014) was an American economist who won the Nobel Memorial Prize in economics in 1992. Becker was initially trained at Chicago as a labor economist in the early 1950s. His PhD dissertation applied microeconomics to an unusual topic for the time, discrimination. Becker notably showed that discriminatory practices were costly and, thus, would not necessarily survive to the process of market competition. In the late 1950s, he moved to Columbia, where he developed other controversial theories mobilizing microeconomics and rational choice applied to topics such as fertility, human capital, the allocation of time, as well as crime and law enforcement (Fuchs 1994). He returned to Chicago in the late 1960s, where he stayed for the remaining of his career. There, he kept on producing trailblazing contributions on topics such as marriage and the family, social interactions, sociobiology, and rational addiction. At the turn of the 1970s, Becker rose to prominence: during the period 1971–1985, he was the most widely cited economist (Medoff 1989).

Innovative and Original Aspects

The most distinguishing feature of Becker’s work is that it contributed to the movement called “economics imperialism” (Lazear 2000; Swedberg 1990), which expanded the scope of economics outside its traditional boundaries (i.e. the analysis of market transactions, wealth, etc.). He was awarded the Nobel Memorial Prize precisely “for having extended the domain of microeconomic analysis to a wide range of human behaviour and interaction, including nonmarket behaviour” (http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1992/). Thus, he was influential in promoting a definition of economics centered on its method rather than on its scope of analysis. Defined as “the combined assumptions of maximizing behavior, market equilibrium, and stable preferences, used relentlessly and unflinchingly” (Becker 1976, p. 5), his “economic approach” was grounded on his research on the allocation of time which eventually led him to build a renewed version of the theory of consumer behavior. Consumer behavior mimics the behavior of producers: the rational agent combines market goods and services with time and skills in order to produce nonmarketable “commodities” that deliver, ultimately, utility (McRae 1978; Fuchs 1994). This is considered as a key methodological foundation of the “Chicago school of economics,” making Becker one of its most influential members, alongside Milton Friedman, George Stigler, and a few others (Emmett 2010). In this strongly policy-oriented approach, the stability of preferences is a central assumption: the economist has to relate changes in behavior to changes in prices *only* (Emmett 2010). Thus, differences in consumption (and more generally in behavior) among individuals are more related to differences in the opportunity cost of time and other “shadow prices” of various activities than to differences in tastes (Lazear 2000). For instance, fertility decisions are dependent on the opportunity cost of time spent to raise a child, as well as the cost of education. Thus, as countries get richer and individuals better educated, the shadow price of children increases, reducing the “demand for large families” (see Becker’s Nobel

Prize Lecture (http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1992/becker-lecture.pdf)).

Impact and Legacy

Becker was influential in the development of traditional questions in labor economics, but in the development of new subfields as well. Among the most important ones are the economics of the family, the economics of education, and law and economics. Regarding the latter subfield, Becker's 1968 paper "Crime and Punishment: an Economic Approach" was highly influential, because it showed how economic analysis could be used to devise optimal laws, especially how to compute the levels of severity of punishment and of probability of conviction that would minimize the social cost of crime (Lazear 2000). But Becker's influence on the field of law and economics was even more pervasive. It derived from his achievements in broadening the scope of application of rational choice theory. Indeed, his approach to economics opened the door for an economic analysis of judicial behavior, as well as provided tools to analyze, from an economics point of view, numerous litigations involving discrimination at work, wage differentials, employment at will, no-fault divorce, inheritance, and taxation (Posner 1993).

Cross-References

► [Economic Analysis of Law](#)

References

- Becker GS (1976) The economic approach to human behavior. The University of Chicago Press, Chicago
- Emmett RB (2010) The Elgar companion to the Chicago school of economics. Edward Elgar, Cheltenham
- Fuchs VR (1994) Nobel Laureate: Gary S. Becker: ideas about facts. *J Econ Perspect* 8(2):183–192
- Lazear E (2000) Economic imperialism. *Q J Econ* 115(1):99–146
- McRae D (1978) The sociological economics of Gary S. Becker. *Am J Sociol* 83(5):1244–1258

- Medoff MH (1989) The ranking of economists. *J Econ Educ* 20(4):405–415
- Posner R (1993) Gary Becker's contributions to law and economics. *J Leg Stud* 22(2):211–215
- Swedberg R (1990) Economics and sociology, redefining their boundaries. Princeton University Press, Princeton

Jean-Baptiste Fleury is assistant professor at the University of Cergy-Pontoise. His research focuses on the history of post-WWII economics with a special interest for economics as a public science. His research also deals with "economics imperialism," the application of microeconomics to topics that traditionally belonged to other disciplines.

Behavioral Law and Economics

Haksoo Ko
School of Law, Seoul National University, Seoul,
Republic of Korea

Definition

Behavioral law and economics is a research field in law and economics, which employs methodologies borrowed from behavioral economics. Behavioral economics in turn combines economics and cognitive psychology and produces research results indicating, in general, that individuals' choices in various decision-making circumstances may depart from what can be predicted from traditional neoclassical economics due to psychological biases. Behavioral law and economics is a relatively new and growing field and has recently been applied to virtually all areas of law, generating interesting and sometimes provocative research results.

Behavioral Law and Economics

Behavioral law and economics is a relatively new field: academic writings began to appear in

earnest in the late century. Although it initially focused on a few legal topics, it has recently been applied to virtually all areas of law, generating interesting, and sometimes provocative, research results.

Behavioral law and economics is in some respects a natural offshoot from the development of behavioral economics. Behavioral economics grew as an academic field through the 1970s and 1980s. Important early contributions include Herbert Simon's work on bounded rationality and prospect theory developed by Daniel Kahneman and Amos Tversky. By the 1990s, many economists and psychologists joined the field and made important contributions. Kahneman and Tversky (2000) contain many of these contributions, and Camerer et al. (2004) contain relatively recent contributions. From around the mid-1990s, Cass Sunstein and Richard Thaler (from the 1980s in Thaler's case) began publishing, individually or jointly, a series of academic articles documenting various forms of cognitive biases and psychological limitations that are relevant to law and economics.

Research in behavioral economics continued vigorously through the 2000s. Some of the relatively new and interesting research results were published through edited volumes such as Frey and Stutzer (2007, delving into the concept of trust and applying developments in neuroscience to behavioral economics, among others) and Diamond and Vartiainen (2007, applying behavioral economics to public economics, development economics, health economics, organizational economics, and other subject areas of economics). Also, there is a vast academic literature related to financial economics, which has grown to form an independent field of research, i.e., behavioral finance. Different from this, a line of recent research in cognitive psychology developed into happiness studies, with some obvious applicability in law. For developments in happiness studies, see Posner and Sunstein (2010) and Frey (2008).

Building upon earlier research results in behavioral economics, by the late 1990s, an academic literature began to emerge that could be labeled as behavioral law and economics. Sunstein (2000) is a showcase containing important results of

academic research in this period. Throughout the 2000s, a growing number of researchers began participating, expanding considerably the overall horizon covered under the rubric of behavioral law and economics. Many academic journals began routinely publishing articles written from behavioral law and economics perspectives. Several notable edited volumes were also published, including Parisi and Smith (2005), Gigerenzer and Engel (2006), Farahany (2009), and perhaps most comprehensively to date Zamir and Teichman (2014). Behavioral law and economics is now applied to almost all areas of law, including contract law, tort law and regulation, property law, criminal law, corporate law, consumer protection, competition law, labor and employment law, environmental law, health law, dispute resolution and procedural law, tax law, and international law.

Behavioral law and economics in general employs methodologies and insights that are developed and used in behavioral economics. Methodologically, behavioral economics (and, by extension, behavioral law and economics) places an emphasis on obtaining results through empirical analyses. Such empirical analyses often involve conducting experiments in a laboratory setting, rather than analyzing large-scale statistical data. Documenting various cognitive biases and demonstrating specific instances where individuals' choices depart from rational choices have been the main contribution of behavioral economics to the existing economics literature. Researchers in behavioral law and economics sometimes conduct original experiments or other empirical studies themselves and, at other times, simply borrow results from behavioral economics and construct their arguments. These researchers often try to draw legal and policy implications from these studies and from various cognitive biases. More recently, researchers started to draw on insights from the social psychology literature at large, extending beyond the confines of cognitive biases. For an example of this line of research, applying social psychology scholarship on interpersonal interactions to the decision-making process in the corporate setting, see Langevoort (2012).

Many and a growing number of researchers have recently made contributions to the literature of behavioral law and economics, facilitating the development of a comprehensive and coherent theoretical framework. Many of the efforts have centered on cataloging circumstances under which an individual's decision-making systematically departs from what rational choice theory would predict. Cognitive biases that are commonly referred to in the academic literature include prospect theory, endowment effects, and framing effects. These biases tend to generate contextual judgment errors, and it has been shown that, due to these biases, an individual's decision for an identical set of choices may be different depending on the contextual circumstances. For instance, it has been demonstrated that an individual's decision may differ whether an identical set of choices is presented in a "gain" frame or a "loss" frame and that individuals tend to show "loss aversion." An implication of these findings is that the Coase Theorem may not work if contextual judgment errors are at play. There are different types of biases which are related to problems of self-assessment such as hyperbolic discounting, optimistic bias, availability heuristic, and impact bias. For an introduction and illustration of various forms of biases, see Jolls (2007), Wilkinson (2008), and Kahneman (2011).

Drawing on research results demonstrating various cognitive biases, some researchers have tried to derive policy prescriptions. Behavioral law and economics has indeed been applied to various legal areas where policies or regulations are relevant. A noteworthy recent example, with strong policy prescriptions applicable to the consumer credit market, is Bar-Gill (2012). Efforts to implement policy prescriptions have in fact gained a substantial momentum in certain countries, most notably the USA and the UK.

In general, an eventual policy goal seen from behavioral law and economics perspectives would be to assist consumers and other decision-makers in order to induce them to make genuinely informed decisions and, by doing so, to let them make choices that they would make in the absence of cognitive biases. This, however, would not necessarily mean providing as much information

as possible. Rather, supplying an adequate amount of information in an adequate choice framework would be important. Policy proposals run a gamut from making the choice framework more simplistic by, e.g., setting an appropriate default rule or making a choice set simpler, to a more direct and active intervention by the government. A major underlying tension in prescribing policy proposals would be between, on the one hand, giving sufficient guides and information to decision-makers inducing them to make decisions in their best interests and, on the other hand, not interfering with their inherent freedom to make choices and not distorting their preferences.

Behavioral law and economics has recently drawn considerable interests among many researchers in law and economics and has produced many noteworthy research results. At the same time, it also has had to face not a small amount of criticism. Such criticism is mostly concerning (1) the theoretical cohesiveness of various psychological biases and (2) the justifiability of policy prescriptions.

About the theoretical cohesiveness of various psychological biases, those who are critical raise a question as to how to link these biases together and to propose a consistent and cohesive theoretic framework. While they in general do not challenge the existence of various cognitive biases, they cast doubt on the capability of behavioral law and economics to go beyond merely compiling various biases and to come up with a theoretical framework which can eventually replace or supplement the framework of rational choice theory.

About policy prescriptions, the challenge would be to justify individual policy proposals. In this area, for the most part, the existence of psychological biases is not to be denied either. The real challenge would be to show a clear logical or empirical link and relevancy between cognitive biases, on the one hand, and policy prescriptions, on the other. In particular, heated debates were raged regarding the feasibility of the government's provision of paternalistic guidance or intervention for the benefit of individual decision-makers, while at the same time leaving these individuals with their freedom to make

choices. Included in this line of criticism is the claim that demonstrated biases may not be robust enough and may manifest themselves differently when exposed to actual market institutions.

Both types of criticism referred to above would perhaps reflect the state of development of behavioral law and economics as an independent research field. That is, while interests in behavioral law and economics grew dramatically in recent decades, it is still a relatively young academic field and as such lacks loads of accumulated research results. This, of course, does not mean that the field does not have a bright future prospect. On the contrary, as more and more results are accumulated, its theoretical and empirical foundations would become more robust and sophisticated and would be able to provide analytic framework that could supplement the more traditional rational choice theory.

References

- Bar-Gill O (2012) *Seduction by plastic*. Oxford University Press, Oxford
- Camerer CF, Loewenstein G, Rabin M (eds) (2004) *Advances in behavioral economics*. Princeton University Press, Princeton
- Diamond P, Vartiainen H (2007) *Behavioral economics and its applications*. Princeton University Press, Princeton
- Farahany NA (ed) (2009) *The impact of behavioral sciences on criminal law*. Oxford University Press, Oxford
- Frey BS (2008) *Happiness: a revolution in economics*. MIT Press, Cambridge
- Frey BS, Stutzer A (eds) (2007) *Economics and psychology: a promising new cross-disciplinary field*. MIT Press, Cambridge
- Gigerenzer G, Engel C (eds) (2006) *Heuristics and the law*. MIT Press, Cambridge
- Jolls C (2007) Behavioral law and economics. In: Diamond P, Vartiainen H (eds) *Behavioral economics and its applications*. Princeton University Press, Princeton, pp 115–155
- Kahneman D (2011) *Thinking fast and slow*. Farrar, Straus and Giroux, New York
- Kahneman D, Tversky A (eds) (2000) *Choices, values, and frames*. Cambridge University Press, Cambridge
- Langevoort DC (2012) Behavioral approaches to corporate law. In: Hill CA, McDonnell BH (eds) *Research handbook on the economics of corporate law*. Edward Elgar, Cheltenham
- Parisi F, Smith VL (eds) (2005) *The law and economics of irrational behavior*. Stanford University Press, Stanford
- Posner EA, Sunstein CR (eds) (2010) *Law and happiness*. University of Chicago, Chicago
- Sunstein CR (ed) (2000) *Behavioral law and economics*. Cambridge University Press, Cambridge
- Wilkinson N (2008) *An introduction to behavioral economics*. Palgrave Macmillan, Hampshire
- Zamir E, Teichman D (2014) *The Oxford handbook of behavioral economics and the law*. Oxford University Press, New York

Behavioral Law and Economics of Contract Law

► Limits of Contracts

Black Economy

► Shadow Economy

Blackmail

Oleg Yerokhin
University of Wollongong, Wollongong,
Australia

Abstract

The criminality of blackmail has been debated vigorously by legal theorists and economists for a considerable amount of time. Yet to date there appears to be no universally agreed upon theoretical rationale for making blackmail illegal. This entry reviews the theory of blackmail developed in the law and economics literature over the last 40 years and some of the critiques that were leveled against it. It argues that the social efficiency considerations emphasized in the economic model of blackmail offer a coherent unifying framework for thinking about this issue.

Introduction

Most of the legal codes treat blackmail as a crime. Despite this fact, legal theorists have had a tough

time explaining why this is the case. In blackmail a perpetrator first threatens to reveal a piece of information about a potential victim or perform a similar act perceived as undesirable by the latter. Next he offers the victim a promise of silence in exchange for a monetary compensation. Both the threat and the offer of a voluntary exchange transaction are perfectly legal actions if taken on their own. Yet combined together they constitute a crime. This paradox of blackmail has attracted a lot of scholarly attention not only because of the perceived logical contradiction in which two rights combined make a wrong but also because of its deeper implications for the legal rules which deal with bargaining and regulation of information. In the words of Katz and Lindgren (1993, p.1565):

It has come to seem to us that one cannot think about coercion, contracts, consent, robbery, rape, unconstitutional conditions, nuclear deterrence, assumption of risk, the greater-includes-lesser arguments, plea bargains, settlements, sexual harassment, insider trading, bribery, domination, secrecy, privacy, law enforcement, utilitarianism and deontology without being tripped up repeatedly by the paradox of blackmail.

The literature on blackmail spans several disciplines, including legal studies, philosophy, and economics, and is therefore too broad to be comprehensively reviewed here. The focus of this entry will be on the theory of blackmail in the law and economics tradition which originates with the works by Ginsburg and Shechtman (1993) and Coase (1988). Not surprisingly, the unifying theme in this strand of the literature is the relationship between the blackmail and economic efficiency. In particular, this body of work has demonstrated convincingly that in general the blackmail transaction is likely to be wasteful and welfare reducing. This happens because the blackmailer faces a set of perverse incentives leading to an equilibrium in which social costs exceed social benefits. It is the likelihood of high social cost generated by the parties involved in the transaction that justifies the criminalization of blackmail.

Despite its success in formulating a coherent approach to resolving the paradox, the economic theory of blackmail has a few gaps which seem to

limit its universality. First, due to the sheer complexity of the phenomenon of blackmail, there are important types of blackmail transaction which do not seem to generate high social cost and thus fall outside the scope of the theory. Second, some commentators have argued that in addition to resolving the paradox of blackmail, a successful theory must also address the related “paradox of bribery”: if blackmail is a crime, why is it not a crime for someone to accept an offer of payment from a potential victim in exchange for the promise not to reveal the damaging information about them? This entry concludes by discussing these challenges to the economic theory of blackmail and possible ways to address them.

The Economic Theory of Blackmail

From an economist’s point of view, the problem of blackmail is closely related to the problem of externalities and bargaining. Indeed, in his 1987 McCorkle Lecture (Coase 1988), Ronald Coase states that the issue of blackmail first came to his attention in the process of writing “The Problem of Social Costs” (Coase 1960). For example, if a party responsible for the creation of a negative externality is not liable for the damage it creates, would it have an incentive to threaten to increase its output beyond the profit maximizing level in order to extract more surplus from the party affected by the externality? Similarly, if the creator of the externality is liable for the damage, wouldn’t the other party have an incentive to exploit this fact by threatening to take an action which would increase the damage (and hence the compensation received) beyond the socially optimal level? Coase discusses both of these possibilities in “The Problem of Social Costs,” but does not comment on the desirability of regulating blackmail. As he explained later (Coase 1988):

My purpose in pointing this out was to show that actions which were undertaken solely for the purpose of being paid not to engage in them, actions which could be called blackmail, would arise whatever the rule of liability, or if you like, the system of rights. I did this not to initiate the discussion of blackmail, but rather to avoid having to do so. In, any case, had I wanted to discuss the subject of

blackmail, the regime of zero transaction cost would have provided a poor setting in which to do so.

The next time we can see blackmail mentioned in the economic context is in the article by Landes and Posner (1975). In that paper the authors offer a general discussion of the private enforcement of the law and argue that the private enforcement of law by the means of blackmail is likely to be inefficient and therefore blackmail should be deemed illegal. The first full treatment of the problem of blackmail emphasizing the role of transaction costs was provided in 1979 by Douglas Ginsburg and Paul Shechtman in their manuscript “Blackmail: An Economic Analysis of the Law” which was later published as Ginsburg and Shechtman (1993). This paper was first to offer a formal argument demonstrating why legalizing blackmail would be detrimental to social welfare. In particular, consider a blackmailer (B) who threatens to collect and then disclose potentially damaging information about a victim (V). Suppose that B’s cost of collecting information is \$200 and that B expects that V would be willing to pay up to \$300 to prevent the disclosure. If blackmail were legal, B would proceed with his plan and earn a profit of \$100. This transaction, however, creates a net social loss because resources with positive opportunity costs are used for the purpose of redistributing existing wealth from one individual to another. Consequently, it makes sense to prevent such transactions from taking place, as “no rational economic planner would tolerate the existence of an industry dedicated to digging up dirt, at a real resource cost, and then reburying it” (Ginsburg and Shechtman 1993).

The preceding analysis assumes that information collected by B has no value to other people, besides being damaging to V. However, the conclusion does not change if we assume that this information has a positive market value, as, for example, in the case of celebrity gossip. If the market price of the information is \$200, then efficiency dictates that only information-gathering projects which cost less than \$200 should be undertaken. Allowing B to threaten the victim, who is willing to pay more than the market price, would violate this condition and

encourage investment in projects with negative social value. Note that if the gathering of information costs less than the market price (\$200), it will be undertaken even if blackmail is illegal. Thus criminalizing blackmail is socially efficient.

While the economic logic behind the criminalization of blackmail is rather straightforward, the above discussion made a number of implicit assumptions which should be critically examined. For example, how does the form of the available contractual arrangement (e.g., legally binding contracts) between the parties affect the equilibrium? Gomez and Ganuza (2002) construct a formal game theoretic model of bargaining under perfect information to address this question. They consider three regulatory regimes: criminal blackmail, legal blackmail without enforceable contracts, and legal blackmail when binding contracts are available. Their results essentially support the intuition developed by Ginsburg and Shechtman, allowing the authors to conclude that criminalization of blackmail improves efficiency by correcting the blackmailer’s incentives to collect and reveal information. As an extension of the model, Gomez and Ganuza (2002) also consider the case in which information revelation is socially valuable and find that the case for the criminalization of blackmail becomes even stronger under these circumstances.

Some Critiques of the Theory

The theory of blackmail discussed above offers a simple and intuitive explanation of the legal treatment of blackmail by appealing to the notion of economic efficiency. Nevertheless, it has faced a number of challenges, some of which deserve a closer look. The first important critique concerns the scope of the theory and its applicability to each possible type of blackmail. It is well understood that blackmail can come in many guises. Hepworth (1975) distinguishes four types of blackmail: *participant* blackmail, which occurs as a result of prior relationship between the parties; *opportunistic* blackmail, arising when information is acquired accidentally by the blackmailer; *commercial research* blackmail, based on the

information which was acquired on purpose; and *entrepreneurial* blackmail, in which the blackmailer himself creates a set of circumstances which implicate the victim. Lindgren (1989) uses this typology to argue that the economic theory of blackmail is incomplete because it cannot be used to justify criminalization of the participant and opportunistic types of blackmail. Indeed, if no resources were used to acquire information, then the efficiency considerations seem to lose their relevance and criminalization of blackmail remains a puzzle. One possible counter-argument to this critique was proposed by Coase (1988), who asserts that the resource costs are never equal to zero because the process of negotiating and securing the blackmail transaction requires expense of real resources. It is however difficult to argue that these transaction costs in the blackmail negotiation are conceptually different from the similar costs arising in the process of negotiating any voluntary exchange contract. Ginsburg and Shechtman (1993) go a step further asserting that even if information is acquired accidentally, the blackmailer would be willing to incur the cost of disseminating the information in order to establish his credibility should a new blackmail opportunity arise in the future. In their view it is this willingness of the blackmailer to invest resources in establishing his reputation (wasteful from the social planner's point of view) that gives rise to efficiency losses in the case of participant or opportunistic blackmail.

Another criticism commonly leveled against the economic theory of blackmail is the fact that it is not able to address the paradox of bribery, which arises when one considers the situation in which a potential victim offers a payment to the informed party in order to ensure that information is not released. As pointed out by DeLong (1993):

it is not unlawful for one who knows another's secret to *accept* an offer of payment made by an unthreatened victim in return for a potential blackmailer's promise not to disclose the secret. What would otherwise be an unlawful blackmail exchange is a lawful sale of secrecy if it takes the form of a "bribe". Lawful bribery poses an obvious challenge to theories that are premised on either the wrongfulness or wastefulness of the blackmail exchange...

Thus it seems that even though bribery and blackmail transactions aim to achieve the same final outcome, the identity of the party making an offer is crucial when deciding whether this transaction constitutes a crime. Interestingly, the legality of bribery can potentially create the same perverse incentives to "dig up dirt" on someone as in the case of legal blackmail. The only difference is that the party possessing the information is not allowed to make an explicit threat but would have to find a way of letting the other party know about the existence of the information and wait for the victim to come up with the offer instead.

While the existing literature on blackmail has not paid much attention to the paradox of bribery, it is not difficult to imagine a set of circumstances in which both the criminal blackmail and legal bribery can be defended on the grounds of economic efficiency. For simplicity, consider the case of accidental blackmail, and suppose that information has a positive social value. It is reasonable to assume that the damage suffered by the victim if information is revealed is known only to him, while the market value of information is common knowledge. Now consider the problem of a social planner who decides on the allocation of the bargaining power between the blackmailer and the victim. As in many similar bargaining settings under asymmetric information, to ensure efficient outcome, it suffices to give all the bargaining power to the informed party, i.e., the potential victim. If the victim's damage is higher than the social value of information (represented by the market price), his take it or leave it offer will ensure a mutually beneficial contract under which information is not revealed. If the damage is lower than the market price, no such deal can be reached and information will be revealed. In both cases a socially efficient outcome is obtained. On the other hand, if the bargaining power resides with the blackmailer, inefficiency can arise due to his lack of knowledge of the victim's reservation price (see Yerokhin (2011) for a detailed argument along these lines). The preceding analysis thus shows that efficiency considerations are able to justify both the criminalization of blackmail and the legality of bribery, although under rather restrictive assumptions that the information

acquisition is costless (e.g., participant or opportunistic blackmail) and that the market price of information reflects its true social value. In any case the situation in which information acquired by the blackmailer has a positive social value raises a set of interesting questions related to the public good nature of information and its implications for the regulation of blackmail. This issue remains largely understudied in the literature. One notable example is Miceli (2011), who offers an analysis along these lines and concludes that the regime of legal blackmail would lead to inefficient suppression of information due to the fact that the blackmailer is not able to extract the whole social surplus.

Conclusion

The economic theory of blackmail originating in the works of Ginsburg and Shechtman (1993) and Coase (1988) provides a simple and intuitive solution to the paradox of blackmail. It achieves this by considering the incentives which would arise in a regime of legal blackmail and their potential impact on economic efficiency. Treating the phenomenon of blackmail as a special case of bargaining about externalities with positive transaction costs, it argues that the blackmail transaction is likely to be welfare reducing and therefore should be penalized. While the basic arguments of the theory are well understood and accepted in the literature, there remain a few open questions as to whether the efficiency-based arguments can deal with all of the different types of blackmail and related phenomena such as bribery. This seems to be a fruitful area for the application of the game theoretic models of bargaining under asymmetric information, which could potentially strengthen the existing theory.

References

- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
 Coase R (1988) The 1987 McCorkle lecture: blackmail. *Univ Virginia Law Rev* 74:655–676

- DeLong S (1993) Blackmailers, bribe takers and the second paradox. *Univ Pennsylvania Law Rev* 141:1663–1693
 Ginsburg DH, Shechtman P (1993) Blackmail: an economic analysis of the law. *Univ Pennsylvania Law Rev* 141:1849–1876
 Gomez F, Ganuza J-J (2002) Civil and criminal sanctions against blackmail: an economic analysis. *Int Rev Law Econ* 21:475–498
 Hepworth M (1975) Blackmail: publicity and secrecy in everyday life. Routledge and Kegan Paul, Boston, pp 73–77
 Katz L, Lindgren J (1993) Blackmail symposium: instead of preface. *Univ Pennsylvania Law Rev* 141:1565
 Landes W, Posner R (1975) The private enforcement of law, *J Legal Studies*, 1
 Lindgren J (1989) Blackmail: on waste, morals and Ronald Coase. *UCLA Law Rev* 36:601–602
 Miceli T (2011) The real puzzle of blackmail: an informational approach. *Inf Econ Policy* 23(2):182–188
 Yerokhin O (2011) The social cost of blackmail. *Rev Law Econ* 7(1):337–351

Blockholders

- [Concentrated Ownership](#)

Bloomington School

Paul Dragos Aligica
 Mercatus Center, George Mason University,
 Arlington, VA, USA

Definition

The Bloomington School of Public Choice and Institutional Theory created by the 2009 Nobel Prize in Economics corecipient, Elinor Ostrom, and movement co-founder, Vincent Ostrom, have contributed to the theory of collective action, the analysis of metropolitan administration, to governance theory, the theory of goods and their production, to the empirical study of natural resources management and of the commons, to the nature and role of social norms, and to the problems of constitutional political economy and institutional design. It also advanced significant

methodological contributions to the study of polycentric, multilevel, dynamic institutional arrangements. In addition to the empirical, theoretical, and methodological research, the Bloomington school has been associated to a broader philosophy of social order and change centered on the “human condition” and the artefactual reality created by fallible but capable human beings in their process of evolution. This facet of the Bloomington school, especially developed in the writings of Vincent Ostrom, combines the distrust and skepticism regarding the “seeing like a state” stance in social and policy analysis, with the advocacy of a “seeing like a citizen” attitude grounded in a deep commitment to democracy and self-governance as normative ideals.

Bloomington School

The notion that there is a particular approach to Public Choice and Institutional Analysis specific to a group of scholars associated to Indiana University Bloomington was first articulated as an intellectual history point by William C. Mitchell in “Virginia, Rochester, and Bloomington: Twenty-five years of public choice and political science,” published in *Public Choice* in 1988: “Aside from the family analogy, it seems that three schools of thought have appeared in public choice and that they are sufficiently different to warrant distinctive labels. Mine are taken from their geographical locations: Virginia (Charlottesville; Blacksburg; Fairfax), Rochester, and Bloomington. At each of these institutions one or two dominant figures led and continue to lead in the effort to construct theories of collective choice: Riker at Rochester, Buchanan and Tullock at various Virginia universities, and the Ostroms at Indiana.”

The Bloomington school created by the 2009 Nobel Prize in Economics corecipient Elinor Ostrom and movement cofounder, Vincent Ostrom, has twofold roots. First in the Public Administration literature and debates. While engaged in the study of metropolitan governance – and especially the heated debates of the 1960s regarding the problems of metropolitan reform via consolidation and centralization –, Vincent

and Elinor Ostrom realized the limits of the mainstream theory of public administration and its policy implications, as expounded and applied to that specific set of problems. Hence, they turned their attention to the cutting edge developments in social sciences, set at that time into motion by James Buchanan and Gordon Tullock and their network of scholars who were advancing the study of “nonmarket decision making,” while they were creating the “no name field” which was later to become the field of Public Choice. The Ostroms turned into key members of the new field, each occupying at one point the position of the President of the Public Choice Society. Out of the combination of classical Public Administration theory with the emerging Public Choice theory and its associated Constitutional Political Economy ramifications grew what is now known as the “Bloomington School” of Public Choice or Institutionalism (Aligica and Boettke 2009).

The Ostroms started their work with an effort to rebuild a Public Choice-based theory of public administration which questioned the assumption, deeply ingrained in the mainstream outlook, according to which large bureaucracies were more efficient than the systems based on competition or bargaining in solving collective action problems and in providing public goods and services. That initial effort had two facets. First, the Ostroms and their associates organized around what was to become the Bloomington “Workshop in Political Theory and Policy Analysis,” started to develop the conceptual and the analytical tools that could explain the structure, functioning, and performance of governance arrangements in conditions of complex, multilevel, institutional diversity. Second, they noted that to advance the understanding of such governance phenomena it was not sufficient to simply note the differences between the two approaches to metropolitan governance (mainstream public administration vs. the public choice) and to suggest – based on the formal features of the theories used to describe them – that one is better than the other. Hence they put together an empirical research program aiming at investigating comparatively the applied level propositions derived from the two paradigms. As they advanced on the theoretical and

empirical paths, the methodological dimension grew naturally. It represents a third facet of their distinctive school, as they had to develop a series of instruments and approaches calibrated to the specific problems and phenomena they had to engage with, as well as to the collaborative, team-based type research that they favored.

Theories of Institutions and Governance

On the theoretical side, the starting point of the Bloomington school was the development of a particular form of institutionalism based on the public choice taxonomies of goods and services (framed by the twin dimensions of “exclusion” and “jointness of use or consumption”). Compared to other schools of institutional theory (each emphasizing as a distinctive feature of their approach, either transaction costs, or agent-principal relationships, or property rights, or rent seeking etc.), the Bloomington scholars – without denying the relevance of the other factors – moved to the forefront the nature of the goods and services, considered the main entry point into institutional analysis (Ostrom and Ostrom 2004). The diversity of institutional arrangements could be thus seen as driven by a functional structure shaped by the nature of goods or services that the institutions in case are supposed to deliver by solving collective action problems under circumstances characterized by diverse configurations of transaction costs, agent-principal relationships, property rights, rent-seeking incentives, etc. Given the contextual variety of factors, the emergence of institutional diversity is natural, and the study and theorization of institutional diversity has become an inherent feature of the Bloomington perspective.

Another distinctive theoretical attribute of the Bloomington School has been the development, around the concept of “polycentrism,” of a theoretical framework for the study of complex and dynamic governance systems, characterized by institutional diversity and social heterogeneity. Polycentricity, understood as a pattern of order defined by overlapping, competing, and cooperating centers of decision-making at different levels, all operating under the sway of an

overarching system of laws and norms, has been theorized in the Bloomington school tradition both as a positive (explanatory) and as a normative theoretical tool. The analysis and assessment of polycentric systems and governance forms entailed the development of an entire set of theories and models such as: Competitive governance (a process set in motion by polycentric structures); Coproduction (the situation in which the quality or even the very production of a good or service is dependent on the consumer’s participation in the production process); Public Entrepreneurship (conceptualized as a function emerging in solving collective action problems in competitive governance situations), Citizenship (a model of social actor having a specific profile and a set of capabilities operating as coproducer of governance in polycentric arrangements); Self-governance (as a normative ideal made possible by polycentric systems), etc. All of the above are pointing out not to a single, most efficient pattern of organization, but to a continual search for relatively more effective ways to perform, as well as to the institutional and social conditions that are shaping that search.

Empirical Research

On the empirical side, the Bloomington scholars have developed, starting at the end of the 1960s, a series of path-breaking research projects dedicated to the comparative study of metropolitan governance in the US, featuring detailed investigations of domains such as police services, water supply, and education and health services. They were also pioneers in the study of public entrepreneurship, public economies, and the “neither markets nor states” domain of institutional diversity emerging at the interface between the private and the public spheres. All these lines of empirical research assumed and implied an ongoing investigation of the production, financing, distribution of goods and services, and of the various problems of collective action situations associated to them. Consequently, Bloomington was increasingly recognized as a major center for collective action research. From all these, grew in the 1990s the program exploring the issue of the commons and

the problem of governing the commons, for which the Bloomington school is best known to the larger public today (Ostrom 1990). That has further reinforced and lead to new inquiries on topics such as: fisheries, forests, irrigation systems, pastures, natural resources management, environmental governance, foreign aid, and knowledge management and governance, making out of Bloomington one of the leading public choice and institutionalist theory empirical research centers in the world.

Methods and Approaches

The challenges created by their empirical and theoretical agenda have led the Ostroms and the scholars associated with them to craft a series of methodological tools and to cultivate an acute awareness of the epistemological problems posed by social science methodology. The metropolitan governance debates motivated them to imagine new ways of measuring and evaluating the delivery and quality of public goods and services, together with the citizens' levels of satisfaction in their consumption. The investigations on common pool resources have familiarized them with the case study research and fieldwork in diverse social and cultural (including non-Western) settings. The study of polycentric institutional arrangements has confronted them with the methodological complexities intrinsic to the multiple levels of analysis (Ostrom 2005).

In their effort to reflect upon, articulate, and codify their research practice, the Bloomington School scholars have come to distinguish between frameworks, theories, and models. A framework identifies the elements and relationships among the elements that need to be considered for analysis – a guideline pointing to the variables to be potentially considered in all types of institutional arrangements. Theories focus on specific elements of the framework, making particular assumptions and hypotheses about those elements and the relationships between them, while models make even more precise specifications about a limited set of parameters and variables, given a theory or a set of theories within a broader

framework. In this respect, a special contribution of the Ostroms has been the development of a framework for the study of institutional arrangements and collective action processes, the IAD (Institutional Analysis and Development) framework: a multitier conceptual map for the identification of action arenas, the resulting patterns of interactions and the ensuing outcomes, as well as for evaluating those outcomes. Last but not least, the Bloomington school has both practiced and theorized the multiple and mixed-methods approach, a “working together” philosophy of social research, based on teamwork and the division of intellectual labor aiming to capture and channel in systematic ways the researchers' complementarities of skills and specialization.

A Philosophy of Self-Governance

In addition to all of the above, the empirical, theoretical, and methodological research of the Bloomington school has always been associated to a broader vision, which advances a philosophy of social order and change centered on the “human condition” and the artifactual reality created by fallible but capable human beings in their process of evolution. This facet of the Bloomington school, especially developed in the writings of Vincent Ostrom, combines the distrust and skepticism regarding the “seeing like a state” stance in social and policy analysis, with the advocacy of a “seeing like a citizen” attitude grounded in a deep commitment to democracy and self-governance as normative ideals. The contribution of the Bloomington School – insist repeatedly the Ostroms – should be seen as part of an intellectual tradition which included *The Federalist* and the Tocquevillian perspectives, as well as the revolt against the Wilsonian Administrative State, the institutional embodiment of the high modernist “seeing like a State” stance in political and governance affairs.

If that is the case, then the crucial question, explains Vincent Ostrom, is the following: “If you and I are to be self-governing, how are we to understand and take part in human affairs?” (Ostrom 1997, 117). To answer it, he wrote, we need to articulate a view “. . . in which a science

and art of association rather than a science of command and control is viewed as constitutive of democratic societies. This fundamental difference of perspective has radical paradigmatic implications in addressing the question ‘Who govern?’ in the plural rather than ‘Who governs,’ in the singular. A minor distinction in language may have radical implications for theoretical discourse in the same way that a shift in perspective from a revolving sun to a spinning and orbiting earth had profound implications for many different sciences, professions, and technologies” (Ostrom 1997, p. 282).

Cross-References

- [Constitutional Political Economy](#)
- [Governance](#)
- [Public Choice: The Virginia School](#)

References

- Aligica PD, Boettke PJ (2009) Challenging institutional analysis and development: the Bloomington school. Routledge, New York
- Ostrom E (1990) Governing the commons. The evolution of institutions for collective action. Cambridge University Press, Cambridge
- Ostrom V (1997) The meaning of democracy and the vulnerability of democracies: A response to toqueville’s challenge. University of Michigan Press, Ann Arbor
- Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
- Ostrom E, Ostrom V (2004) The quest for meaning in public choice. *Am J Econ Sociol* 63(1):105–147

Böhm, Franz

Stefan Kolev

Wilhelm Röpke Institute, Erfurt and University of Applied Sciences Zwickau, Zwickau, Germany

Abstract

The purpose of this entry is to delineate the political economy and legal philosophy of

Franz Böhm. To reach this goal, a history of economics approach is harnessed. First, the entry concisely reconstructs Böhm’s life, intellectual evolution, and public impact. Second, it presents the specificities of his theories of market power, of competition as a disempowerment instrument, and of private law society.

Biography

Franz Böhm (1895–1977) was a German legal scholar who co-initiated the Freiburg School of ordoliberalism, but also a key political figure during the postwar decades of the Federal Republic in asserting the politico-economic agenda of the Social Market Economy in general and of antitrust legislation in particular. This introduction aims at embedding Böhm in his time and at depicting his role in several contexts in science as well as at the interface between science and society, before subsequently turning to his contributions to political economy and to legal philosophy.

Böhm was born in Konstanz and grew up in the capital of Baden, Karlsruhe, in the family of a high government official and later minister of education. After participation in the war, he studied law at Freiburg and, before finishing his dissertation, left for Berlin in 1925 to join the antitrust section of the Ministry of the Economy. In 1931, Böhm returned to Freiburg to finalize his dissertation and to subsequently start a habilitation project. The dissertation became the cornerstone of his habilitation, “Competition and the Struggle for Monopoly,” submitted in April 1933 and reviewed by the lawyer Hans Großmann-Doerth (1894–1944) and the economist Walter Eucken (1891–1950) (Eucken-Erdsiek 1975, pp. 12–14; Vanberg 2008, pp. 43–44). Both assessed Böhm’s piece as a success: Großmann-Doerth praised Böhm’s attempt to justify the seminal role of “performance-based competition” against the anticompetitive pressure groups within the industry, while Eucken applauded Böhm’s efforts to base his legal case on economic theory (Hansen 2009, pp. 46–48). With the almost immediate start of joint seminars, the three scholars established what later became known as the Freiburg School

of Ordoliberalism or as the Freiburg School of Law and Economics (Böhm 1957; Vanberg 1998, 2001; Goldschmidt and Wohlgemuth 2008a). Their cooperation steadily intensified, and the book series “Order of the Economy” initiated in 1936 constituted a milestone – its introduction under the title “Our Mission” became the programmatic manifesto of the incipient ordoliberal understanding of the role of law and economics in science and in society (Böhm et al. 2008; Goldschmidt and Wohlgemuth 2008b). Böhm received a temporary professorship at Jena in 1936, but in 1937 he was suspended from teaching after criticism of National Socialism (Vanberg 2008, pp. 43–44; Hansen 2009, pp. 88–128). Together with Eucken, he participated in the Freiburg Circles, intellectual resistance groups whose interdisciplinary discourse envisioned solutions for the age after National Socialism (Rieter and Schmolz 1993, pp. 95–103; Nicholls 1994, pp. 60–69; Grosseketler 2005, pp. 489–490).

Unlike Eucken, with whom in 1948 he co-founded the “ORDO Yearbook of Economic and Social Order,” Böhm was blessed with a long life, which enabled him to become an essential figure in the politico-economic and legal developments during the early decades of the Federal Republic. Böhm’s career at Freiburg in 1945 was a brief one: in the last months of the war, he received a position in the institute of the late Großmann-Doerth, became vice-rector of the university, but already in October 1945 he left to Frankfurt, the place to become focal for his further development. After a short period as advisor to the American authorities on decartelization and as minister of education, in early 1946 he received a call to the chair of private law, trade and business law at the University of Frankfurt which he would hold until 1962 (Zieschang 2003, pp. 227–228; Vanberg 2008, p. 44). Together with high administrative positions at the university, Böhm was active in Ludwig Erhard’s Frankfurt-based economic administration and simultaneously worked on new proposals for antitrust legislation (Möschel 1992, pp. 62–65; Hansen 2009, pp. 264–272; Glossner 2010, pp. 104–105). From 1953 to 1965, he was member of the

Bundestag, dedicating the first years especially to the protracted and tiresome debates with the Federation of German Industry (BDI) on various proposals for antitrust legislation – eventually passed and coming into effect in 1958 as the “Act against Restraints of Competition” (GWB), a fundamental document for the further development of European competition policy (Giocoli 2009). While Böhm’s focus in politics was directed at antitrust law, parallel efforts regarding labor relations and inner-company co-determination are also noteworthy (Biedenkopf 1980). Acting as Chancellor Adenauer’s envoy, Böhm was also a key figure in negotiating the first compensation agreements between the Federal Republic and Israel and remained important for the relations to Israel all his life (Hansen 2009, pp. 425–461).

Power in the Economy as an Enemy to Liberty

In 1928, during his years as an antitrust official in Berlin, Böhm formulated an article that would prove seminal for his further intellectual development. In “The Problem of Private Power: A Contribution to the Monopoly Debate,” he extensively discussed both theoretical and practical notions regarding the power which stems from cartels and monopolies and juxtaposed this type of power with the power and coercion which stem from government, also comparing the respective abilities of private and public law to deal with them (Böhm 2008). Böhm’s analysis of the legal practice of the preceding decades, following the fundamental decision of the Imperial Court of 1897 legalizing cartels and making them legally enforceable (Möschel 1989, pp. 143–145; Nörr 2000, pp. 148–156), led him to the diagnosis that the treatment of monopolies and cartels had been highly inadequate and that therapies to the ensuing problems of power concentration and “re-feudalization of society” (Tumlrir 1989, pp. 130–131) were overdue – and it was both the diagnosis and the therapy which he expanded upon in his habilitation and in his contribution to the “Order of the Economy” book series (Böhm 1933, 1937). The core problem he was struggling

with was to what extent the “rules of the game” of private law should be indifferent to (or even affirmative of) power concentrations as visible in monopolies and cartels or whether special attention was to be invested in designing rules which counteract such concentrations (Sally 1998, pp. 115–116).

Competition as a Disempowerment Instrument

Böhm’s key early contribution to the incipient political economy of the Freiburg School was the concept of the “economic constitution” (Tumlir 1989, pp. 135–137; Vanberg 2001, pp. 39–42, 2008, pp. 45–46). This concept not only fortified the interdisciplinary character of the scholarly community between lawyers and economists at Freiburg, but – with the semantic proximity between “constitution” and “order” – also provided a cornerstone for the development of the seminal analytical distinction of “economic order” versus “economic process,” a core element of the “Freiburg Imperative” (Rieter and Schmolz 1993, pp. 103–108) which was also at the root of many debates with laissez-faire liberals, most notably Ludwig von Mises (Kolev et al. 2014, pp. 6–7; Kolev 2016).

The search for “rules of the game” of the economic order, which adequately handle the problems of private power on markets, was further augmented by Böhm on another key domain: the ordoliberal notion of competition. Böhm contributed here at least in two respects: on the nature of competition and on the role of competition. On the nature of competition, his conceptual apparatus is based upon the notion of “performance-based competition” (*Leistungswettbewerb*), a procedural view on the desirable competitive process aimed at superior performance for the customers – which can be seen as a counterweight to “complete competition” (*vollständiger Wettbewerb*), the end-state view of competition (close to the neoclassical understanding of perfect competition) also present in ordoliberalism (Vanberg 2001, pp. 46–47; Kolev 2013, pp. 63–65; Wohlgenuth 2013, p. 166). On the role of

competition, Böhm innovated with his notion of *Entmachtungsinstrument*, i.e., competition as “the greatest and most ingenious disempowerment instrument in history” (Böhm 1961, p. 21) – a concept which clarifies how opening the doors of markets to competition (and keeping these doors open) creates choice options for the opposite side of the market and thus destroys the detrimental impact of power concentration.

Private Law Society

Later in his career, Böhm presented what Chicago economist Henry Simons called in the 1930s a “positive program,” i.e., a vision of the desirable order as opposed to primarily negative phenomena like the issues of power. Böhm called his positive program “private law society” (Böhm 1966). This system very much resembled Hayek’s legal philosophy as presented in the same period of the 1960s and 1970s, with Hayek referring to Böhm’s notion in *Law, Legislation and Liberty* (Hayek 1976, p. 158). Böhm’s ideal consists in creating protected domains especially of economic liberty for the individual, including the protection of property rights and enabling private contracts as cooperation of equals – domains which are to be secured by general rules of private (synonymously: civil) law (Streit and Wohlgenuth 2000, pp. 226–227; Sally 1998, pp. 115–117).

Society in Böhm’s analysis is an intermediate entity between the individual and the state, but it is distinctly separate from the state, thus opposing Carl Schmitt’s stance regarding the obsolescence of the distinction between society and state (Tumlir 1989, pp. 131–132). Society is an indispensable entity: it contains the market as one of its systems, as well as the sets of rules which enable cooperation between the individuals, but also their embeddedness in and subordination to the “rules of the game” (Nörr 2000, pp. 158–160). For Böhm, private law is a historical achievement of paramount importance for a free society to overcome the privilege-based order of feudalism and to establish equality before the law – which is also the reason why he chose this name for his ideal,

since he saw the general character of private law rules as the key obstacle to the abovementioned “re-feudalization” of society of his day, i.e., the permanent struggle for power and the successful regaining of privileges (understood as the very opposite of general rules) for individuals or groups in the sense of rent-seeking, in his analysis the greatest threat to the order of a free society of equals (Zieschang 2003, pp. 107–117).

Böhm’s impact and heritage go beyond the realm of ordoliberalism and the Social Market Economy. In addition, he was highly successful in being formative for generations of younger legal scholars, as becomes evident from the contributions in the numerous Festschriften and edited volumes dedicated to him (Mestmäcker 1960; Coing et al. 1965; Sauermann and Mestmäcker 1975; Konrad-Adenauer-Stiftung 1980; Ludwig-Erhard-Stiftung 1995) as well as from a special issue of *The European Journal of Law and Economics* in 1996 (Backhaus and Stephen 1996).

Cross-References

- Austrian School of Economics
- Constitutional Political Economy
- Eucken, Walter
- European Union Anti-cartel Policy
- Hayek, Friedrich August von
- Mises, Ludwig von
- Ordoliberalism
- Political Economy
- Röpke, Wilhelm

References

- Backhaus JG, Stephen FH (eds) (1996) Franz Böhm (1895–1977): pioneer in law and economics. Spec Issue Eur J Law Econ 3/4
- Biedenkopf K (1980) Der Politiker Franz Böhm. In: Konrad-Adenauer-Stiftung (ed) Franz Böhm: Beiträge zu Leben und Wirken, Forschungsbericht 8. Archiv für Christlich-Demokratische Politik, Bonn, pp 53–62
- Böhm F (1933) Wettbewerb und Monopolkampf: eine Untersuchung zur Frage des wirtschaftlichen Kampfrechts und zur Frage der rechtlichen Struktur der geltenden Wirtschaftsordnung. Carl Heymanns, Berlin
- Böhm F (1937) Die Ordnung der Wirtschaft als geschichtliche Aufgabe und rechtsschöpferische Leistung. Kohlhammer, Stuttgart
- Böhm F (1957) Die Forschungs- und Lehrgemeinschaft zwischen Juristen und Volkswirten an der Universität Freiburg in den dreißiger und vierziger Jahren des 20. Jahrhunderts. In: Wolff HJ (ed) Aus der Geschichte der Rechts- und Staatswissenschaften zu Freiburg i.Br. Eberhard Albert, Freiburg, pp 95–113
- Böhm F (1961) Demokratie und ökonomische Macht. In: Institut für ausländisches und internationales Wirtschaftsrecht (ed) Kartelle und Monopole im modernen Recht. Beiträge zum übernationalen und nationalen europäischen und amerikanischen Recht. C. F. Müller, Karlsruhe, pp 1–24
- Böhm F (1966) Privatrechtsgesellschaft und Marktwirtschaft. ORDO Jahrb Ordn Wirtsch Ges 17:75–151. English translation In: Peacock A, Willgerodt H (eds) (1989) Germany’s social market economy: origins and evolution. St. Martin’s Press, New York, pp 46–67
- Böhm F (2008) Das Problem der privaten Macht. Ein Beitrag zur Monopolfrage. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, pp 49–69
- Böhm F, Eucken E, Großmann-Doerth H (2008) Unsere Aufgabe. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, pp 27–37. English translation In: Peacock A, Willgerodt H (eds) (1989) Germany’s social market economy: origins and evolution. St. Martin’s Press, New York, pp 15–26
- Coing H, Kronstein H, Mestmäcker E-J (eds) (1965) Wirtschaftsordnung und Rechtsordnung: Festschrift zum 70. Geburtstag von Franz Böhm. C. F. Müller, Karlsruhe
- Eucken-Erdsiek E (1975) Franz Böhm in seinen Anfängen. In: Sauermann H, Mestmäcker E-J (eds) Wirtschaftsordnung und Staatsverfassung: Festschrift für Franz Böhm zum 80. Geburtstag. Mohr Siebeck, Tübingen, pp 9–14
- Giocoli N (2009) Competition versus property rights: American antitrust law, the Freiburg school, and the early years of European competition policy. J Compet Law Econ 5(4):747–786
- Glossner CL (2010) The making of the German post-war economy. Political communication and public reception of the social market economy after world war II. Tauris Academic Studies, London
- Goldschmidt N, Wohlgemuth M (2008a) Entstehung und Vermächtnis der Freiburger Tradition der Ordnungsökonomik. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, pp 1–16
- Goldschmidt N, Wohlgemuth M (2008b) Zur Einführung: Unsere Aufgabe (1936). In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger

- Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, pp 21–25
- Grossekettler H (2005) Franz Böhm (1895–1977). In: Backhaus JG (ed) *The Elgar companion to law and economics*. Edward Elgar, Cheltenham, pp 489–498
- Hansen N (2009) Franz Böhm mit Ricarda Huch: Zwei wahre Patrioten. Droste, Düsseldorf
- Hayek FA (1976) *Law, legislation and liberty*, vol 2: The mirage of social justice. University of Chicago Press, Chicago
- Kolev S (2013) *Neoliberale Staatsverständnisse im Vergleich*. Lucius & Lucius, Stuttgart
- Kolev S (2016) Ludwig von Mises and the “Ordo-interventionists” – more than just aggression and contempt? Working paper 2016–35. Center for the History of Political Economy at Duke University, Durham
- Kolev S, Goldschmidt N, Hesse J-O (2014) Walter Eucken's role in the early history of the Mont Pèlerin Society. Discussion paper 14/02. Walter Eucken Institut, Freiburg
- Konrad-Adenauer-Stiftung (ed) (1980) Franz Böhm: Beiträge zu Leben und Wirken, Forschungsbericht Nr. 8. Archiv für Christlich-Demokratische Politik, Bonn
- Ludwig-Erhard-Stiftung (ed) (1995) *Wirtschaftsordnung als Aufgabe: Zum 100. Geburtstag von Franz Böhm*. Sinus, Krefeld
- Mestmäcker E-J (ed) (1960) Franz Böhm: Reden und Schriften. C. F. Müller, Karlsruhe
- Möschel W (1989) Competition policy from an ordo point of view. In: Peacock A, Willgerodt H (eds) *German neo-liberals and the social market economy*. St. Martin's Press, New York, pp 142–159
- Möschel W (1992) Wettbewerbspolitik vor neuen Herausforderungen. In: Walter Eucken Institut (ed) *Ordnung in Freiheit: Symposium aus Anlass des 100. Jahrestages des Geburtstages von Walter Eucken*. Mohr Siebeck, Tübingen, pp 61–78
- Nicholls AJ (1994) *Freedom with responsibility. The social market economy in Germany, 1918–1963*. Clarendon, Oxford
- Nörr KW (2000) Franz Böhm and the theory of the private law society. In: Koslowski P (ed) *The theory of capitalism in the German economic tradition: historism, ordo-liberalism, critical theory, solidarism*. Springer, Berlin, pp 148–191
- Rieter H, Schmolz M (1993) The ideas of German ordoliberalism 1938–1945: pointing the way to a new economic order. *Eur J Hist Econ Thought* 1(1):87–114
- Sally R (1998) *Classical liberalism and international economic order: studies in theory and intellectual history*. Routledge, London
- Sauermann H, Mestmäcker E-J (eds) (1975) *Wirtschaftsordnung und Staatsverfassung: Festschrift für Franz Böhm zum 80. Geburtstag*. Mohr Siebeck, Tübingen
- Streit ME, Wohlgemuth W (2000) The market economy and the state: Hayekian and ordoliberal conceptions. In: Koslowski P (ed) *The theory of capitalism in the German economic tradition: historism, ordo-liberalism, critical theory, solidarism*. Springer, Berlin, pp 224–271
- Tumlir J (1989) Franz Böhm and the development of economic-constitutional analysis. In: Peacock A, Willgerodt H (eds) *German neo-liberals and the social market economy*. St. Martin's Press, New York, pp 125–141
- Vanberg VJ (1998) Freiburg school of law and economics. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*, vol 2. Palgrave Macmillan, London, pp 172–179
- Vanberg VJ (2001) The Freiburg school of law and economics: predecessor of constitutional economics. In: Vanberg VJ (ed) *The constitution of markets. Essays in political economy*. Routledge, London, pp 37–51
- Vanberg VJ (2008) Zur Einführung: Franz Böhm (1895–1977). In: Goldschmidt N, Wohlgemuth M (eds) *Grundtexte zur Freiburger Tradition der Ordnungsökonomik*. Mohr Siebeck, Tübingen, pp 43–48
- Wohlgemuth M (2013) The Freiburg school and the Hayekian challenge. *Rev Aust Econ* 26(3):149–170
- Zieschang T (2003) *Das Staatsbild Franz Böhms*. Lucius & Lucius, Stuttgart

Bondholder's Trustee

Flavio Bazzana

Department of Economics and Management,
University of Trento, Trento, TN, Italy

Abstract

The bondholder's trustee is an agent that will manage the bond contract between the issuer and the investors. The main reason for the existence of the bondholder's trustee is to reduce the general conflict of interests between debtholders and shareholders. Another important role of the bondholder's trustee is in the case of default, because reducing the expected costs of financial distress may decrease potential losses incurred by the bondholders. However, only recently the inadequacy of the current bondholder's trustee in different legal cultures has been criticised and, in the German case, also amended.

Definition

A bondholder's trustee is an agent – normally a financial institution – that has been empowered by

a bond issuer to act on behalf of the bondholders collectively.

The main duty of the bondholder's trustee is to manage the contract between the bond issuer and the investors that purchase the bond, in compliance with any governmental regulations that may apply. That is, the trustee may collect the initial payments from the investors and manage the interest and capital payments made by the issuer in accordance with the terms and conditions of the bond issue. The trustee will also periodically inform the bondholders about the economic and financial conditions of the issuer. Another important function of the trustee is the management in case of the default of the issuer, both for a failure of the bond payments – the capital and/or the interest – and when the issuer fails to respect particular terms of the contracts, the covenants. Covenants are specific clauses in debt contracts that regulate business policy and, when violated, provide debtholders the possibility of putting precise actions into force. Normally, these actions are the renegotiation of the conditions of the debt, i.e., the costs or the maturity, or the anticipated refund of the debt.

The main reason for the existence of both the bondholder's trustee and covenants is to reduce the agency costs of public corporate debt borne from the general conflict of interests between debtholders and shareholders (Jensen and Meckling 1976; Viswanath and Eastman 2003; Bazzana and Palmieri 2012). In fact, especially given a higher value of debt, the managers of a firm may make some decisions in favor of the shareholders that damage the debtholders, reducing the firm's value. Such actions result in wealth transfer from the debtholders in favor of the shareholders, which may also be partially controlled by a bondholder's trustee. Myers (1977) and Smith (1993) identify four main sources of conflicts: (i) dividend payments, (ii) claim dilution, (iii) asset substitution, and (iv) underinvestment (Bratton 2006). In the first case, when a firm is in financial turmoil, shareholders do not have incentives to invest and tend to withdraw liquidity from the firm, selling assets and consequently increasing dividend distribution. Claim dilution concerns the negative impact of a new debt on the value of the old debt. The cost of debt depends on the risk

of the firm, so the spread in every debt issue only incorporates the firm's current risk. The spread is normally fixed during the maturity of the debt, whether it be bonds or loans. Therefore, if a firm with a new issue increases the level of debt and, consequently, the level of risk, the spread of the old debt does not incorporate this new risk, leading to a decrease in the value of the old bond. The case of asset substitution depends on the shareholders' limited liability. When a firm is in financial distress, the shareholders may lose the net capital only and thus may decide to sell the firm's assets and substitute them with a riskier project. The total risk of the firm increases, but due to the asymmetry of its distribution, only the debtholders suffer from its increase, leading to a transfer of wealth. Finally, the case of underinvestment occurs when the shareholders decide not to undertake profitable investment projects. This event may occur because when the firm is in financial distress, the debtholders will obtain most of the benefits of the profitable project.

The agency costs of other types of debts, private debt (bank loans) and foreign public debt, are reduced without the use of the trustee and, in the latter case, without the use of covenants as well. In the case of private debt, the bank may collect more information and may process it in a more efficient way with respect to the case of public debt, reducing agency costs (Diamond 1984). The bank may also ask the firm to provide a guarantee in the form of collateral (Berger and Udell 1990). In addition to such methods, the bank may add some covenants to the loan contract to reduce the agency costs of debt (Demiroglu and James 2010). In the case of sovereign public debt, also in response to problems arising from the default of Argentine bonds – the unanimous approval of all bondholders on the change of the payment terms – the solution was the universal adoption of collective action clauses (CACs). With CACs only a qualified majority of bondholders is necessary to amend the payment terms that are binding also for nonparticipating bondholders (Haldane et al. 2005; Haeseler 2010).

However, it is in the case of default or in the case of technical default – the violation of a covenant – that the bondholder's trustee has the

most important role. Because the expected costs of financial distress have a negative impact on the value of the firm and, consequently, on the value of the bonds, reducing these costs may decrease potential losses incurred by the bondholders (Elkamhi et al. 2012). One of the costs that may be reduced by the bondholder's trustee is coordination costs, which depend strictly on the number of bondholders of every bond issue. In fact, if the bondholders are widely dispersed, they face serious difficulties in coordinating action, and because of the small claims, they have no incentive to solely bear the monetary and time costs to reduce dispersion. The bondholder's trustee, acting on behalf of the bondholders, may play a crucial role in reducing these costs (Bazzana and Palmieri 2012; Bazzana and Broccardo 2013).

However, only recently following the growing incidence of default on corporate bonds, the inadequacy of the current bondholder's trustee in different legal cultures has been criticized and, in the German case, also amended. According to Anglo-American norms – under the Trust Indenture Act in the United States – an “indenture trustee” (the bondholder's trustee) must be appointed for all public bond issues greater than 10 million dollars. Before a default, the responsibilities of the indenture trustee are very simple: distributing payments to investors, monitoring covenants, and other ministerial duties. If a default (including a technical default) occurs, the indenture trustee must act as a “prudent man” on behalf of the bondholders. This rule, especially during a default, is not simple to follow: for some actions such as changing the primary characteristic of the bond, e.g., its maturity, principal, or interest terms, unanimous bondholder consent is necessary. Also for these reasons, some solutions have been proposed to increase the power of the indenture trustee (Amihud et al. 2000), as have some alternatives that make similar adjustments (Schwarcz and Sergi 2008).

Also in Italian law, the capacity of the *rappresentante comune* (common representative) of the *assemblea degli obbligazionisti* (bondholder's meeting) – delineated in the Italian Civil Code – to manage the default situation efficiently is limited. In fact, before the default, the

information ability of this legal institution was very limited; for example, access to the company's financial documents was restricted, and after the default, the length of the procedure and the different quorum needed to resolve the bondholder's meeting may significantly postpone a formal decision. A possible solution to solve such inconveniences that follows the idea of Amihud et al. (2000) for Anglo-American norms and does not change the Italian Civil Code may be the use of the new hybrid contracts introduced in Italy with the 2003 corporate law reform (Bazzana and Palmieri 2012). It is possible to insert in these debt instruments mandatory representation to provide an investment firm the ability to act with full power on behalf of the bondholders, especially in default situations.

The German situation is different because German bondholder legislation was changed in 2009 with the introduction of the Debt Securities Act, which amended the Debt Securities Act of 1899. The representative structure is similar to those used in Italian law, featuring a bondholder meeting and the *Gemeinsamer Vertreter* (common representative), but with a greater power. In fact, under the German structure, the bondholder meeting may reach a binding majority decision regarding the bond's primary characteristics, such as changing the maturity or reducing the principal, with only a qualified 75% majority. The new law also facilitates the exclusion of individual bondholders' actions.

Cross-References

- Banks
- European Integration
- Law and Finance

References

- Amihud Y, Garbade K, Kahan M (2000) An institutional innovation to reduce the agency costs of public corporate bonds. *J Appl Corp Financ* 13:114–121
- Bazzana F, Broccardo E (2013) The role of bondholder coordination in freeze-out exchange offers. *J Financ Manag Market Institut* 1:67–84

- Bazzana F, Palmieri M (2012) How to increase the efficiency of bond covenants: a proposal for the Italian corporate market. *Eur J Law Econ* 34:327–346
- Berger A, Udell G (1990) Collateral, loan quality, and bank risk. *J Monet Econ* 25:21–42
- Bratton W (2006) Bond covenants and creditor protection: economics and law, theory and practice, substance and process. *Eur Bus Org Law Rev (EBOR)* 1:39–87
- Demiroglu C, James C (2010) The information content of bank loan covenants. *Rev Financ Stud* 23:3700–3737
- Diamond D (1984) Financial intermediation and delegated monitoring. *Rev Econ Stud* 51:393–414
- Elkamhi R, Ericsson J, Parsons CA (2012) The cost and timing of financial distress. *J Financ Econ* 105:62–81
- Haeseler S (2010) Trustee versus fiscal agents and default risk in international sovereign bonds. *Eur J Law Econ* 34:425–448
- Haldane AG, Penalver A, Saporta V, Shin HS (2005) Analytics of sovereign debt restructuring. *J Int Econ* 65:315–333
- Jensen MC, Meckling WH (1976) Theory of the firm: Managerial behavior, agency costs and ownership structure. *J Financ Econ* 3:305–360
- Myers SC (1977) Determinants of corporate borrowing. *J Financ Econ* 5:147–175
- Schwarcz SL, Sergi GM (2008) Bond defaults and the dilemma of the indenture trustee. *Ala Law Rev* 59:1037–1073
- Smith CW Jr (1993) A perspective on accounting-based debt covenant violations. *Account Rev* 68:289–303
- Viswanath P, Eastman W (2003) Bondholder-stockholder conflict: contractual covenants vs court-mediated ex-post settling-up. *Rev Quant Financ Account* 21:157–177

Bosman Ruling

Miriam Marcén

Universidad de Zaragoza, Zaragoza, Spain

Abstract

This entry explains the Bosman case and the Bosman ruling. I focus on the judicial process itself and on the decision of the European Court of Justice. The application of the Bosman ruling and its initial obstacles are also described. Finally, I present the consequences and implications of that ruling.

The Bosman Ruling

The Beginning of the Bosman Case

The indefatigable fight of an hitherto-unknown Belgian midfield football player, Jean-Marc

Bosman, who spent the last years of his professional sporting life going from court to court in defense of his rights, achieved one of the most recognized decisions of the European Court of Justice (for a revision of the ECJ, see Vauvel 2014), known as the Bosman ruling. Jean-Marc Bosman was a 25-year-old professional player in the RFC Liège when his contract expired in June 1990. He tried to force a transfer to another club, but under the rules of the Belgian Football Association and the UEFA (Union of European Football Associations), even with a contract complete, clubs could still charge a fee for a player to move as compensation for training expenses. The Belgian club demanded a fee for its player that appeared to be far out of reach for the French club, the USL Dunkerque, who were interested in signing the Belgian player. The impossibility of agreement between the Belgian and the French clubs prevented J.M. Bosman from playing in France in August 1990. After Bosman's unsuccessful transfer, he took legal action.

That was the beginning of the Bosman case, which beat the system of modern football. The judicial process was long, and it did not end in the Belgian courts because the case involved an international, cross-border employment dispute. He brought his case to the ECJ in Luxembourg, against the Belgian Football Association, RFC Liège, and UEFA (Union Royale Belge des Societes de Football Association ASBL & Others vs. Jean-Marc Bosman, ECJ Case No.: C-415/93 (1995) ECR I-4921), citing the 1957 Treaty of Rome, which guaranteed the freedom of movement of workers across European Union (EU) countries (for a revision of the European Community Law, see Heine and Sting 2015). Five years of legal battle was finally settled on December 15, 1995, when the ECJ ruled in his favor. The ECJ extended the application of Article 48 (on the prohibition of country restrictions of the free movement of workers) to the private sector. The Bosman ruling established that a player shared every other worker's right to move to an employer in another EU country without impediment. The ruling of the ECJ involved the liberalization of the migration of professional sportsmen and sportswomen within the EU and

the abolition of transfer fees after the expiration of contracts (Antonioni and Cubbin 2000). The restrictions on the number of EU players any club could field were also deemed illegal under the Bosman ruling.

During the whole judicial process, UEFA resolutely opposed Bosman's arguments. UEFA's proposition was that sporting activities constitute an exception to the provisions of the Treaty of Rome, since they could not be considered as economic activities and therefore were not subject to the Treaty regarding freedom of movement. With respect to transfer fees, UEFA argued that there should be compensation for the training and development of players since, if it did not exist, it would not be worth investing in young players (Marcén 2016). The ECJ rejected UEFA's arguments since, as the Court explained, transfer costs after a contract expiration are not the only means to assure the training and development of young players. UEFA ended up accepting the Bosman ruling in March 1996.

The Application of the Bosman Ruling and its Consequences

The application of the Bosman ruling was not exempt from certain difficulties. EU national leagues and international competitions were in the midst of the 1995/1996 season in Europe. In order to avoid the alteration of the competitions, with the transfer window closed in many countries, there was an agreement, known as "the gentlemen's agreement," which involved completion of that season (1995/1996) with the rules under which it had begun, that is, applying "rule 3 + 2." This meant that only three non-national players could be fielded, plus two additional foreign players who had played professionally in the host country for a period of five uninterrupted years, including 3 years in junior teams. No club violated that agreement during that season. In the next season, 1996/1997, the free movement of EU national football players and the abolition of transfer fees after the expiration of contracts were applied without exception (Marcén 2016).

The Bosman ruling resulted in more freedom of contracts and fewer restrictions on the mobility of players (Feess and Mühlheusser 2002), and the economic implications have been examined by several researchers. In the post-Bosman period, an unsurprising increase in player migration in professional football has taken place (Frick 2009). EU players after the ruling were not considered foreigners, making signing them more attractive. In addition, Bosman's victory made nonnational EU players more attractive to be signed up, since the nationality quotas were no longer applicable to EU national players. These movements internationalized the squads, although the effect of the Bosman ruling appears to vary over time (Marcén 2016). In the immediate aftermath of the ECJ ruling, the clubs had some limitations on signing new players, since most players still held contracts. In the case of the English Premier League, it took around 2 years until the number of its non-British players rose significantly (Lowrey et al. 2002). Additionally, not all clubs had the ability to scout foreign players at the time of the Bosman ruling, which generated some doubts about the signing of players, which, once again, delayed the impact of the ruling.

After overcoming the initial obstacles, clubs could – and did – field teams without a single native player. The arrival of the new players had an impact on the quality of the football leagues (Ericson 2000; Flores et al. 2010). As suggested in Álvarez et al. (2011), foreign players produced an improvement in the skills of the local native players through contact with new techniques and practices, which could then further improve the quality of the national teams (Binder and Findlay 2012). Other researchers have pointed out that the Bosman ruling decreased the competitive balance (Kesenne 2007; Vrooman 2007). The increasing demand for international players because of the open access to EU national football players, but also because of the elimination of transfer fees after the expiration of contracts, which decreased the costs of signing up international players, forced clubs to pay higher wages (Kesenne 2007) for nonnative players and native players alike, to deter the best players from leaving. This was possible, partly, through the dramatic increase

in television revenues, but this would not remain the case forever. With the power in their hands, players saw their wages go through the roof. With the decrease in TV revenues, and without the transfer fees that smaller clubs received in the pre-Bosman period, only the larger and richer clubs could afford to pay the salaries of the superstar football players. The small clubs also lost the incentives to develop home-grown talent, since their best young players could leave for free at the end of their contracts. Then, other consequence of the Bosman ruling was that, within each national competition, the differences between the small and the big clubs greatly increased. Similarly, the decreasing competitive balance was also observed across EU countries between the “Big 5” leagues and the rest of the European national leagues.

Nonetheless, the dynamic analysis of the impact of the Bosman ruling on the participation of native players in their national league, presented by Marcén (2016), using Spanish data, reveals that the impact of the Bosman ruling was greatest three or four seasons after the Bosman ruling, in line with the argument that the clubs had some initial restrictions, but after seven or eight seasons, no effect can be observed. This can be explained by a readaptation of the clubs and players to the conditions of the post-Bosman labor market. As explained by Antonioni and Cubbin (2000), the bargaining power of the clubs could be, at least in part, recovered by maintaining relations with their player under what Dietl et al. (2008) call the “shadow of the transfer system,” by excluding the advent of contract expiration through extended contracts. With the Bosman ruling, clubs did not lose the power to decide who is fielded, and so the threat of no play in any match in the whole season (imposed when players do not accept a new contract before the expiration of their existing contract) was credible and led many players to accept the renewal of their contracts.

The impact of the Bosman ruling is everywhere in the modern-day sport. The principles applied in the Bosman case extended to other non-EU nationals, those originating from “third countries,” making them equal to European Union

players (Hendrickx 2005). There are several examples of sportsmen attempting to attain Bosman objectives, such as Kolpak, a Slovak handball player in the German handball league. In 2000, Slovakia was not a member of the European Union, and therefore the Bosman ruling did not apply to its citizens. However, Slovakia had an Association Agreement with the European Union. Based on that agreement, in 2003 the European Court of Justice ruled in favor of Kolpak (Case C-438/00, *Deutscher Handballbund v. Maros Kolpak*), giving equal rights for Slovak sportsmen residing and lawfully employed by a club established in a Member State and EU players (Marcén 2016). The Russian Simutenkov and the Turk Nihat, football players participating in the Spanish League, also obtained rights equal to EU players in 2005 and 2008, respectively (Hendrickx 2005; Penn 2006). Even after more than 20 years of the Bosman ruling, its shadow has not disappeared, since it is unclear how the pending exit of the United Kingdom from the European Union, known as “Brexit,” will affect the mobility of workers and the large number of foreign players participating in the English Premier League.

References

- Álvarez J, Forrest D, Sanz I, Tena JD (2011) Impact of importing foreign talent on performance levels of local co-workers. *Labour Econ* 18(3):287–296
- Antonioni P, Cubbin J (2000) The Bosman ruling and the emergence of a single market in soccer talent. *Eur J Law Econ* 9(2):157–173
- Binder JJ, Findlay M (2012) The effects of the Bosman ruling on national and Club teams in Europe. *J Sports Econ* 13:107–129
- Dietl HM, Franck E, Lang M (2008) Why football players may benefit from the “shadow of the transfer system”. *Eur J Law Econ* 26:129–151
- Ericson T (2000) The Bosman case effects of the abolition of the transfer fee. *J Sports Econ* 1(3):203–218
- Feess E, Mühlheusser G (2002) Economic consequences of transfer fee regulations in European football. *Eur J Law Econ* 13:221–237
- Flores R, Forrest D, Tena JD (2010) Impact on competitive balance from allowing foreign players in a sports league: evidence from European soccer. *Kyklos* 63(4):546–557
- Frick B (2009) Globalization and factor mobility: the impact of the “bosman-ruling” on player migration in professional soccer. *J Sports Econ* 10:88–106

- Heine K, Sting A (2015) European Community law. In: Encyclopedia of law and economics. Springer, New York
- Hendrickx F (2005) The Simutenkov case: Russian players are equal to European Union players, the international sports law journal (ISLJ) 2005/3
- Kesenne S (2007) The peculiar international economics of professional football in Europe. *Scott J Polit Econ* 54(3):388–399
- Lowrey S, Neatrou S, Williams J (2002) The Bosman ruling, football transfers and foreign footballers. Centre for the Sociology of Sport, Leicester
- Marcén M (2016) The Bosman ruling and the presence of native football players in their home league: the Spanish case. *Eur J Law Econ* 42(2):209–235
- Penn DW (2006) From Bosman to Simutenkov: the application of non-discrimination principles to non-EU nationals in European sports. *Suffolk Transl Law Rev* 30(1):203–231
- Vauvel R (2014) Court of justice of the European Union. In: Encyclopedia of law and economics. Springer, New York
- Vrooman J (2007) Theory of the beautiful game: the unification of European football. *Scott J Polit Econ* 54(3):314–354

agents faced with the costs of acquiring, absorbing and processing information. It also presents an overview of BR's current applications in various fields of economic activity.

Definition

Bounded rationality (BR) is the idea that when individuals make decisions, they are “bounded” or limited because of inadequate information, cognitive limitations inherent in the human mind, and time constraints. This type of rationality fittingly describes broad areas of social and economic action in which rational utility-maximizers faced with complex situations are required to make less-than-perfect choices (satisficing rather than optimizing). The term was coined by the US economist and Nobel laureate Herbert Simon who proposed it as a more realistic version of the “perfect rationality” assumed by the neoclassical model (Simon 1982).

Bounded Rationality

Aspasia Tsaoussi

School of Law, Aristotle University of
Thessaloniki, Thessaloniki, Greece

Abstract

Bounded rationality (BR) is the idea that when individuals make decisions, they are “bounded” or limited because of inadequate information, cognitive limitations inherent in the human mind and time constraints. This type of rationality describes broad areas of social and economic action, in which rational utility-maximizers faced with complex situations are required to make less-than-perfect choices (satisficing rather than optimizing). BR is one of the cornerstones of rational choice theory and it informs many scientific fields, spanning from mathematic and economic psychology to political economy and managerial economics. This entry discusses the general aspects of BR, focusing on how cognitive biases affect the decision-making of rational

General Aspects of Bounded Rationality

BR is one of the cornerstones of rational choice theory (Simon 1955; Tversky and Kahneman 1986; Heap et al. 1992). Although its origins can be traced back several decades (more specifically, to the early stages of game theory), the modeling of human behavior in terms of choice flourished into a challenging new area of interdisciplinary research with the work of pioneering cognitive psychologists and behavioral economists in the early 1980s (Tversky and Kahneman 1981; G  th 2000). Today BR informs and populates many scientific fields, spanning from mathematic and economic psychology to political economy and the law.

According to Simon's original thesis, rational actors function under three insurmountable obstacles, which limit their potential to reach “perfectly” rational decisions: (1) only incomplete and often unreliable information is available regarding possible alternatives and their consequences, (2) the human mind has a reduced capacity to evaluate and process available information,

and (3) in many situations, only a limited amount of time is available in which to make a decision. These three limitations constitute the core around which contemporary BR theories have been developed.

Imperfect information or information overload affects the decision-making of all agents who are faced with the costs of acquiring, absorbing, and processing information, e.g., consumers (Reis 2006), managers (Basel and Brühl 2013), judges (Tsaoussi and Zervogianni 2010), and spouses-to-be (Garcia-Retamero et al. 2010). The problem is aggravated when rational actors enter into contracts. Time limitations serve to enhance the impact of the previous two factors. The outcome in the social world is the increase in the frequency of systematic errors in judgment but also the intensity and gravity of errors. Hence, the scientific study of how humans make suboptimal decisions overlaps extensively with the study of why humans make mistakes. BR models have been proposed to study the behavior of NGOs, global investors, teachers, bankers, and even soccer players and goalkeepers.

The limitations of human cognition are known as “cognitive biases.” The basic premise is that because the human brain cannot process all information, it routinely makes mistakes and it resorts to heuristics in order to make decisions (Baron 2007). The current literature covers an impressive array of cognitive biases. Behavioral biases, decision-making biases, and belief biases represent a main strand of research, both in social psychology (Kahneman 2003) and behavioral economics. Examples include the anchoring effect (the tendency to rely too heavily, or “anchor,” on one trait or piece of information when making decisions, usually the first piece of information that we acquire on that subject), the bandwagon effect (the tendency to do or believe things because many other people do or believe the same), and the framing effect (drawing different conclusions from the same information, depending on how or by whom that information is presented). Memory biases form another broad category within the cognitive bias family. Finally, social biases have been shown to impact decision-making. Characteristic examples include in-group

bias (the tendency of individuals to give preferential treatment to those they perceive to be members of their own group) and the “just-world phenomenon” (the tendency for people to believe that the world is just and therefore people get what they deserve).

Heuristics are experience-based cognitive shortcuts which humans have taken throughout history when problem-solving, discovering, and innovating. They are most helpful in settings when information is too costly or difficult to obtain. The best-known example is the availability heuristic, a mental shortcut that relies on immediate examples which come to mind. Heuristics involve making judgments on the basis of simple, fast, and efficient rules-of-thumb. Criticism has been leveled at decision theorists for straying from Simon’s original thesis and emphasizing the inability of individuals to optimize. Instead of searching for theoretically optimal models, it has been shown that simple heuristics help agents reach wiser decisions (Gigerenzer and Reinhard 2001). Simple heuristics that use limited information have been successfully applied to environments from stock market investment to judging intentions of other organisms to choosing a mate (Todd 2007). Behavioral economists are equally optimistic, having faith in the power of heuristics to increase the effectiveness of decision-making. A popular idea is that choice architectures may be modified in light of agents’ BR and that decisions which affect health, wealth, and happiness may be “improved” (Thaler and Sunstein 2008).

Specific Applications of Bounded Rationality

The applications of BR in economics are as rich as they are diverse (Rubinstein 1998). Over the last few decades, BR research has spilled over into other scientific disciplines, covering the full spectrum from brain research and preventive medicine to the role of emotions, ecological economics, and education policy. A sizable portion of the existing literature is theoretical-normative in scope and nature, consisting of two types of published

articles: those which function as critical reviews of BR or the perfect rationality hypothesis or both, and those which introduce new theories aspiring to expand the borders of the existing paradigm. However, BR has attracted the attention of many economists who conduct mostly empirical work and inject the field with the dynamic energy of their methodologies. There is also a plethora of articles combining theory and empirical research, and many of them fall within the larger “game theory” umbrella.

Game theoretical models that test several parameters of BR in different settings abound in recent years (Stahl and Haruvy 2008). To name one, a pivotal issue in game theory is that of the “finite horizon paradoxes.” The most noticeable examples are the finitely repeated Prisoner’s Dilemma game (Rapoport and Dale 1967), Rosenthal’s centipede game (Rosenthal 1981), and Selten’s chain store paradox (Selten 1978). In these games the standard game theoretic solutions (the application of backward induction and subgame perfection) yield results that are counter-intuitive and inconsistent with traditional equilibrium analysis. The paradoxes are explained only by the players’ lack of full rationality (Rosenthal 1989). The very popular Prisoner’s Dilemma game and the less known centipede game both present a conflict between the players’ self-interest and mutual benefit. If it could be enforced, both players would prefer to cooperate throughout the game. However, a player’s self-interest or both players’ distrust can interfere and create a situation where both do worse than if they had blindly cooperated. Selten’s chain store paradox introduces a three-level theory of decision-making where decisions can be made on different levels of rationality.

Another thread of BR research deals with Stackelberg games, which are natural models for many important applications involving human interaction, such as oligopolistic markets and security domains. Here one player (the leader) commits to a strategy, and the follower makes her decision knowing the leader’s commitment. Existing algorithms for Stackelberg games efficiently find optimal solutions (leader strategy), but they assume that the follower plays optimally.

Nevertheless, in many instances, agents face human followers (adversaries) who may deviate from their expected optimal response because of the cognitive limitations in observing the leader strategy (Pita et al. 2010).

The problem of imperfect information has been addressed by law and economics scholars in a growing body of literature which has extensively explored asymmetry of information in contracts, in various markets, and in particular manifestations: adverse selection, moral hazard, and the principal-agent problem. The basic premise is that no buyers can accurately assess the value of a product through examination before sale is made, and all sellers can more accurately assess the value of a product prior to sale (Akerlof 1970). Applications in economic life are plentiful. To use an illustrative example, moral hazard is viewed as a root cause of the subprime mortgage crisis in the USA which, according to economic analysts, triggered the 2008 recession.

Another “classic” BR application is in financial markets. Misperceptions of risk have been identified in empirical work as one of the major causes of the global financial crisis and related phenomena such as the bank run and stock market crashes. Satisficing (the fact that satisfactory decisions are chosen over the best decisions) is widely applied in marketing. Also, BR models can help reduce uncertainty in supplier selection decisions (Riedl et al. 2013). BR models have been developed to investigate the origin and structure of loss aversion and optimism in marketplaces (Yao and Li 2013).

In recent decades, economic organization has been increasingly approached under the lens of BR models (Foss 2001). Theorists from institutional economics (Williamson 1985; Hart 1995) have looked at the role of BR in the formation of contracts, attributing (e.g., Gifford 1999) contractual incompleteness to the trade-off between devoting scarce attention to managing existing contracts and writing new ones. The key assumption of BR in organizational economics is that managers are fallible, so they commit errors of evaluation (Sah and Stiglitz 1985). More recent analyses focus on the effects of first-mover advantage on organizational size and structure, finding,

for example, that smaller and decentralized organizations are preferable since they are speedier in making decisions. Following Simon, who often used BR in context with the theory of the firm, later models focused on BR as limited attention that needs to be economized and tested for different organizational responses to it. BR has also been applied to observe market fluctuations.

Finally, the BR assumption has proven fruitful in managerial economics. Some approaches (e.g., Tisdell 1996) draw on game theory, transaction costs theory, and evolutionary economics to critically examine decisions made by individuals, taking into account learning possibilities, and decisions by groups and economic organizations, studying optimal communication within organizations. Others study managerial decision-making in international business under the prism of BR, focusing on the role of decision-makers in foreign direct investment and related strategies (Aharoni et al. 2011).

The increased attention devoted by economists to BR assumptions has allowed them to make more accurate predictions about human behavior. BR analyses have enriched mainstream economic thinking, suggesting improvements to existing regulatory and policy frameworks. The main BR hypothesis and its implementation in different areas of economics have added sophistication to the classic view of human rationality, infusing it with greater adaptive power to the rapid technological changes of the new information society. Furthermore, by inquiring into the logic guiding the action of economic agents in complex interactive systems such as markets and games, BR proposes normative insights to rational actors about how to make better decisions. Lastly, through the looking glass of BR, legislators are able to minimize the negative unintended consequences of lawmaking, judges can reduce judicial error, and lawyers can more efficiently represent the interests of their clients.

Cross-References

- [Behavioral Law and Economics](#)
- [Games](#)

- [Good Faith and Game Theory](#)
- [Rationality](#)

References

- Aharoni Y, Tihanyi L, Connelly BL (2011) Managerial decision-making in international business: a forty-five-year retrospective. *J World Bus* 46(2):135–142
- Akerlof GA (1970) The market for ‘lemons’: quality uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Baron J (2007) *Thinking and deciding*, 4th edn. Cambridge University Press, New York
- Basel JS, Brühl R (2013) Rationality and dual process models of reasoning in managerial cognition and decision making. *Eur Manag J* 31(6):745–754
- Foss NJ (2001) Bounded rationality in the economics of organization: present use and future possibilities. *J Manag Gov* 5(3–4):401–425
- Garcia-Retamero R, Takezawa M, Galesic M (2010) Simple mechanisms for gathering social information. *New Id Psychol* 28(1):49–63
- Gifford S (1999) Limited attention and the optimal incompleteness of contracts. *J Law Econ Organ* 15(2):468–486
- Gigerenzer G, Reinhard S (eds) (2001) *Bounded rationality: the adaptive toolbox*. MIT Press, Cambridge
- Güth W (2000) Boundedly rational decision emergence – a general perspective and some selective illustrations. *J Econ Psychol* 21(4):433–458
- Hart O (1995) *Firms, contracts, and financial structure*. Clarendon, Oxford
- Heap SH, Hollis M, Lyons B, Sugden R, Weale A (1992) *The theory of choice: a critical guide*. Blackwell, Oxford
- Kahneman D (2003) Maps of bounded rationality: psychology for behavioral economics. *Am Econ Rev* 93(5):1449–1475
- Pita J, Jain M, Tambe M, Ordóñez F, Krau S (2010) Robust solutions to Stackelberg games: addressing bounded rationality and limited observations in human cognition. *Artif Intell* 174(15):1142–1171
- Rapoport A, Dale PS (1967) The “end” and “start” effects in the iterated Prisoner’s dilemma. *J Confl Resolut* 11:354–462
- Reis R (2006) Inattentive consumers. *J Monet Econ* 53(8):1761–1800
- Riedl DF, Kaufmann L, Zimmermann C, Perols JL (2013) Reducing uncertainty in supplier selection decisions: antecedents and outcomes of procedural rationality. *J Oper Manag* 31(1–2):24–36
- Rosenthal RW (1981) Games of perfect information, predatory pricing, and the chain store paradox. *J Econ Theory* 25(1):92–100
- Rosenthal RW (1989) A bounded-rationality approach to the study of noncooperative games. *Int J Game Theory* 18:273–292

- Rubinstein A (1998) Modeling bounded rationality. MIT Press, Cambridge, MA
- Sah RK, Stiglitz JE (1985) Human fallibility and economic organization. *Am Econ Rev* 75(2):292–296
- Selten R (1978) The chain store paradox. *Theory Decis* 9(2):127–159
- Simon HA (1955) A behavioral model of rational choice. *Q J Econ* 69(1):99–118
- Simon HA (1982) Models of bounded rationality: behavioral economics and business organization, vol 2. MIT Press, Cambridge, MA
- Stahl DO, Haruvy E (2008) Level-n bounded rationality and dominated strategies in normal-form games. *J Econ Behav Organ* 66(2):226–232
- Thaler RH, Sunstein CR (2008) *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven
- Tisdell C (1996) Bounded rationality and economic evolution: a contribution to decision making, economics and management. Edward Elgar, Cheltenham
- Todd PM (2007) How much information do we need? *Eur J Oper Res* 177(3):1317–1332
- Tsaoussi A, Zervogianni E (2010) Judges as satisficers: a law and economics perspective on judicial liability. *Eur J Law Econ* 29(1):333–357
- Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458
- Tversky A, Kahneman D (1986) Rational choice and the framing of decisions. *J Bus* 59(4):251–278
- Williamson OE (1985) *The economic institutions of capitalism*. Free Press, New York
- Yao J, Li D (2013) Bounded rationality as a source of loss aversion and optimism: a study of psychological adaptation under incomplete information. *J Econ Dyn Control* 37(1):18–31

Broken Window Effect

Joël J. van der Weele¹, Mataka P. Flynn² and Rogier J. van der Wolk^{2,3}

¹Department of Economics, University of Amsterdam, Amsterdam, The Netherlands

²Faculty of Law, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

³Faculty of Law, Universiteit van Amsterdam, Amsterdam, The Netherlands

Abstract

The broken windows effect refers to the hypothesis that there is a positive effect of urban disorder on the incidence of more

serious crimes, where the term “broken windows” represents a range of disorders within communities. The hypothesis has been the subject of an intensive academic debate and has had an important effect on law enforcement in the USA, where it increased the focus on community policing and zero tolerance. This essay reviews the evidence for the existence of the broken windows effect and the effectiveness of the associated policing strategies.

Introduction

James Wilson and George Kelling coined the term “broken windows effect” (BWE) in a magazine article in 1982. They argued that broken windows, a catch-all term for disorder and incivility within a community like graffiti, vandalism, littering, etc., can lead to further disorder and increases in criminal behaviour. Their argument originated in experiences with Newark police practices and also referenced an “experiment” by Stanford psychologist Zimbardo (1973), who abandoned cars in different neighborhoods and in various states of disrepair to study their subsequent vandalization.

Kelling and Wilson (1982) asserted that broken windows send a signal of indifference and lack of enforcement, leading to increased fear of crime and weakening of social controls, thus paving the way for bigger transgressions. To prevent such processes, the authors argued that it is crucial for the police to engage in the prevention and policing of disorder and minor crimes like panhandling.

The BWE had a substantial influence on the practice of law enforcement in several major US cities, inducing a shift to disorder policing and “zero tolerance”. Specifically, in 1993, under the guidance of Police Commissioner William Bratton and Mayor Rudolph W. Giuliani, New York City (NYC) implemented an enforcement strategy designed to “reclaim the public spaces of New York” (Bratton and Knobler 1998, p. 228). This led to a more than 50% increase in arrests for misdemeanors, without a substantial increase in reporting of misdemeanors (Harcourt 1998).

These police initiatives coincided with a large drop in crime in NYC, sparking a vigorous debate involving both academics and practitioners (Eck and Maguire 2000). Proponents of the BWE argue that the initiatives like those in NYC described above were at least partly responsible for drop in crime (Kelling and Bratton 1998). Critics, however, emphasize rival explanations like the decline in the crack epidemic and argue that crime was decreasing in NYC prior to the implementation of novel policing strategies and decreased in other large cities without aggressive policing strategies (Harcourt 1998; Fagan et al. 1998).

Evidence

The existence of a BWE is an empirical question. In what follows, we will discuss some of the most influential contributions to the empirical debate and provide some thoughts about directions for future research.

To test the broken windows hypothesis, a number of studies have tried to establish a correlation between disorder, possibly as a consequence of variation in police strategy, and more severe crimes like robberies or violent crimes. Many studies focus on the New York 1993 crime initiative and use variations in misdemeanor arrests across precincts or boroughs as a measure of disorder. Different studies vary in the length of those time-series and the control variables used.

For instance, Kelling and Sousa (2001) use variations across NYC precincts over time with regard to both misdemeanor arrests and crime rates and demonstrate a statistically significant and substantial negative relationship between both. This result is echoed by Corman and Mocan (2005), who analyze a dataset of monthly time series citywide and control for policing variables such as felony arrests and the size of the police force.

Other studies have found smaller or no correlations between disorder and crime. For instance, Sampson and Raudenbush (1999) employed trained observers to drive through all streets

throughout 196 census tracts in Chicago. They selected 15,141 street sides, which were then videotaped and coded for neighborhood disorder. The authors found that the correlations between the documented disorder and most predatory crime rates disappeared when controlling for structural neighborhood conditions like poverty.

Apart from yielding conflicting results, the correlational studies cited above (and similar ones like it) are susceptible to confounding explanations. For example, Harcourt and Ludwig (2006) attempt to replicate the findings by Kelling and Sousa (2001) and Corman and Mocan (2005) and find that the patterns are also consistent with mean reversion. That is, high spikes in crime are followed by crime drops regardless of disorder policing. Alternatively, like Sampson and Raudenbush (1999) point out, both disorder and crime may depend on unobserved neighborhood characteristics, resulting in spurious correlations.

To overcome such problems, an increasing number of studies investigate the impact of experimental or random variation in disorder policing strategies. This makes it possible to exclude the confounds mentioned above and make causal statements about the effect of policing disorder on crime. For instance, Braga et al. (1999) conducted a randomized control trial, focussing on problem-oriented policing strategy to control physical and social disorder. The strategy resulted in a significant reduction in crime incidents in hotspots for violent crime, without much evidence for displacement to nearby locations.

Harcourt and Ludwig (2006) used data from an experiment carried out by the US Department of Housing and Urban Development known as Moving to Opportunity (MTO). Officials assigned around 4,600 low-income families from communities largely consisting of public housing and characterized by high rates of crime and social disorder to housing in more reputable and high-status neighborhoods (Orr et al. 2003). Assignment to the program was random, allowing identification of the impact of disorder in the new neighborhood on the relocated individuals' criminal activity, without confounding effects of

individuals' background. Harcourt and Ludwig (2006) compared the crime rates of those that moved and those that stayed in their lesser neighborhood and found that the rate of neighborhood order or disorder has no noticeable effect on criminal behavior.

Braga et al. (2015) provide a meta-analysis of 30 (quasi) experimental studies for a range of different disorder policing strategies. The authors conclude that aggressive zero-tolerance strategies against individuals do not produce significant effects on crime; however, approaches that involve the community to solve problems at particular locations tend to be quite effective. This is in line with the conclusions of an earlier review by Weisburd and Eck (2004). It also reinforces the findings of Taylor (2001), who used longitudinal data from 66 neighborhoods in Baltimore to determine the relationship between disorder and crime. On the basis of those results, Taylor argues that different types of disorder might require different policies and cautions against simple-minded crackdowns that undermine the social fabric of communities.

The differential effect of various policing strategies point to another empirical concern, namely, the identification of the different channels through which disorder can affect crime. Kelling and Wilson (1982) maintained that the effects of disorder manifest themselves through an increased fear of crime associated with neighborhood disorder. In turn, this leads to reduced social controls as people move away and are less inclined to engage in crime prevention or intervention. This sequence of events contrasts with more traditional explanations of incapacitation and deterrence associated with increased police arrests, presence contact, and surveillance.

Following this logic, Weisburd et al. (2015) argue that to test the BWE, it is not enough to look at the effect of disorder policing on crime rates without taking into account the effects on fear or crime perceptions. To do so, the authors identify six studies which allow the identification of such effects. In a meta-analysis of these studies, they do not find consistent evidence that disorder policing reduces fear of crime. Moreover, the effect sizes differ starkly across studies, and

confidence intervals around the estimated effects are large, implying a need for more evidence.

Conclusion

The original formulation of the BWE by Kelling and Wilson has resulted in several decades of lively debate on the proper role of police in society and has provided a big stimulus for empirical research on community policing. Although the evidence for the BWE seems underwhelming at present, it is hard to draw firm conclusions, as much of the available research is not well suited for the identification of causal patterns, or to distinguish between the BWE and more traditional theories.

To move the discussion forward, research designs should focus on the use of random variation in policing strategies. Moreover, a better understanding of the links between disorder and crime requires investigation of the specific social mechanisms under consideration. This implies a need to formulate detailed theories from which one can derive testable hypotheses. The set of outcome variables will have to go beyond crime reports and arrest rates to include perceptions of safety, trust, willingness to intervene, and other variables related to community enforcement (Sampson et al. 1997). These studies will require careful coordination between academics and police forces. Whether the BWE turns out to be real or not, this approach will allow the evidence-based policing that makes society safer.

Cross-References

- [Causation](#)
- [Cost of Crime](#)
- [Crime: Economics of, the Standard Approach](#)
- [Crime: Expressive Crime and the Law](#)
- [Crime, Incentive to](#)
- [Criminal Sanctions and Deterrence](#)
- [Empirical Analysis](#)
- [Prosocial Behaviors](#)
- [Public Enforcement](#)

References

- Braga A et al (1999) Problem-oriented policing in violent crime places: a randomized controlled experiment. *Criminology* 37(3):541–580
- Braga A, Welsh B, Schnell C (2015) Can policing disorder reduce crime? A systematic review and meta-analysis. *J Res Crime Delinq* 52(4):567–588
- Bratton W, Knobler P (1998) Turnaround: how America's top cop reversed the crime epidemic. Random House, New York
- Corman H, Mocan N (2005) Carrots, sticks and broken windows. *J Law Econ* 48:251
- Eck J, Maguire E (2000) Have changes in policing reduced violent crime? An assessment of the evidence. In: Blumstein A, Wallman J (eds) *The crime drop in America*, rev. edn. Cambridge University Press, New York, pp 207–265
- Fagan J, Zimring F, Kim J (1998) Declining homicide in New York City: a tale of two trends. *J Crim Law Criminol* 88(4):1277–1324
- Harcourt B (1998) Reflecting on the subject: a critique of the social influence conception of deterrence, the broken windows theory, and order-maintenance policing New York style. *Mich Law Rev* 97(2):291–389
- Harcourt B, Ludwig J (2006) Broken windows: new evidence from New York City and a five-city social experiment. *Univ Chic Law Rev* 73:271–320
- Kelling G, Bratton W (1998) Declining crime rates: insiders' views of the New York city story. *J Crim Law Criminol* 88(4):1217–1232
- Kelling G, Sousa W (2001) Do police matter?: an analysis of the impact of New York city's police reforms. Civic report no. 22. Manhattan Institute, New York
- Kelling G, Wilson Q (1982) Broken windows. *Atl Mon* 127(2). <http://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/>
- Orr L et al (2003) Moving to opportunity: interim impacts evaluation. U.S. Department of Housing and Urban Development Office of Policy Developments and Research report. http://www.abtassociates.com/reports/2003302754569_71451.pdf
- Sampson R, Raudenbush S (1999) Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. *Am J Sociol* 105(3):303–651
- Sampson R, Raudenbush S, Earls F (1997) Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science* 277:918–924
- Taylor R (2001) Breaking away from broken windows: Baltimore neighborhoods and the Nationwide fight against crime, guns, fear and decline. Westview, Boulder
- Weisburd D, Eck J (2004) What can police do to reduce crime, disorder, and fear? *Ann Am Acad Pol Soc Sci* 593(1):42–65
- Weisburd D et al (2015) Understanding the mechanisms underlying broken windows policing: the need for evaluation evidence. *J Res Crime Delinq* 52(4):589–608
- Zimbardo P (1973) A field experiment in Autoshaaping. In: Ward C (ed) *Vandalism*. Architectural Press, London, pp 85–90

Budgetary Institutions

Ringa Raudla

Faculty of Social Sciences, Ragnar Nurkse School of Innovation and Governance, Tallinn University of Technology, Tallinn, Estonia

Abstract

Budgetary institutions encompass two different types of institutional arrangements: fiscal rules and budget process rules. Fiscal rules entail substantive constraints on public spending, taxation, deficit and debt, usually in the form of explicit quantitative targets. Budget process rules, in turn, entail procedural aspects of public budgeting: they establish the competencies of the actors involved in the budget process and outline the procedures that govern the preparation, adoption and implementation of the budget. This entry gives an overview of the different forms and characteristics of these institutions and also of their impact on budgetary outcomes.

Definition

Budgetary institutions encompass two different types of institutional arrangements: fiscal rules and budget process rules. Fiscal rules entail substantive constraints on public spending, taxation, deficit, and debt, usually in the form of explicit quantitative targets. Budget process rules, in turn, entail procedural aspects of public budgeting: they establish the competencies of the actors involved in the budget process and outline the procedures that govern the preparation, adoption, and implementation of the budget.

Introduction

Because of the recent fiscal crises in Europe and elsewhere, budgetary institutions have received more attention than ever. On the one hand, faulty

budgetary institutions have been blamed for their failure to secure fiscal discipline. On the other hand, these institutions have been viewed as triggering unnecessarily painful austerity measures. This entry gives an overview of two main types of budgetary institutions: fiscal rules and budget process rules. Fiscal rules entail substantive constraints on public spending, taxation, deficit, and debt, usually in the form of explicit quantitative targets. Budget process rules, in turn, entail procedural aspects of public budgeting: they establish the competencies of the actors involved in the budget process and outline the procedures that govern the preparation, adoption, and implementation of the budget.

Law and Economics analysis can be particularly helpful in structuring the analysis of budgetary institutions. On the one hand, Law and Economics approach can be utilized for examining more systematically the impacts of different types of fiscal rules and budget process rules on the functioning of the economy as a whole. On the other hand, Law and Economics can provide useful conceptual and analytical tools for understanding how exactly different types of budgetary institutions shape the behavior of actors involved in the budget process and what kind of factors influence the actual achievement of the goals set for these institutions. For example, Law and Economics approaches could be used for undertaking detailed examinations of how fiscal rules and budget process rules structure the incentives of the different actors involved in the budget process and how these configurations of incentives influence the actors' interactions and the adjustments in their behavior. Law and Economics perspectives could also be used for examining the actors' actual motivations for establishing fiscal and budgetary process rules and for uncovering the unintended effects of these rules.

In the next sections, both of these two approaches will be drawn upon. First, different forms and characteristics of budgetary institutions will be outlined, followed by a discussion of their impact on the behavior of the budget actors and on the fiscal outcomes (for a more detailed overview, see Raudla 2010a).

Fiscal Rules

Different Types of Fiscal Rules

Output-oriented or substantive fiscal rules, also called numerical fiscal rules, are rules that pertain directly to particular aggregate parameters of the budget. Kopits and Symansky (1998, p. 2) define a fiscal rule as “a permanent constraint on fiscal policy, expressed in terms of a summary indicator of fiscal performance.” The four main types of fiscal rules are deficit rules (e.g., balanced budget requirements or deficit limits), debt rules (e.g., debt ceilings), expenditure rules (e.g., ceilings on general public expenditure growth or on certain types of government spending), and tax revenue rules (e.g., limits on the increase in the overall tax burden) (Corbacho and Schwartz 2007). Fiscal rules can be stipulated either in constitutions or organic budget laws (framework laws governing budgetary decision-making of the government).

Balanced budget rules and deficit constraints have received the most extensive attention in discussions over fiscal rules thus far (von Hagen 2007). The general category “balanced budget requirement,” however, may involve rather different rules in terms of scope and content (Poterba 1995). One can differentiate between balanced budget rules in three different aspects. First, the rules vary with respect to the *stage* of the budget process they apply to: the executive may be required to submit a balanced budget, the legislature may be required to approve a balanced budget or the government can be required to balance the budget at the end of the year. Second, balanced budget rules may vary in terms of the *type of funds* they cover: the rule may apply only to general funds but may also cover capital funds or special revenue funds. Inasmuch as the balanced budget rule applies only to the current budget, it is equivalent to the “golden rule,” which requires that borrowing may only be used to fund investment. Third, balanced budget requirements can vary with regard to when and how the legislature can *override or suspend* the requirement: the provisions can entail specific circumstances and special majority requirements.

Fiscal rules pertaining to *tax and expenditure limits* have received relatively less attention

compared to debt and deficit rules so far, and most of the literature until recently was primarily concerned with the tax and expenditure limits (TEs) adopted in the US states during the “tax revolt” of the 1970s. Since the beginning of 1990s, a number of European countries have adopted expenditure rules, bringing about resurgence in the interest of such rules and their impacts (von Hagen 2008). Tax and expenditure ceilings can be established either in nominal or real terms, and limits can be tied to growth in personal income, inflation, or population. Expenditures can also be limited to a percentage of projected revenues, with the rest of the revenues channelled to a rainy-day fund. In terms of coverage, expenditure limits can set a ceiling on primary expenditure, wage (and pension) expenditure, interest payments, and debt service.

How Can Fiscal Rules Shape Fiscal Policy?

Fiscal rules can be expected to influence fiscal policy by enhancing the *accountability* of policy-makers and by mitigating the *common-pool problems* inherent in budgetary decision-making.

First, since fiscal rules establish clear benchmarks with which actual policies can be compared, they can be viewed as enhancing the accountability of the legislature and the executive (von Hagen 2007). If, for example, the allowed budget deficit is fixed at a certain percent of GDP, voters can use this indicator as a yardstick for evaluating the prudence of policy choices undertaken by their representatives (Schuknecht 2005). Further, fiscal rules can heighten attention to budgeting and attract closer monitoring by interested voters and interest groups.

Second, fiscal rules can mitigate the *common-pool problems* involved in public budgeting and taxation. It is often argued that the “fiscal commons” or “budgetary commons” are subject to similar tragedies as the natural commons, leading to a deficit bias in fiscal policy-making (Buchanan and Wagner 1977; Tullock 1959; Velasco 2000; Wagner 2002). The gist of the idea of budgetary commons is that the participants involved in budgeting internalize full benefits of a spending proposal but bear only a fraction of the cost, since

it is financed from the common tax fund. The divergence between *perceived* and *actual* costs of programs, in turn, would lead the herders on the budgetary commons to demand higher levels of particularistic expenditures than would be efficient, leading to increasing expenditures, and, in a dynamic context, to higher deficits (for an overview of different models of budgetary commons, see Raudla 2010b). The smaller the fraction that each grazer has to bear, the smaller become the perceived costs compared with the actual costs and the more severe the common-pool problems of budgeting are likely to be. Fiscal rules are considered to enhance the coordination of budgetary and tax policy-making by providing the actors involved with a clear focal point that can facilitate the internalization of decision externalities entailed in spending or borrowing.

Trade-Offs Involved in Designing Fiscal Rules

Kopits and Symansky (1998) emphasize that if fiscal rules were to be *effective*, they should be well defined (with regard to the indicator to be constrained, institutional coverage and specific escape clauses), transparent, simple, flexible (to accommodate exogenous shocks), adequate (with respect to the specified goal), enforceable, consistent (both internally and with other policies), and efficient. They point out, though, that no fiscal rule can fully combine all these features. Hence, there are usually significant trade-offs that have to be made in establishing fiscal rules, which may underline their overall effectiveness.

The most important trade-off that has to be made in designing fiscal rules is between *simplicity* and *flexibility*. In order to strike an optimal balance between these two features, one has to keep in mind how the fiscal rule is foreseen to be *enforced* (Schick 2003). In other words, if we think that a fiscal rule is necessary, then what the best rule would be depends on what we consider to be the main enforcement mechanism for the rule. If we assume that the main reason politicians stick to the fiscal rule is that they are afraid of the *electoral backlash* when they deviate from the rule, then simplicity and transparency of the rule is a precondition for it to work. Simplicity and

transparency of a fiscal rule imply that the rule is established in a constitution and entails a straightforward numerical target which has to be achieved without any exceptions.

The main problem with such very simple rules (like is the case with the Maastricht deficit target of 3% of GDP) is that they may prevent macro-economic stabilization via automatic stabilizers and fiscal stimulus. Hence, such a simple and transparent budget balance rule may needlessly prolong an economic downturn and this could prove self-defeating in the long run (Anderson and Minarik 2006). Conversely, during good times, a simple headline balance rule may encourage cyclically loose fiscal policies because it may not give sufficient guidance about how large surpluses the government should run.

One possible solution to alleviate pro-cyclicality in fiscal policy-making is to make the rule more flexible and to require the fiscal policy to adhere to a cyclically adjusted balance or a structural balance, which would allow possible deviations in the case of severe economic recessions and other emergencies. However, in the case of such more sophisticated rules (as outlined in the Fiscal Compact, for example) the general public may not be able to evaluate whether the government has complied with the fiscal rule or not (Alesina and Perotti 1999; Schuknecht 2005). Indeed, even economists may not be able to say with full confidence what the cyclically adjusted balance actually is at any point in time because of the difficulties and uncertainties involved in calculating the potential output and the revenue and expenditure elasticities (Larch and Turrini 2010). Indeed, the ex ante, real-time, and ex post assessments of the structural and cyclically adjusted balances may diverge significantly. Furthermore, if targets are set in cyclically adjusted terms, it may give rise to moral hazard in forecasting (Anderson and Minarik 2006).

Thus, in the case of such more “sophisticated” fiscal rules, relying on the electorate as the main enforcement mechanism is not feasible anymore. Alternative mechanisms entail enforcement by the constitutional courts, financial markets, and independent fiscal councils. All these

enforcement mechanisms, however, have their own shortcomings.

First, the (constitutional) courts may not have either the willingness or the competence to evaluate the structural or cyclical balances and intervene in fiscal policy-making (for a theoretical discussion on the role of courts in fiscal policy, see Raudla 2010a). Furthermore, even if the constitutional courts are willing to pass decisions on the violations from the fiscal rule, the legislature may simply choose to ignore those judgments.

Second, the experience with financial markets as enforcers of fiscal discipline is not too promising either: they can often be either too “slow” or too “neurotic” or both, first too slow to react and then overreact. In other words, the financial markets do penalize fiscal profligacy, but they do it in a rather discontinuous fashion and only with significant time lags and often only at extreme stage (Debrun et al. 2009).

Third, the use of independent fiscal councils as an institutional device for helping to enforce fiscal rules (see Calmfors and Wren-Lewis 2011; Debrun et al. 2009; Wyplosz 2005) appears to be the most attractive option, at least theoretically. How well the fiscal council can act as an enforcer of fiscal rules depends, of course, on what exactly their mandate is, how their independence from the political system is guaranteed, what resources they have for conducting independent analyses, etc. Besides monitoring the government’s compliance with the fiscal rule(s), the Fiscal Council could also contribute to economic policy discussions in the public sphere and raise the level of public debate on macroeconomic issues. If well designed, the Council can serve as an “interface” between the general public and the government.

In some discussions over fiscal councils, it has been proposed that independent fiscal authorities or fiscal agencies could even *replace* fiscal rules (Debrun et al. 2009; Wyplosz 2005). Such proposals draw on the experience with delegating monetary policy-making to independent Central Banks and argue that, in a similar fashion, fiscal policy-making could be handed over to independent (and unelected) fiscal authorities who would be better able (than elected politicians) to strike

a balance between long-term fiscal sustainability and short-term flexibility.

In the case of supranational fiscal rules (like is the case in the European Union), the national governments may face sanctions for violating the established fiscal rules, which, at least in principle, should enhance their compliance with the rules. As the experience with the Stability and Growth Pact shows, however, the supranational body, if controlled by member states themselves, may be unwilling to actually impose the sanctions (and often with a good reason). To what extent the Fiscal Compact (enacted in 2013) which tries to make the enforcement of the prescribed rules more credible – by involving the European Commission and the European Court of Justice – will fare better remains to be seen (e.g., Bird and Mandilaras 2013).

In the end, however, the effectiveness of fiscal rules depends on political leaders: the rules will work if politicians want them to work and will not work if the political commitment is lacking. If politicians don't want to comply with the fiscal rules, they will usually find a way to evade them, either explicitly or implicitly by engaging in creative accounting and off-budget operations. As Schick (2003) has noted, in countries where fiscal rules are most needed, they may be least workable and where conditions are most hospitable to fiscal constraints, they may be the least needed.

Budget Process Rules

Procedural budgetary institutions (or budget process rules) – which can be contrasted with the output or target-oriented fiscal rules – and their influence on budgetary outcomes have received increasing attention in the political economy literature since the beginning of 1990s. Procedural budgetary institutions are a set of rules that guide the decision-making process in formulating, approving, and implementing the budget. Using Ostrom's (1986) terminology, these budgetary institutions entail position, authority, aggregation, information, and payoff rules that guide and constrain the behavior and decisions of political actors in the action arena of budgeting. In other

words, budget process rules can be seen as the "rules of the game" that shape the interaction of the participants involved in budget-making process by distributing strategic influence among them and regulating the flow of information.

How Can Budget Process Rules Influence Budgetary Decision-Making?

Like the case with fiscal rules, it is argued that budget process rules help to alleviate common-pool problems of budgeting. As mentioned above, common-pool problems of budgeting arise from the fact that constituencies who benefit from a particular program have to bear only a fraction of the costs involved in providing this program. Given the divergence between the perceived costs and benefits involved in targeted spending, the elected officials (representing geographically or socially defined constituencies) either in the legislature or executive have incentives to demand excessive levels of expenditures on targeted programs.

As the literature on fiscal governance argues, however, properly designed budget processes can lead the participants in the budget process to internalize the costs of spending decisions and the budget deficit. In discussions over different kinds of fiscal governance arrangements (or budget process rules), a *centralized* budget process is often contrasted with a *fragmented* budget process (Alesina and Perotti 1996, 1999; Fabrizio and Mody 2006; von Hagen 2002). A fragmented process is seen as more likely to give rise to excessive spending and deficits, while a centralized process is considered to be conducive to fiscal discipline.

As von Hagen (2002) emphasizes, the institutional elements of centralization can be relevant at all different stages of the budget process: the planning and drafting of the budget by the executive, the adoption of the budget by the parliament, and the implementation of the budget.

The executive phase can be characterized as centralized, when it promotes the setting of spending and deficit (or surplus) targets at the outset; conversely, it can be designated as fragmented, when the resulting budget is merely a sum of uncoordinated bids from individual ministries.

At the legislative approval stage, the process can be described as fragmented when the following elements are present: the parliament can make unlimited amendments to the draft budget submitted by the executive, spending decisions can be made by legislative committees with narrow and dispersed authorities, and there is little guidance of the parliamentary process either by the executive or by the speaker. Elements of centralization at the adoption stage give the executive or the budget committee an important role in setting the agenda of budget proceedings in the legislature, limit legislative amendments to the budget, require amendments to be offsetting, postulate the legislature to vote on the total budget size before the approbation of single provisions, and raise the political stakes of rejecting the executive's budget (e.g., by making this equivalent to a vote of no confidence).

At the implementation stage, elements of centralization assure that the adopted budget would actually be the basis for the spending decisions of the executive. Thus, implementation can be regarded as centralized when it is difficult to change the existing budget document or to adopt supplementary budgets during the fiscal year, when transfers of funds between chapters are forbidden or limited, and when unused funds cannot be carried over to the next year's budget. Further elements of centralization in the implementation stage include extensive powers of the finance minister to monitor and control spending flows during the fiscal year (e.g., by blocking expenditures and imposing cash limits on spending ministers) and sanctioning the disbursement of funds by a central unit in the ministry of finance (von Hagen 2002; Hallerberg et al. 2007).

Different Modes of Fiscal Governance

Hallerberg and von Hagen (1999) distinguish between two modes of fiscal governance that help to resolve the common-pool problems of budgeting by inducing decision-makers to internalize the budgeting externalities: the *delegation* approach and the *fiscal contracts* (also called the *commitment*) approach.

The *delegation* approach involves lending significant strategic powers to the finance minister,

who can be assumed to take a more comprehensive view of the budget than other actors and who would have incentives to internalize the costs of public spending programs (see also Alesina and Perotti 1999). In the delegation approach, the preparation phase of the budget procedure is characterized by strong agenda-setting powers of the finance minister vis-à-vis the spending ministers: the finance minister makes binding proposals for broad budgetary categories, negotiates directly with the individual ministries, approves the bids submitted to the final cabinet meeting, and can act as a veto player over budgetary issues in the cabinet. In the legislative approval phase of the process, the delegation approach limits the rights of the legislature to amend the budget (in order to avoid major changes in the executive's budget proposal) and gives the executive strong agenda-setting powers vis-à-vis the legislature. In the implementation phase, the minister of finance has strong monitoring powers regarding the actual use of budget appropriations by the spending ministers and the authority to prevent and correct any deviations from the budget plan (Alesina and Perotti 1996, 1999; von Hagen 2002, 2008; von Hagen and Harden 1995; Hallerberg et al. 2009).

Under the *fiscal contracts* approach, at the beginning of the annual budget cycle, all members of government negotiate multilaterally and commit themselves to a key set of budgetary parameters or fiscal targets (usually spending targets for each ministry) that are considered to be binding for the rest of the budget cycle. It is assumed that the process of collective negotiations would lead the different actors involved to understand and take into account the externalities entailed in budgetary decisions. During the rest of the budget preparation process, the minister of finance has the responsibility to evaluate the consistency of the budget proposals submitted by the spending ministers with the agreed targets. At the legislative stage, the fiscal contracts approach puts less emphasis on constraining legislative budgetary amendments and more emphasis on the legislature's role in monitoring the compliance of the executive's budget with the fiscal targets, by giving the legislature significant rights to demand budgetary information from the executive. In the

implementation phase, the minister of finance would have strong monitoring and control rights to prevent and correct deviations from the adopted budget; while the delegation approach emphasizes the need for managerial discretion by the minister of finance, allowing him to react flexibly to changing budgetary circumstances, the contracts approach is characterized by contingent rules for dealing with unforeseen events (von Hagen 2002, 2007, 2008; von Hagen and Harden 1995; Hallerberg and von Hagen 1999; Hallerberg et al. 2009).

These two ideal types of fiscal governance are usually contrasted with the *fiefdom* approach to budgeting (Hallerberg 2004), which is characterized by fragmented, uncoordinated, and ad hoc decision-making in all phases of the budgetary process. Under the fiefdom approach, budgetary decision-making is expected to be plagued by common-pool problems, leading to lower levels of fiscal discipline than the delegation or contracts approach. In addition, Hallerberg (2004) has pointed to a *mixed* type of fiscal governance, which employs elements of delegation in the preparation phase but is followed by a contract-like adoption phase, with the parliament playing an equally strong role in the passage of the budget alongside the executive.

Which Form of Fiscal Governance Is Most Effective?

Which of the fiscal governance mechanisms is more effective in solving the common-pool problems of budgeting? Hallerberg and von Hagen (1999) conjecture that the delegation approach is likely to be more successful in safeguarding fiscal discipline because enforcement and punishment mechanisms are more credible under the delegation approach. They argue that under the delegation approach, the defecting minister can be punished by dismissal – which is a harsh punishment for an individual but does not bring about severe consequences for the government as a whole. Under the contracts approach, however, the defecting minister cannot usually be dismissed so easily, since the distribution of portfolios is decided in the coalition agreement; thus, ousting a minister from another party could be regarded as

an interference with the internal issues of the other parties in the coalition. Hallerberg and von Hagen (1999) note that the main sanction under the contracts approach – when a minister deviates from the agreed budget targets – is the breakup of the coalition. That, however, would be a harsh punishment for the government as a whole. Further, this “punishment” would only be credible if there are other parties in the parliament with whom alternative coalitions could be formed by the party who seeks to penalize the coalition partner (s) for reneging on the agreement. Thus, given that the parties do not always have incentives to monitor and punish each other, it is more difficult to uphold successfully the negotiated targets than it is to maintain a strong finance minister.

In the more recent discussion on fiscal governance, it is argued that in the case where the government is subjected to supranational fiscal rules, having a commitment (or contracts) type of fiscal governance can facilitate the adherence to fiscal rules (Hallerberg 2004; von Hagen 2008). As Hallerberg (2004, p. 194) explains, the domestic fiscal targets and external fiscal rules can be mutually reinforcing: the external fiscal rule can enhance the commitment of coalition partners to fiscal targets, and the institutions that contract states use for monitoring the adherence to the fiscal contract by coalition partners foster compliance with the externally imposed fiscal targets.

Concluding Remarks

The experiences with the recent fiscal crises indicate that a lot of theoretical and empirical work still remains to be done in order to better understand how budgetary institutions actually influence fiscal policy-making and budgetary outcomes.

In addition, there are many issues that have not been sufficiently addressed by studies on budgetary institutions so far. For example, the existing studies on budgetary institutions have not paid enough attention to the question of which institutional aspects would be conducive to maintaining fiscal discipline and which aspects would be conducive to undertaking fiscal adjustments in the

face of fiscal shocks. Also, it would be worth investigating how fiscal shocks or other developments in the fiscal environment bring about changes in budgetary institutions. Finally, when examining the impacts of budgetary institutions, the resulting spending mix – e.g., in terms of particularistic (pork-barrel) expenditures vs. broader public programs (e.g., social insurance) – should also be considered, in addition to the effects on fiscal discipline.

In further analyses of budgetary institutions, systematic use of Law and Economics approaches could be particularly useful, both for developing the understanding about the impacts of budgetary institutions on the functioning of the economy as a whole and also on the microlevel behaviors of the budget actors.

References

- Alesina A, Perotti R (1996) Fiscal discipline and the budget process. *Am Econ Rev* 86(2):401–407
- Alesina A, Perotti R (1999) Budget deficits and budget institutions. In: Poterba JM, von Hagen J (eds) *Fiscal institutions and fiscal performance*. University of Chicago Press, Chicago
- Anderson B, Minarik JJ (2006) Design choices for fiscal policy rules. *OECD J Budget* 5(4):159–203
- Bird G, Mandilaras A (2013) Fiscal imbalances and output crises in Europe: will the fiscal compact help or hinder? *J Econ Policy Reform* 16(1):1–16
- Buchanan JM, Wagner RE (1977) *Democracy in deficit*. Academic, London
- Calmfors L, Wren-Lewis S (2011) What should fiscal councils do? *Econ Policy* 26(68):649–695
- Corbacho A, Schwartz G (2007) Fiscal responsibility laws. In: Kumar MS, Ter-Minassian T (eds) *Promoting fiscal discipline*. IMF, Washington, DC
- Debrun X, Hauner D, Kumar MS (2009) Independent fiscal agencies. *J Econ Survey* 23(1):44–81
- Fabrizio S, Mody A (2006) Can budget institutions counteract political indiscipline? *Econ Policy* 21(48):689–739
- Hallerberg M (2004) *Domestic budgets in a united Europe: fiscal governance from the end of Bretton Woods to EMU*. Cornell University Press, Ithaca
- Hallerberg M, von Hagen J (1999) Electoral institutions, cabinet negotiations, and budget deficits in the European Union. In: Poterba JM, von Hagen J (eds) *Fiscal institutions and fiscal performance*. University of Chicago Press, Chicago
- Hallerberg M, Strauch R, von Hagen J (2007) The design of fiscal rules and forms of governance in European Union countries. *Eur J Political Econ* 23(2):338–359
- Hallerberg M, Strauch R, von Hagen J (2009) *Fiscal governance in Europe*. Cambridge University Press, Cambridge
- Kopits G, Symansky S (1998) *Fiscal policy rules*. IMF, Washington, DC
- Larch M, Turrini A (2010) The cyclically adjusted budget balance in EU fiscal policymaking. *Intereconomics* 45(1):48–60
- Ostrom E (1986) An agenda for the study of institutions. *Public Choice* 48(1):3–25
- Poterba JM (1995) Balanced budget rules and fiscal policy: evidence from the States. *National Tax J* 48(3):329–337
- Raudla R (2010a) Constitution, public finance, and transition. Peter Lang, Frankfurt am Main
- Raudla R (2010b) Governing budgetary commons: what can we learn from Elinor Ostrom? *Eur J Law Econ* 30:201–221
- Schick A (2003) The role of fiscal rules in budgeting. *OECD J Budget* 3(3):7–34
- Schuknecht L (2005) Stability and growth pact: issues and lessons from political economy. *Int Econ Econ Policy* 2(1):65–89
- Tullock G (1959) Some problems of majority voting. *J Polit Econ* 67(6):571–579
- Velasco A (2000) Debts and deficits with fragmented fiscal policymaking. *J Public Econ* 76(1):105–125
- von Hagen J (2002) Fiscal rules, fiscal institutions, and fiscal performance. *The Econ Social Rev* 33(3):263–284
- von Hagen J (2007) Budgeting institutions for better fiscal performance. In: Shah A (ed) *Budgeting and budgetary institutions*. The World Bank, Washington, DC
- von Hagen J (2008) European experiences with fiscal rules and institutions. In: Garrett E, Graddy A, Jackson HE (eds) *Fiscal challenges: an interdisciplinary approach to budget policy*. Cambridge University Press, Cambridge
- von Hagen J, Harden I (1995) Budget processes and commitment to fiscal discipline. *Eur Econ Rev* 39(3–4):771–779
- Wagner RE (2002) Property, taxation, and the budgetary commons. In: Racheter DP, Wagner RE (eds) *Politics, taxation, and the rule of law: the power to tax in constitutional perspective*. Kluwer, Boston
- Wyplosz C (2005) Fiscal policy: institutions versus rules. *National Inst Econ Rev* 191:64–78

Calabresi: Heterodox Economic Analysis of Law

Alain Marciano^{1,2} and Giovanni Battista Ramello^{3,4}

¹MRE and University of Montpellier, Montpellier, France

²Faculté d'Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France

³DiGSPES, University of Eastern Piedmont, Alessandria, Italy

⁴IEL, Torino, Italy

Abstract

Guido Calabresi is one of the founders of the law and economics movement. His approach, however, corresponds to a form of economic analysis of law that, we claim, is heterodox. We show why in this short notice.

Biography

Born on October 18, 1932, in Milan, Guido Calabresi migrated with his family in the USA in 1939. After having received his Bachelor of Science degree (summa cum laude) from Yale College in 1953, majoring in economics, his Bachelor of Arts from Magdalen College at Oxford University in 1955, he got his Bachelor of Laws (LL.B.) magna cum laude from Yale Law

School in 1958. Calabresi started by clerking for Justice Hugo Black, who was then US Supreme Court Associate, from 1958 to 1959, and then joined the Yale Law School (instead of the University of Chicago Law School where he was offered full professorship). Calabresi was appointed US Circuit Judge of the US Court of Appeals for the Second Circuit in 1994. He still serves as Sterling Professor Emeritus and Professorial Lecturer in Law at the Yale Law School.

Law and economics emerged just after World War II, gained structure in the 1950s, took further shape in the 1960s, and established itself in the 1970s, essentially under the influence of economists and legal scholars from the so-called Chicago school. The use of economics as a parsimonious tool to tackle otherwise difficult and complex legal problems has now become so frequent that a number of scholars regard this as possibly the most important novelty in all of modern legal scholarship (Denoza 2013; Mattei 1994).

It was Richard Posner who, formally, “invented” the economic analysis of law at the beginning of the 1970s when he published the discipline’s eponymous masterpiece (Posner, 1972), launched the *Journal of Legal Studies*, and started to write articles in which he explained that economics is an important tool that can be used to analyze (in particular) legal phenomena. This accounts for the North American, Anglo-Saxon origins of law and economics. But the field was from the outset a melting pot in the broadest sense – not just because it was a field of

studies created at the intersection of economics and law but also because it grew out of a blend of North American and European cultures (Ramello 2016). Ronald Coase, another founder of law and economics, was born in England. And Guido Calabresi was born in Italy, where he spent part of his childhood there. His family moved to the USA to escape fascism and brought with them a lively Italian and European bourgeois environment. Even if attempting to deduce impacts from historical backgrounds is generally a risky business (Kalman 2014, p.15), there is strong evidence that Calabresi did indeed blend his family's European culture with that of the USA and that this in turn had an influence on his scholarship. Once, when asked what he considered to be the most important part of his legal education, Calabresi replied:

"I am a refugee!" And of course, how can I not have been influenced by the fact that we were antifascists and that we left Italy because my father had been jailed and beaten in 1923 and he was a democrat with a small 'd'; that we were very, very rich there and came here with nothing because it was against the law under penalty of death? If I write about capital punishment or if I make a decision, I am not going to be writing to push an agenda but, on the other hand, I would be pretty foolish not to be aware of the fact that that is in my background (Benforado and Hanson 2005, p. 75).

It is not our purpose, here, to look for and find specific traces of European elements in his work, but we nonetheless suggest that Calabresi's work evidences this European heritage. One major aspect of this heritage is what could be called the "comparative" dimension or the adoption of a "comparative viewpoint." This was in particular the case with "Some Thoughts on Risk Distribution and the Law of Torts" (Some Thoughts), Calabresi's first article. Written in the 1950s when Calabresi was still a student, it was published in 1961. That was almost when Coase published his own path-breaking article, "The Problem of Social Cost" (1960). These two works are comparative in the sense that they were using the insights of another discipline to improve another one. While Coase was trying to improve his understanding of economic phenomena by relying on court cases, Calabresi was

proposing to use economics to improve one's understanding of economic phenomena. In other words, with "Some Thoughts," Calabresi was probably the first to apply economic methods to analyze legal questions. He really initiated an original approach: of course different from what most legal scholars did but also from what economists – Coase, for instance, or Aaron Director – started to do. Calabresi is not only a founder of the law and economics movement but also of an economic analysis of law.

Commentators on Calabresi's works perceived this innovation at the time of publication. For instance, Walter Blum and Harry Kalven observed the novelty of Calabresi's perspective as soon as they started to study and comment on his work and noted that he had "crystallized the economic analysis of liability" (1967, 240). Posner himself also underscored the change of direction initiated by Calabresi with respect to Coase in his 1970 review of Calabresi's *The Costs of Accidents* (1971). Moreover, in 1971 Frank Michelman, also commenting on *The Costs of Accidents*, noted that Calabresi "provide[s] a conceptual apparatus for describing, comprehending, and evaluating systems of accident law." Strictly speaking, Calabresi's approach was and remains a form of economic analysis of law, because he uses economics as a tool for analyzing legal issues (see Marciano 2012).

Yet Calabresi himself continues to insist that his work should be viewed as a form of law and economics rather than as an economic analysis of law and that he prefers to see his contribution grouped with Coase's rather than with Posner's, whom he strongly disagrees with and even opposes. In his most recent book, he anchored the distinction between his and Posner's approach – to put it differently, between law and economics and an economic analysis of law – in the opposition between Jeremy Bentham and John Stuart Mill. He also insists on the need for an economic approach to law which should rest on a broader cognitive framework than the one used by neoclassical economists. Let us note this very claim was already present in "Some Thoughts," which is indeed remarkable not just because it is foundational for law and economics but also

because it is foundational for Calabresi. In this article, in accordance with the economic analysis of law, Calabresi used economics to guide legal action. But, on the other hand, he recognized that in certain settings “traditional economic theory [can] be of little help,” he equally acknowledges the role of laws in fostering economic efficiency, as in the law and economics view. This twofold orientation not only places economics and law on an equal footing, it also treats them both as instruments serving higher goals connected with basic individual liberties, which the market alone is not always able to promote.

For instance, in a 2014 article, Calabresi explained the public function of torts: The liability rule (in both torts and in takings and eminent-domain law) is not used principally, much less solely, to approach the result that would occur in a free market of consensual exchanges (were such a market available) but is instead used approach inalienability (i.e., a fully collective result) in those instances when a criminal law solution is not desired. Calabresi expands on this argument by showing that the liability rule (of the collectively set price) is used to achieve goals that are neither purely libertarian nor purely collectivist but are properly viewed as social democratic.

To understand Calabresi’s approach, one must take into account the distinction between choice and consent. Usually, at least in neoclassical economics, individual choices are supposedly made under certain conditions, to which the choosers are assumed to implicitly consent. Consent is thus never discussed or considered in any way distinct from choice. The role of law is precisely to defend consent and to intervene in an efficient way whenever this principle is violated.

Whereas standard economic analyses of law assume that choice means consent, Calabresi insists on the discrepancy between choice and consent, which arises in many practical situations involving legal intervention. To him, the conditions of choice should not be treated as trivially exogenous features of the setting in which legal action – possibly guided by economic efficiency – is played out but rather as a fully fledged part of the decision set, which the legal system must carefully consider. From this viewpoint it follows

that the role of law and economics is to provide a method for examining these complex issues and arriving at solutions that consider not only social welfare (and, by implication, efficiency) but also other matters connected to individual rights and liberties (see Marciano and Ramello, 2014).

Thus, Calabresi raised the problem of the potential lack of consent – arising from monopoly, individuals’ lack of rationality, and their vulnerability to external pressures, or systemic imperfections – with the implication that individuals do not always make choices that correspond to their preferences. This in its turn enabled law and economics scholars to contribute by providing a wider framework for decision-making that uses the efficiency criterion but also explicitly combines it with other principles, such as societal welfare and individual liberties. The implication is that the questions social scientists have to tackle cannot always be reduced to optimal allocation of resources and instead frequently require enquiring about the “starting points,” conditions of choice, and consent to those conditions.

Impact and Legacy

Calabresi is one of the founders of the law and economics movement and was even one of the first to suggest that economics could be used to analyze legal phenomena. One of his main insights and legacy is to have explained that and how law and economics complement each other. To him, economics is concerned with choice under certain given conditions that, as we have noted, may not be satisfactory. What economics provides is only a framework, which needs to be normatively qualified by judges and the legal system. Therefore, if for Calabresi, “law and economics proves the more challenging and worthwhile endeavor” than the economic analysis of law, it is because he envisages law and economics as a back-and-forth dialogue between the two disciplines (Kalman 2014). This equal footing of law and economics is what the economic analysis of law tends to preclude, because it essentially downgrades economics to a mere

problem-solving technology. To be sure, Calabresi sees economics as providing road signs – “road signs that are not too misleading to be worth spending time on” – that judges and lawmakers can then use to serve a higher good than simply fostering efficiency.

Cross-References

- [Coase, Ronald](#)
- [Economic Analysis of Law](#)
- [Posner, Richard](#)

References

- Benforado A, Hanson J (2005) The costs of dispositionism: the premature demise of situationist economics. *Md Law Rev* 64:24–84
- Blum W, Kalven H Jr (1967) The empty Cabinet of Dr. Calabresi: auto accidents and general deterrence. *Univ Chic Law Rev* 34(2):239–273
- Calabresi G (1961) Some thoughts on risk distribution and the law of torts. *Yale Law J* 70(4):499–553
- Calabresi G (1971) The costs of accidents: a legal and economic analysis. Yale University Press, New Haven
- Calabresi G (2014) A broader view of the cathedral: the significance of the liability rule, correcting a misapprehension. *Law Contemp Probl* 77(2):1–14
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Denoza, F. (2013), ‘il modello dell’analisi economica del diritto: come si spiega il tanto successo di una tanto debole teoria?’, 11(2):*Ars Interpretandi*, 43–67
- Kalman L (2014) Some thoughts on Yale and Guido. *Law Contemp Probl* 77(2):15–43
- Marciano A (2012) Guido Calabresi’s economic analysis of law, Coase and the Coase theorem. *Int Rev Law Econ* 32:110–118
- Marciano A, Ramello GB (2014) Consent, choice and Guido Calabresi’s Heterodox economic analysis of law. *Law Contemp Probl* 77(2):97–116
- Mattei, U. (1994), ‘Efficiency in Legal transplants: an Essay in Comparative Law and Economics’, *International Review of Law and Economics*, 14:13–19
- Posner RA (1972) *Economic analysis of law*. Little, Brown and Co., New York
- Michelman F (1971) Pollution as a tort: a non-accidental perspective on Calabresi’s costs. *Yale Law J* 80:647
- Ramello GB (2016) The past, present and future of comparative law and economics. In: Ramello G (ed) *Comparative law and economics*. Edward Elgar Publishing, Cheltenham, pp 3–22

Cameralism

Andre Wakefield

Pitzer College, Claremont Colleges, Claremont, CA, USA

Definition

Cameralism was an aspiring profession during the seventeenth and eighteenth centuries; it thrived in the small territories of the Holy Roman Empire. Academic cameralists, using law and medicine as their models, constructed a system of auxiliary sciences – largely natural, economic, and technological sciences – to support the training of future state servants in the German lands. This system of professional knowledge, known as the cameral sciences, was taught at German universities during the eighteenth century. As a professional model, cameralism ultimately lost out to jurisprudence, but the discourse that it spawned extended well beyond the German lands into Austria-Hungary, Scandinavia, and the Italian states.

Introduction

Historians of economic thought often treat their discipline like physics or chemistry, which is to say, they regard it as a positive science. In this, they follow Milton Friedman. “Economics as a positive science,” he famously argued, “is a body of tentatively accepted generalizations about economic phenomena that can be used to predict the consequences of changes in circumstances” (Friedman 1953). To regard economics as a positive science has implications for the way we write its history. Chemistry had its phlogiston; astronomy had its Ptolemaic system; and economics had its preclassical period. Lavoisier, Copernicus, and Adam Smith play the heroes, relegating older theories to the dustbin of history. The narrative of positive science has, for quite some time, motivated dictionaries and encyclopedias like this one. As the great Inglis Palgrave put it, such compendia “show what has actually been written

in former times, and hence will enable the reader to trace the progress of economic thought” (Palgrave 1987).

In a world where Adam Smith and his intellectual progeny play the heroes, it becomes clear what is left for English mercantilists, German cameralists, and other “backward” theorists: they play the foils against which the stories of disciplinary progress get written. H. C. Recktenwald’s entry in the *New Palgrave* did just that. “Analytical economics, insights into the laws of the market and the study of the interaction between market and state,” he explained, “are relatively unknown in the simple textbooks of the cameralists, which show otherwise sound common sense” (Recktenwald 1987).

Cameralism has been variously defined as a German variant of mercantilism, a university science, a theory of government and society, a baroque science, a political science, an early modern economic theory, and an administrative technology (Roscher 1874; Tribe 1988; Small 1909; Lindenfeld 1997; Schumpeter 1954). Karl Marx just called it a “silly mish-mash of notions inflicted on aspiring bureaucrats” (Marx and Engels 1961–1974). There is some truth in all of it. The dominant narrative, however, has long treated cameralism as a subset of English mercantilism. As Recktenwald put it, cameralism “is the specific version of mercantilism, taught and practised in the German principalities (Kleinstaaten) in the 17th and 18th centuries” (Recktenwald 1987). Nineteenth- and early twentieth-century writers discovered broad lines of agreement between mercantilism and cameralism (Roscher 1874). Certainly, cameralism shared important family resemblances with mercantilism – a commitment to statebuilding, in both political and economic terms, through policies such as import substitution and the industrialization of raw materials (Reinert 2005). Cameralism has also been analyzed as an important early model for and an inspiration for alternative approaches to public finance (Backhaus and Wagner 2004).

But there have been dissonant voices along the way. Writing in 1909, the American sociologist Albion Small suggested that historians of economics had mischaracterized German cameralists.

“Cameralism,” he argued, “was an administrative technology. It was not an inquiry into the abstract principles of wealth, in the Smithian sense” (Small 1909). Small had a good point because cameralists wrote a lot and most of it did not involve what we would call “economics.” Magdalena Humpert’s bibliography of cameralist literature included more than 14,000 printed sources (Humpert 1937). Not many of those pages (numbering in their millions) included general discussions about balance of trade or bullionism. Open a cameralist text and you will be more likely to find chapters describing lead smelting, gardening, brewing beer, raising pigs, forestry, and hard-rock mining than, say, general principles of trade. These particulars have most often been ignored in accounts of cameralism as a political or economic theory, but they highlight an important fact: cameralism owed much to the *Kammer*, or fiscal chamber, a specialized collegial body dedicated to administering the sovereign finances (Zielenziger 1914).

More recently, Keith Tribe redefined cameralism as “a university science,” placing great weight on the context and practice of university instruction and rejecting the significance of administrative practice for the production of cameralist texts. Instead, Tribe looked to the context of pedagogy and discursive formations. “The two prime influences on these texts,” he argued, “were the actual teaching situation and the Wolffian philosophy which informed their style and was itself very largely a product of pedagogic practice” (Tribe 1988). Of paramount importance, then, were the “discursive conditions” under which the texts were produced. Tribe’s approach represented a radical departure from earlier scholarship on the subject for he treated cameralism as a self-contained academic discourse, separating the production of the cameral sciences completely from the context of fiscal administration. He also greatly expanded the traditional canon of cameralism by examining hundreds of cameralist texts, many of which were used in university instruction.

Tribe’s intervention, in turn, prompted other scholars to think more systematically about the relationship between the cameral sciences and

administrative practice. Perhaps, as Tribe argued, there was no necessary relationship between the two; or perhaps, as Small and others had long maintained, the cameral sciences reflected administrative practice. For some of the more prominent cameralist authors, however, it turns out that there was a relationship between discourse and administrative practice, though not a transparent one. The cameral sciences, that is, did not simply *reflect* everyday practice in the bureaus. Rather, published cameralist texts – canonical works by Veit Ludwig von Seckendorff and Johann von Justi among them – were characterized by a pragmatic utopianism that painted the world as it should be, even as they purported to describe the world as it was. The increasing tendency to treat cameralism as a unique historical formation has challenged the historiographical tradition that relegated the cameral sciences to marginal status in the history of economic thought (Sandl 1999; Tribe 1988; Wakefield 2009).

Historical Development

By the seventeenth century most German territories, large and small, had developed *Kammer* to manage the intimate affairs of princes, dukes, kings, and emperors (Heß 1962; Klinkenborg 1915). By the second half of the seventeenth century, members of the *Kammer* began to be recognized as a distinct group. People started calling them cameralists. Every responsible fiscal official was expected to know his way around a mine or a barley field, because those were the appropriate “ordinary” sources of revenue for his prince, such as income from the mines. (“Extraordinary” sources of revenue, such as direct taxation in times of crisis, were seen as illegitimate and even despotic in many German territories.) Cameralism was structured by the material and institutional realities of fiscal administration in the territories of the Holy Roman Empire.

In the wake of the Thirty Years War, the German lands of the Holy Roman Empire were a mess, devastated, and depopulated. The Peace of Westphalia (1648) recognized more than 300 sovereign territories, ranging widely in size, wealth,

and power. For the next 200 years, the Empire served, in the words of Mack Walker, as an “incubator,” protecting smaller territories against aggressive incursions from more powerful neighbors (Walker 1971). The economic and political structure of the Empire at once protected and limited the states within it. Cameralists had to accept these limitations, as the ruler of each territory became a kind of entrepreneur seeking to profit from the natural and human resources in his territory.

Insofar as cameralists sought to systematize the daily work of fiscal administration, they faced great obstacles, because the logic of every *Kammer* was distinct, attuned to the local resources of a particular territory or region. The Holy Roman Empire, with its hundreds of kingdoms, duchies, principalities, and bishoprics, presented a staggering diversity of administrative structures, geography, and economic activities. Accordingly, cameralists filled their books with endless detail about the territories in which they lived and worked. This has led authors to suggest that the cameral sciences were descriptive sciences, models of “practical reasoning” that avoided the utopian thinking of nineteenth-century economics (Lindenfeld 1997). It was not, however, always that straightforward. Sometimes, utopian thinking masqueraded as practical, utilitarian knowledge. Cameralists liked to publish “practical” treatises about how to brew beer or raise cattle, for example, and they often made it sound easy. But practical success in agriculture or manufacturing was never easy, which is why failure was the rule when it came to new state ventures. In this respect, cameralists were utopian pragmatists, imagining fields full of healthy crops and fat cows, even as the people drank miserable beer and struggled to feed themselves.

In 1727 Frederick William I of Prussia established the first academic chairs in cameralism, resolving to initiate lectures on “*Cameralia, Oeconomica* and *Polizeisachen*” at his universities in Halle and Frankfurt an der Oder (Stieda 1906; Schrader 1894). “To that end,” declared a cabinet order from Berlin, the king had decided to establish a “special Profession, so that students could acquire a good foundation in

these sciences before they are employed in state service.” The first “professor of *Cameralia*” was Simon Peter Gasser, a Prussian War and Domains Councilor. Students at the University of Halle were encouraged to attend his lectures, and those who received good recommendations from Gasser could expect special consideration when the time came to appoint new officials. The authorities in Berlin sketched an outline of topics for Gasser’s lectures. Frederick William’s new “profession” of cameralism demanded sweeping knowledge of Prussia’s material circumstances, its productive potential, and the complicated landscape of its rights and privileges.

Cameralism as Profession

After the formal establishment of academic cameralism in 1727, cameralists throughout the Holy Roman Empire sensed an opportunity to establish themselves professionally. We should not imagine these men as members of a political economic school, like the physiocrats, or as some early modern version of the Chicago School of Economics. Cameralist reformers had bigger dreams. They imagined their subject not as a discipline, such as history or mathematics, but as an entirely new academic profession. Cameralism, in other words, would be modeled on law and medicine. Johann von Justi, the most prominent of cameralist proselytizers, was very clear about this in his groundbreaking 1755 cameralist textbook, *Staatswirtschaft*. He suggested that the existing professional faculties of theology, law, and medicine be supplemented by a cameralist faculty (Tribe 1988). The new professors would need to be skilled in areas ranging from forestry and manufactures to taxes and chemistry. “The professor of chemistry would be chosen so that he could lecture on assaying and smelting, and not just the preparation of medicines. . . the teacher of mechanics would be able to lecture on mining machinery, and the professor of *Naturkunde* would need adequate knowledge about the essence of ores and of deposits.” There would be six professors in all, “to which one might add a teacher of civil and military

engineering.” Not only would this new faculty train skilled future officials, but it would offer “advice for the many institutions and undertakings of the state, for which one must often turn to foreigners at great expense” (Justi 1755a).

Behind the recitation of cameralist principles and material detail in hundreds of textbooks, then, there was a roiling debate about what it meant to be a cameralist. It was a struggle not so much over abstract principles of wealth creation as it was over professional identity. When “aspiring cameralists” flocked to places like Göttingen and Lautern to hear lectures in the cameral sciences, they were not just studying economic policies and the principles of good police, but they were also learning how to behave as members of the *Kammer*. It was not enough to know about budgeting and accounting, one had to be fashionable as well. The classic markers of scholarly culture – knowledge of Latin, learned disputation, reference to authoritative sources, and reading the textbook from a lectern – were rejected in favor of more gentlemanly approaches. Justi, when he arrived in Göttingen, was very specific about this. “I have employed a special teaching style in my courses.” He would, contrary to common practice, lecture for only thirty minutes. “Then I got down from my lectern and, standing together with my listeners, I spent the rest of the hour in free and sociable conversation about the lecture material” (Justi 1755b).

It would be a mistake, therefore, to view the cameral sciences as nothing more than a set of political economic principles. For Justi, as for other cameralist reformers, the new profession represented a new way of life and a new epistemology. One could not simply teach students to balance the books and learn about revenue sources; equally important was the need to behave like a proper servant of the *Kammer*. One had to know how gentlemen acted at court, how to make polite conversation, and how to avoid being a tedious pedant. Cameralism was thus a perfect fit with the new model universities of eighteenth-century Germany, notably Halle and Göttingen (two centers of the cameral sciences). Gerlach Adolf von Münchhausen, Hanoverian minister and first curator of the University of Göttingen,

sought from the very beginning to attract young noblemen and wealthy students to his university. Like the ideal classroom of Justi's reveries, Münchhausen's university aimed to attract fashionable and wealthy students. Münchhausen focused on building nice streets, coffee houses, and impressive academic buildings as a way to attract the right kind of student. For him, utility meant the ability to attract wealthy students to Göttingen from around the Holy Roman Empire, and even from England. For this, one would need famous professors and fashionable knowledge. From this perspective the cameral sciences, fashionable sciences designed to appeal to wealthy noblemen, were perfect. Münchhausen brought Justi to Göttingen in 1755, the same year in which his *Staatswirtschaft*, the most influential of cameralist textbooks, appeared in print (Wakefield 2009).

Cameralism, as imagined by Justi and Münchhausen, was not a stand-alone science like economics or sociology; it was a system of professional education. During the latter half of the eighteenth century, reformers worked to create cameralist faculties throughout the lands of the Holy Roman Empire. Cameralist reformers managed to alter university curricula, establish new academies, and found separate university faculties (Stieda 1906; Tribe 1988; Klippel 1995). In many cases, they even instituted examinations and succeeded in making access to coveted state offices contingent on academic study of the cameral sciences (Bleek 1972). In Göttingen, Münchhausen worked for decades to build a system of "auxiliary sciences" that would create the structure necessary for a cameralist faculty. This involved, most of all, building a system of natural sciences that could serve to train aspiring cameralists. Münchhausen ran into trouble with the other higher faculties – notably law and medicine – in his efforts to harness auxiliary sciences in the service of cameralism. Eventually, though, Münchhausen built a system of sciences, ranging from "economic botany" to "technology," which served as auxiliary sciences to train future state servants (Wakefield 2009).

Göttingen was not alone. In Lautern, 200 miles to the southwest, Friedrich Casimir Medicus

founded a freestanding cameralist academy (*Kameral-Hohe-Schule*). Lautern represents the mature example of a professionalizing cameralist curriculum. Hoping to avoid the stubborn traditional faculties and their privileges, Medicus decided to sidestep them altogether by appointing permanent professors to teach subjects such as chemistry, economic botany, technology, and agriculture. For Medicus, it was crucial to have one or two professors dedicated entirely to the natural sciences, because they were the "true foundation upon which all the knowledge of the future state administrator rests, and without which he will never be able to make one sure step forward. One must firmly guarantee that no young man who has failed to study these with zeal is allowed to pass on to the Science of Sources" (Wakefield 2009). Lautern's focus on cultivating the "source sciences" made sense as a strategy for developing the many small and landlocked territories of the Holy Roman Empire. For these principalities and duchies, the constant, intensive improvement of very limited territorial domains – what Sophus Reinert has called "*ersatz imperialism*" – proved more appealing than on the restless, expansionist ambitions of colonial enterprises (Reinert 2011).

Conclusion and Future Directions

The historiography on cameralism stretches back at least two centuries and, as could be expected, that literature records many shifts in approach, definition, and methodology. There is, in other words, no single agreed-upon definition of cameralism; instead, we have a shifting and multifaceted debate about the nature of cameralism and its significance. The most widespread approach to cameralism treats it as a variety of mercantilism, taught and practiced in the German lands of the eighteenth century. Others have treated cameralism as an administrative technology, specifically adapted to the small German territories of the Holy Roman Empire. Still others have defined it as a university science, subject to the pedagogical and discursive conditions present in eighteenth-century German academic settings.

It has also been analyzed as a literature that functioned as public relations for the early modern fiscal policy states of central Europe.

The lack of general agreement about what, exactly, cameralism was (or was not) provides fertile ground for further research. There is much to be done. Recent work has tended to emphasize that cameralist discourse was not limited to the German lands of the Holy Roman Empire, the traditional focus of analysis. Rather, studies in the circulation and translation of texts have revealed that cameralism reached far beyond the German lands, and it was cameralist discourse, not English mercantilism or Smithian political economy, that enjoyed the widest circulation across large swaths of central Europe, Scandinavia, and Italy (Reinert 2011; Lluch 1997). Another burgeoning area of research connects cameralism to technology and the natural sciences. Long seen as peripheral to the “essence” of cameralist discourse, the natural sciences – especially Linnean natural history, chemistry, and mining sciences – have gained increasing attention as core parts of the cameralist enterprise (Koerner 1999, Smith 1994, Wakefield 2000).

References

- Backhaus J, Wagner R (2004) Society, state, and public finance: setting the analytical stage. In: Backhaus J, Wagner R (eds) *Handbook of public finance*. Kluwer, Boston
- Bleek W (1972) *Von der Kameralausbildung zum Juristenprivileg*. Colloquium Verlag, Berlin
- Friedman F (1953) *Essays in positive economics*. University of Chicago Press, Chicago
- Heß U (1962) *Geheimer Rat und Kabinett in den ernestinischen Staaten Thüringens*. Weimar, Böhlau
- Humpert M (1937) *Bibliographie der Kameralwissenschaften*. Köln, Schroeder
- Justi JHG (1755a) *Staatswirthschaft*, 2 vols. Breitkopf, Leipzig
- Justi JHG (1755b) *Abhandlung von den Mitteln die Erkenntniß in den Oeconomischen und Cameral-Wissenschaften dem gemeinen Wesen recht nützlich zu machen*. Göttingen
- Klinkenborg M (1915) Die kurfürstliche Kammer und die Begründung des Geheimen Rats in Brandenburg. *Hist Z* 114(1915):437–488
- Klippel D (1995) Johann August Schlettwein and the economic faculty at the University of Gießen. In: Bernard Delmas B, Demals T, Steiner P (eds) *La diffusion internationale de la Physiocratie, XVIIIe-XIXe*. Grenoble: Presses Universitaires, 1995, pp. 345–365
- Koerner L (1999) *Linnaeus: nature and nation*. Harvard University Press, Cambridge, MA
- Lindenfeld D (1997) *The practical imagination: the German sciences of state in the nineteenth century*. University of Chicago Press, Chicago
- Lluch E (1997) Cameralism beyond the germanic world. *Hist Econ Id* 5:85–99
- Marx K, Engels F (1961–1974). *Karl Marx, Friedrich Engels. Werke*. 39 vols. Dietz, Berlin
- Palgrave RHI (1987) Introduction to the first edition. In: Eatwell J, Milgate M, Newman P (eds) *The New Palgrave: a dictionary of economics*, 3 vols, vol 1. Macmillan, London, p xi
- Recktenwald HC (1987) Cameralism. In: Eatwell J, Milgate M, Newman P (eds) *The New Palgrave: a dictionary of economics*, 3 vols, vol 1. Macmillan, London, pp 313–314
- Reinert E (2005) A brief introduction to Veit Ludwig von Seckendorff (1626–1692). *Eur J Law Econ* 19:221–230
- Reinert S (2011) *Translating empire: emulation and the origins of political economy*. Harvard University Press, Cambridge
- Roscher W (1874) *Geschichte der Nationalökonomik in Deutschland*. Oldenburg, Munich
- Sandl M (1999) *Ökonomie des Raumes. Der kameralwissenschaftliche Entwurf der Staatswirthschaft im 18. Jahrhundert*. Böhlau, Köln
- Schumpeter J (1954) *History of economic analysis*. Oxford University Press, New York
- Schrader W (1894) *Geschichte der Friedrichs-Universität zu Halle*. 2 vols. Berlin: Dümmler
- Small A (1909) *The cameralists, the pioneers of German social polity*. University of Chicago Press, Chicago
- Smith P (1994) *The business of alchemy: science and culture in the Holy Roman Empire*. Princeton: Princeton University Press
- Stieda W (1906) *Die Nationalökonomie als Universitätswissenschaft*. Teubner, Leipzig
- Tribe K (1988) *Governing economy. The reformation of german economic discourse, 1750–1840*. Cambridge University Press, Cambridge
- Wakefield A (2000) Police chemistry. *Sci Context* 13:231–267
- Wakefield A (2009) *The disordered police state: german cameralism as science and practice*. University of Chicago Press, Chicago
- Walker M (1971) *German home towns: community, state, and general estate 1648–1871*. Cornell University Press, Ithaca
- Zielenziger K (1914) *Die alten deutschen Kameralisten*. Fischer, Jena

Further Reading

- Dittrich E (1974) *Die deutschen und österreichischen Kameralisten*. Wissenschaftliche Buchgesellschaft, Darmstadt

- Nielsen A (1911) Die Entstehung der deutschen Kameralwissenschaft im 17. Jahrhundert. Fischer, Jena
- Rosenberg H (1958) Bureaucracy, aristocracy, and autocracy. The prussian experience 1660–1815. Harvard University Press, Cambridge
- Sommer L (1920–25) Die Österreichischen Kameralisten. 2 vols. Konegen, Vienna
- Troitzsch T (1966) Ansätze technologischen Denkens bei den Kameralisten des 17. und 18. Jahrhunderts. Duncker & Humblot, Berlin

Additionally, a historical phase of society's development is called capitalism.

Introduction

In general, the economic systems could be described according to coordination mechanism – which can be market or plan coordination and according to ownership, which can be state or private ownership. Each society has a set of production factors that – independent from the social and economic order – should be used for an optimal production and allocation of goods. The main task of any economic and social system is to match production and the preferences of its members. Therefore, it is relevant to decide which goods will be produced and how they will be allocated. To manage this allocation and production decisions, many approaches have been tried in history. There were and still are feudalistic systems without industrial production and an autocratic government that rules economy regarding its own advantage. Besides that there were and still are socialistic systems that used central planning of production to solve the allocation production problems. In addition to the feudal, socialist (or communistic) way of using of the production factors, capitalism is another possible system of providing goods and is formed as a combination of private ownership and market coordination mechanism. Capitalism is described by its elements.

Cap-and-Trade

► Emissions Trading

Cap-and-Trade Approach

► Transferable Discharge Permits

Capitalism

Sylvia Ručinská¹, Ronny Müller¹ and Jannik A. Nauwerth²

¹Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

²Faculty of Business and Economics, University of Technology Dresden, Dresden, Germany

Abstract

Capitalism is a social and economical system which applies the use of production factors of the economy as a whole. Capitalism can also be understood as a manufacturing process, which is focused on profit maximization and is based on the principle of the invisible hand. Capitalism is associated with the private production of goods and the individual benefit of surplus. These attributes differ from other systems, especially from socialism or communism. Differences occur in the explanation of capitalism, depending on the socioeconomic origin of the explanation.

Elements of Capitalism

Capitalism can be described by the following typical elements: private property, freedom, free markets and free competition, private capital and profits, and free prices and wages (Darcy 1970).

The basic of any capitalistic economy is **private property and ownership**. The production factors are not owned by state, and individuals should decide, with respect to their benefit, whether to demand or supply goods. Therefore, individuals have to determine their offered work

force and their accumulation of capital. It is assumed that any individual knows best about their preferences and is able to satisfy them most suitably. Besides that firms produce goods and individuals benefit from the profits. Firms demand labor force and capital to facilitate production and innovations.

In contrast to central planned economies there is no need of a general aggregation of preferences. Therefore, capitalistic economies are based on the sum of decentralized decisions made by individuals. Thus an important characteristic of capitalism is individualism and individual satisfaction of needs.

Other main requirements for capitalistic economies are **free markets** and **free competition**, to achieve efficiency and entrepreneurship. Capitalism is based on the functioning of market mechanism, which is built on the principle of the invisible hand. The principle was first introduced by A. Smith, who stated that individualism and meeting individual prospects by every individual lead to efficiency and benefit the whole economy (Smith 2012). Free competition without restriction by access to information, market access, market structure, and interventions of the government leads the firms to invest in new technology, innovation, and new skills (Hodgson 2003). Also prices and wages are determined by the free market. This idea also assumes that markets are well functioning and a significant state intervention isn't needed. Precondition for free market and private ownership is **freedom**. For some authoress, this issue is important enough that they equate capitalism and a market system (Lindblom 1980).

In capitalism, societal and social conditions are mainly influenced by **capital**. This is due to its high mobility and its universal applicability. Moreover, it is possible to accumulate it unlimited in contrary to labor force. Individuals with a negligible amount of capital can only influence production decisions by consuming.

This leads to a main difficulty of capitalism. The market result is highly influenced by initial endowment of an individual. On the one hand this is essential for an individualistic society, on the other hand this can cause unjustified market

outcome. Therefore, politics are addressed to avoid societal not accepted market results.

Irrespective of the characteristics above, there is no general definition of capitalism. The explanations of capitalism are highly influenced by ideological background and target group.

Types of Capitalism

First steps of capitalism's development can be found in the period, which is associated with the development of manufacture production, although simple forms of the so-called protocapitalistic trading appeared already in 3300 BC (Mischer 2014). The exact year of its beginning cannot be specified, but we can subdivide the development of capitalism into three stages: early capitalism, industrial capitalism, and late capitalism (Sombart 2001, 2011). **The early capitalism** includes the time from the sixteenth century to the beginning of industrial revolution. This period is characterized by absolutist monarchs and mercantilism. Besides that, France is the world's leading economy. Due to the political absolutism, economy was widely regulated. To avoid another absolutist monarch benefiting from foreign trade, exporting goods was favored over importing. In the absence of foreign trade and a dynamic system, economic growth was negligible. The production was mainly concentrated on agricultural goods, embedded in a rigid and hierarchical feudalism. A well-known theory describing this period by Malthus (2007) connects the absence of economic growth with population growth. The so-called Malthusian trap states that any improvement in technology results in higher population, rather than improving the economic situation. Despite this rigid environment, first small steps were undertaken to improve economy. In this regard, international trade and banking became more important, and besides that, manufactories arose and thereby announced the industrial revolution.

With the beginning of Industrialization, the age of **industrial capitalism** began and the United Kingdom took over the world economic leadership. Determining an exact beginning depends

on point of view; however, the invention of the first spinning machines in the United Kingdom depicted a cornerstone. After that, industrial development spread over continental Europe before attaching America. One can observe multitudes of social and economic changes in this time. Obviously there were technological improvements, for example, the steam engine, industrial use of electricity, and chemical industry. Thus, by replacing agriculture with industrial production, a structural change in economy occurred. As a consequence, the production factors labor and capital became more important and the factor land became less important. Moreover, people moved from rural areas to cities, in particular towards coastal regions. This was due to a higher productivity of labor in industry production than in agriculture. Selling labor force to a company promised more than working at subsistence farming. In conjunction with the urbanization, a decrease in death and birth rate took place. Beyond this, the institutional background changed to the benefit of the people; especially democratic processes were initiated. Furthermore, public goods as schooling and security were implemented step by step. Moreover, property rights in material and intellectual dimensions were implemented. According to the structure of economies and political systems, many shapes of capitalism took place.

The late capitalism began at the end of the nineteenth century and the USA became the leading economy. Former small firms developed to large companies by taking over competitors and expansions. Thus, the number of monopolies and cartels rose. In addition, the linkages between companies increased in dimensions of business connections and ownership structure. Moreover, large banks were founded and financial markets became more important in corporate finance. Parallel to this, economic crisis took place in recurring periods. International trade was based on the so-called gold standard. This means that a currency rested on a fixed amount of gold. Thus, everyone could convert paper money in gold. Therefore, importing led to an outflow and exporting to an inflow of gold. According to that, the price level was driven by foreign trade.

This leads to the so-called impossible trinity in economic theory. It declares that a stable foreign exchange rate, free capital movement, and an independent monetary policy cannot be achieved at the same time. In recent past, there were approaches to ease this impossibility, but they were of limited duration.

At the beginning of the twentieth century, world economy was shocked by the two World Wars and the "Great Depression." As a result of these shocks and uncertainty about the future, capitalism produced market results that were not commonly accepted. In some countries socialist systems arose, and the iron curtain took place. The capitalistic economies changed into more regulated systems with frequent interventions. However, they managed to preserve the main mechanics of capitalism. The so-called New Deal implemented by Roosevelt in the 1930s depicted the beginning of market regulation. The plan's aim was to increase purchasing power and restore confidence in economy and included gains in public expenditure. After the World War II, the economies of Western Europe reconstructed their political system by focusing on welfare. To avoid unemployment, public pension schemes and redistribution of money became part of the systems. Moreover, the Asian countries, especially China, improved industrial production. By remembering the expansive foreign trade in gold standard, a modern adaption was implemented in 1944, the "Bretton-Woods-System." This was intended to raise the efficiency of international trade and thus improve economic welfare. To achieve fixed but flexible foreign exchange rates, the US dollar set as anchor currency that was bound to gold. The other members had to follow the monetary police of the USA to fix the foreign exchange within narrow bounds. But this approach crashed in 1970s followed by regional currency snakes; sometimes this period is referred to as embedded liberalism. With the end of "Bretton-Woods," political policy tended to liberalization of economies. This includes a slow reduction of subsidies, lower barriers in trade, and reducing market regulations especially in financial markets.

Theoretical View on Capitalism

The modern capitalism based on the “classical” economy represented by Adam Smith and his “Invisible Hand” (Smith 2012), which is used to explain that pursuing own interest promotes the society. Every individual works for their own consumer needs. However, other people are benefiting from this, because they can buy products with higher quality and quantity. Thus, society is promoted by people regarding their own interest and morality. To use full potential from trade he declared the necessity of free markets and the absence of monopolies and cartels.

Karl Marx developed an opposing theory of capitalism (Callinicos A 2012). He established the so-called Historical Materialism to consider society and economy. It states that social and economic change was not driven by ideas but material foundation of people. According to this, human development is a consequence of material equipment and the mode of production. It decomposes history in five parts. These are primitive communism, slave society, feudalism, capitalism, and socialism/communism. To climb up this scale, it is necessary that mode of production is no longer appropriate. The working masses that do not participate in wealth overcome the regime and implement a new system with suitable mode of production and governmental system. This will repeat until communism is reached. A capitalistic society is split in two opposing parts. On the one hand there are owners of capital and on the other hand the working people. Capitalists maximize their income and workers have to sell their working power to them. Marx assumed capitalists selling goods at market price, but workers income is at subsistence level anyway. According to that, capitalists receive the surplus value, and in a competitive environment capitalists have to reinvest it in production. But depending on subsistence wages, the workers do not have enough money to buy all products. Therefore, crises arise and clear the market from overproduction. Thus, Marx predicted that workers, oppressed by frequent crises and increasing social inequality, will overcome capitalism.

Max Weber used a sociological approach to discuss capitalism. He considered capitalism and its characteristics in the Western civilizations (Weber 2008). He works out that rational behavior of individuals increases, but furthermore social acting exists. In addition, he describes that capitalism is driven by expectations over future income from trade. For him, a main character of occidental capitalism is Protestantism. He states that especially Lutheranism and Calvinism lead to another work ethic. In this opinion, this ethic leads to a religious foundation of higher work effort and diligence. Max Weber introduced the concept of “political capitalism” (Holcombe 2015).

Joseph Schumpeter 2008 assumed that capitalism is not a permanent order in economics (Schumpeter 2008). He argued that improvements in efficiency and innovation will lead to monopolies and cartels. As a result from this concentration of economic power, he predicted the destruction of the societal order capitalism is based on. Therefore, similar to Marx he assumed capitalism as not permanent.

The British economist of the twentieth century John Maynard Keynes assumed that the demand is the crucial variable in an economy (Keynes 1997). His theory based on the idea that capitalism is not able to maintain itself. Economic crisis would destroy it if the government does not intervene. In conclusion, Keynes promoted a “strong state” and countercyclical interventions supporting demand, to reduce effects of crises.

Austrian economists Friedrich August von Hayek and Ludwig von Mises partially to Keynes’s theory advocated market economy. Hayek assumed that any public administration is inefficient and expensive. Therefore, he rejected a government policy based on interventions and refused any subsidies to firms or demand. Hayek promoted a “slim state” without interventions, in opposite to Keynes.

Keynes and Hayek were proponents of capitalism and both refused any socialist or feudal economic system. Nowadays, modified versions of their theories are frequently used in economic discussions.

In addition to the Keynesian's ideas, the monetary school of thought with Milton Friedman was established in the twentieth century. In his opinion, government should provide property rights, competition, monetary constitution, and support for underage and depended people (Friedman 2002). Opposing to Keynes he states a natural rate of unemployment based on structural frictions in labor markets. Besides that, he states a narrow connection between the quantity of money and inflation. Thus, an appropriate central banking could dispose the problems of inflation and deflation (Friedman 2005) and stabilize economy. To reduce unemployment in the long run he recommended reforms that reduce those structural frictions (Friedman 1968). Friedman suggested a "slim state" that drives markets only by monetary policy. Especially the idea of widening the monetary supply in financial crisis goes back to him.

In the past and also nowadays there are discussions if capitalism as a system could be stabile. The reason of capitalism's crisis is its very nature. Orientation on profit and surplus value means a concentration of money in the hands of a small group of people and also a dissatisfaction of a big group of people, what has to lead to riots and change of the system.

Conclusion/Future Perspective

The undisputed advantages of capitalistic economy are its high productivity, the efficiency, and its adaptability. In the age of capitalism, a level of wealth has been achieved which had never been reached before.

However, there are major challenges that need to be resolved in the future. There are serious problems in environmental pollution due to globalization (Bhagwati 2007). Moreover, there are difficulties in distribution of wealth (Piketty 2014). Besides that, the structure of welfare states is discussed widely with respect to downsize welfare systems. Additionally, durable unemployment is a problem especially in Europe.

References

Monographs

- Bhagwati J (2007) In defense of globalization. Oxford University Press, Oxford
- Callinicos A (2012) The revolutionary ideas of Karl Marx. Haymarket Books, Chicago
- Darcy RL (1970) Primer on social economics. Colorado State University, Colorado
- Friedman M (2002) Capitalism and freedom: fortieth anniversary edition. University of Chicago Press, Chicago
- Friedman M (2005) The optimum quantity of money. Aldine Transaction, Piscataway
- von Hayek FA (1967) Prices and production. Kelley (Augustus M.) Publishers, New York
- Keynes JM (1997) The general theory of employment, interest and money. Prometheus Books, Amherst (New York)
- Lindblom C (1980) Jenseits von Markt und Staat. Eine Kritik der politischen und ökonomischen systeme. Klett-Cotta, Stuttgart
- Malthus TR (2007) An essay on the principle of population. Dover Publications, Mineola (New York)
- Piketty T (2014) Capital in the twenty-first century. Belknap, Cambridge
- Schumpeter J (2008) Capitalism, socialism and democracy. Harper Perennial Modern Classics, New York
- Smith A (2010) The theory of moral sentiments. Digireads Publishing, New York
- Smith A (2012) The wealth of nations. Wordsworth Editions, Hertfordshire
- Sombart W (2001) Der moderne Kapitalismus, vol 1. Adegi Graphics LLC, New York
- Sombart W (2011) Der moderne Kapitalismus, vol 2. Adegi Graphics LLC, New York
- Weber M (2008) The protestant ethic and the spirit of capitalism. Digireads Publishing, New York

Journals

- Friedman M (1968) The role of monetary policy. Am Econ Rev 58(1):1–17
- Hodgson GM (2003) Capitalism, complexity, and inequality. J Econ Issues XXXVII(2):471–478
- Holcombe RG (2015) Political capitalism. Cato J 35(1)
- Mischer O (2014) Geschichte einer Wirtschaftsordnung. GeoEpoche der Kapitalismus 69:160–167

Caribbean Piracy

► Piracy, Old Maritime

Cartels and Collusion

Bruce Wardhaugh

School of Law, Queen's University Belfast,
Belfast, UK

Abstract

This entry provides an introductory account of cartels and collusion and the means used by European and American law to control such practices. The welfare-reducing and welfare-enhancing features of these cartel and other cartel-type arrangements are discussed to demonstrate the need for considered regulation. Both horizontal and vertical arrangements are analyzed, given their different uses and effects in the economy. However, as some forms of collusive activity are welfare enhancing, these are discussed in an effort to show why regulating such behavior must be done with care. Criminal, administrative, and private sanctions are compared as means of control of such agreements. Other topics briefly discussed the nature of (legally) permitted and prohibited collusive information exchange and noneconomic concerns which may justify collusive behavior.

Introduction

JEL codes: K.21, L.40, L.41, L.42

Adam Smith famously wrote, “People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public or in some contrivance to raise prices” (Smith 1776, Bk I, ch x). By this observation, Smith concisely identifies the essence of collusive activity leading to a cartel, which results in an agreement to raise (or fix) prices and harm the consumer in so doing. However, cartel activity can take forms other than naked price-fixing, these include output restrictions, customer allocation (including bid-rigging agreements), and geographic exclusivity of

operations. Whatever practice of this sort the parties choose, the practice can give the parties a degree of monopoly power over their customers (Neils et al. 2011, p. 288) and thus isolate the parties from the competitive rigors of the market. Indeed, the European Court of Justice in *ICI* (at para. 64) has referred to this type of activity as “knowingly substitute[ing] practical cooperation for the risks of competition.”

Collusive activities can take place among ostensive competitors at the same level of the supply chain (horizontal arrangements or cartels), at different levels of the supply chain (vertical arrangements or cartels), and among members at the same level of the supply chain with the information necessary to collude passed through a member (typically a distributor or wholesaler) at a different level in the supply chain (these arrangements are known as “hub-and-spoke” or “A-B-C” cartels). Yet not all forms of collusion among competitors are regarded as harmful, indeed certain agreements may be beneficial to consumers. Examples of such activities include: joint ventures (particularly those involving research and development to take advantage of synergies of the participants who each may possess specialized knowledge or access to intellectual property) (Motta 2004, pp. 202–205; Neils et al. 2011, pp. 295–297), cooperative standard setting to ensure that customers can take advantage of network effects (Motta 2004, p. 207), and joint purchasing agreements to take advantage of quantities of scale. But at the same time, if the parties to such an agreement possess sufficient market power, such arrangements can lead to anti-competitive effects (Neils et al. 2011, p. 293).

This entry briefly summarizes the results of some of the vast literature on cartels and other forms of economic collusion. More complete bibliographies of this literature can be found elsewhere (see, e.g., Motta 2004; Neils et al. 2011; Kaplow 2013; Wardhaugh 2014). The entry will first discuss the nature of harm which is commonly ascribed to such economic collusion. The discussion of harm ends by outlining the means by which these sorts of activities are controlled or sanctioned. The entry next turns to a discussion

of the types of cartels (horizontal, vertical, hub and spoke) which one sees in today's marketplace. As the effects of these types of agreements can be different, the varying means by which various legal regimes treat such agreements are considered. The entry next examines the features of industries in which cartelization occurs. The fifth section of the entry briefly discusses collusion and information exchange and explains the economic and legal issues. The final section of the entry considers noneconomic considerations which may be taken into account in the evaluation of collusion.

Harms from Cartels

In a competitive market, goods are sold at the marginal cost of their production. The intersection of the cost-and-demand curves therefore determines the quantity produced and the price of the good. As the analysis of cartels is for all intents and purposes identical to the analysis of monopolies (Faull and Nikpay 2014, pp. 21–22), the insights learned from analysis of monopoly power are readily transferable to the analysis of cartels (Stigler 1964). Indeed it is illuminating to view members of a cartel as “divisions” or “branches” of a single-firm monopolist.

The standard economic analysis of the harm from cartels (and monopolies) views them as problematic for the following five reasons:

1. Cartels appropriate consumer surplus to themselves, at the expense of the consumer (Motta 2004, pp. 41–42).
2. Cartels cause deadweight social loss (Motta 2004, pp. 43–44).
3. The creation and preservation of a cartel involves the waste of valuable social resources (Posner 1975).
4. Cartel activity retards the development of new products and processes, thereby depriving consumers of these possible innovations (Motta 2004, pp. 45–47).
5. Participation in a cartel exacerbates managerial slack or “X-inefficiency” (Leibenstein 1966).

In addition to this economic analysis of cartel harm, there are more normative analyses which identify at least some of the harm caused by these arrangements which occurs when cartelists are not playing by the expected rules of the marketplace and do so in a clandestine manner (Whelan 2007, 2014; MacCulloch 2012; Wardhaugh 2012, 2014). Each of these harms is analyzed below.

Appropriation of Consumer Surplus

Assuming the demand curve for a particular good is downward sloping, if the monopolist (or cartel) reduces production, the price of the good will rise. A monopolist (and hence a cartel) will set its production of goods to reflect the marginal revenue. This latter amount of production is less than the amount that would be produced were the price set at marginal cost. As fewer goods are produced by a cartel, their price will rise. Given that consumer surplus is the difference between the price the consumer paid for a good and the maximum price the consumer would pay for a good (reservation price), the elevation in price by a cartel represents a reduction in consumer surplus. Further, as producer surplus (the difference between price for which a good is sold and the cost to produce the good) is the “opposite side of the coin” of consumer surplus, the consumer's loss is therefore the producer's (cartelist's) gain.

It is this appropriation of consumer surplus which has led some to characterize cartel activity as a form of theft. In this vein, a former European Competition Commissioner has used the term “rip-off” to describe the activities of cartels (Kroes 2009). These words are echoed by Whish (2000, p. 220) who claims, “on both a moral and practical level, there is not a great deal of difference between price-fixing and theft.”

Creation of Deadweight Loss

Deadweight loss is frustrated (non-)consumption. In the above case, while the harm done is that the consumer paid “too much” (i.e., a super-competitive price) for the product, the purchaser was able to consume the product, as the cartelized price for the good was less than the consumer's reservation price. If under competitive conditions the price of the good would have been below a

potential consumer's reservation price, but if the cartelized price of the good exceeds the reservation price, the good is not purchased. This frustrated consumption is a social deadweight loss.

Costs of Establishing and Maintaining Cartels

Posner (1975) argues that the realization of and ongoing maintenance of a monopoly position (or cartel) involves the expenditure of resources. In the case of monopolies, this can involve lobbying costs associated with regulation (to keep out competition) and other rent-seeking activities. In the case of cartels, such costs include the costs of keeping the arrangement clandestine and even the costs associated with verifying members' compliance with the terms of the agreement (Marshall and Marx 2012, pp. 130–137). It is not an infrequent practice for cartel members to outsource this “audit function” to a third party perceived as neutral to the participants (Marshall and Marx 2012, pp. 134–135). The resources expended on these cartel-preserving activities are viewed as a form of wasted or nonsocially beneficial expenditure.

Reduced Innovation of Products and Productive Processes

One of the rewards of participation in a cartel is a guaranteed return without the requirement (or effort) of engaging in the competitive process. Following Motta (2004, pp. 45–51) we note that given the agreement among cartelists not to compete, there is no incentive for any cartel member to develop new (i.e., “improved”) products and to spend resources improving their products or designing new processes to produce the products more efficiently. Indeed, in contrast to a monopolist (who may be concerned with a potential competitor developing a substitute for the monopolized produce and thus may therefore wish to make some investment lest this sort of unwanted entry occurs), given an agreed “standstill” on development, members of a cartel have even less incentive to invest in new products or productive efficiencies.

Managerial Slack

Managerial slack arises from the agency nature of the owner-manager/employee relationship in a firm (Leibenstein 1966; Jensen and Meckling

1976; Motta 2004, p. 47). In effect while owners (shareholders) care about the return on their investment, they delegate the day-to-day operation of the firm to managers. However, managers (and employees) will maximize their own utility function when carrying out their responsibilities. While they may care about the overall profitability of the firm (particularly when incentivized to do so by their remuneration package), they will also care about other matters, such as the effort which they are required to exert in the performance of their duties. By insulating agents in the firms from the rigors of the competitive process, while at the same time ensuring that sales or profit targets are met, cartel behavior is thus not merely a manifestation of managerial slack, but is also an active response for those who wish to pursue a “quiet life,” as John Hicks once remarked.

Normative Concerns

In addition to the economic harms isolated above, the legal literature also locates the harm occasioned by cartel activity in the effect which such conduct has on the competitive process (e.g., Whelan 2007, 2014; MacCulloch 2012; Wardhaugh 2012, 2014). Given that there is an understanding by participants in a market transaction that the exchange occurs under conditions of competition, by participating in a collusive arrangement with its ostensive “competitors,” a cartelist violates this expectation, thereby perceived as taking advantage of this position in the marketplace or as bringing into disrepute the fairness of the marketplace by clandestinely not playing by its rules. Recent amendments to the UK's *Enterprise Act 2002* reflect this normative position, as sections 188A and B of that Act (as amended by the *Enterprise and Regulatory Reform Act 2013*) exempt from criminal liability individuals who openly enter into such an agreement or otherwise provide affected customers of the details of such arrangements.

Control of Cartels

Given the pernicious nature of many of these agreements, competition regimes attempt to

control such collaboration. In the EU, Article 101 of the *Treaty for the Functioning of the European Union* (TFEU) regulates agreements among competitors. Paragraph 1 of that Article prohibits:

... all agreements between undertakings, decisions by associations of undertakings and concerted practices which may affect trade between Member States and which have as their object or effect the prevention, restriction or distortion of competition within the internal market, and in particular those which:

- (a) directly or indirectly fix purchase or selling prices or any other trading conditions;
- (b) limit or control production, markets, technical development, or investment;
- (c) share markets or sources of supply;
- (d) apply dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage;
- (e) make the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.

However, paragraph 3 of the same Article exempts from prohibition agreements which improve the production or distribution of goods or technical progress while ensuring consumers receive a fair share of this benefit. This exemption permits agreements which promote, inter alia, research and development, distribution, technology transfer, and licensing. While the EU bases its appraisal of the legality of a mooted scheme on self-assessment (Regulation 1/2003, recitals 4 and 5) rather than the former regime of prior notification and clearance (Regulation 17/1962), the Commission also publishes Guideline and Block Exemption (e.g., those in EU 2013) to provide prospective collaborators with a safe harbor for their proposed agreement. Such a safe harbor is typically available only if the market shares of the participants to the agreement are below certain thresholds, to ensure that the agreement does not permit its participants to obtain a degree of monopoly power.

In contrast, section 1 of the *Sherman Act* (the relevant US law) prohibits “every contract, combination in the form of a trust or otherwise, otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations ...” The US Supreme Court has

subsequently interpreted the words of the Act to provide for three classes of agreements, which are given differing degrees of scrutiny, depending on the economic harm which the agreement could potentially cause. The most “pernicious” (*Northern Pacific Railway* at 5) agreements are prohibited per se (as the Supreme Court recognized that such arrangements always tend to restrict competition and reduce output (*CBS* at 19–20)), and violations are proven merely through proof that the agreement fell within the prohibited category. Horizontal price-fixing, horizontal fragmentation/division of the market, and concerted refusals to deal all fall into this category.

At the other end of the scale, “rule of reason” analysis examines the entire commercial and economic context of the agreement to determine its anticompetitive effects (*Continental TV*). If the impugned activity unreasonably restrains competition, then the activity is prohibited, as the *Sherman Act* has been interpreted as prohibiting only “unreasonable” restraints on trade (*Standard Oil* at 57). Vertical resale price maintenance agreements are now analyzed under the rule of reason approach (*Leegin*). The third category, “quick look” analysis, sits between per se illegality and a rule of reason analysis. This intermediate analysis is used when a practice is not regarded as per se illegal, but “an observer with even a rudimentary understanding of economics could conclude that the arrangements in question would have an anticompetitive effect on customers and markets” (*California Dental* at 779). With quick look analysis, once the harm has been identified, the burden of proof shifts on the defendant to show the pre-competitive effects (ibid at 770). The Supreme Court employed this form of analysis in examining television restrictions on US University football games, holding that the ostensive justification (to protect live viewing of the games) did not outweigh the anticompetitive effect of broadcast restrictions (*NCAA*).

Cartel activity is controlled administratively (the EU’s means) and/or through the use of criminal sanctions. Canada was the first jurisdiction to introduce criminal penalties for cartel activity in 1889. The USA followed with the *Sherman Act* 1 year later. As of 2013 about 25 jurisdictions

have criminalized some form of this activity (Stephan 2014). Of the EU member states, the UK, Ireland, Estonia, Greece, and Germany (for bid rigging) have some form of criminal penalty (ibid). Administrative sanctions are also employed as an ex ante deterrent. In 2013, the EU Competition Commission meted out fines in the amount of €1 882 975 000 for cartel activity (EU 2014). In addition to using criminal sanctions against hard-core cartel activity, the US authorities will also use administrative sanctions as part of their anti-cartel armory.

While the EU's fine totals may appear to be an impressive deterrent, however, in Becker's (1968) sense, they may be suboptimal. There is evidence of recidivism in Europe (Connor and Helmers 2007; Connor 2010). Nevertheless, in the USA, where the mean prison sentence for cartel activity is 25 months (DOJ 2014), there has been no instance of recidivism since July 23, 1999, when the first non-American was imprisoned for anti-trust violations (Werden et al. 2011, p. 6). There has been academic discussion of the merits of introducing criminal sanctions for hard-core activity into Europe (see, e.g., Cseres et al. 2006; Beaton-Wells and Ezrachi 2011). However, this is an unlikely prospect, given that cultural attitudes in Europe may not support this use of criminal law and its harsh sanctions (Stephan 2008; Brisimi and Ioannidou 2011).

Civil damages for cartel violations are available both in Europe and in the USA. In the USA, s 15 of the *Clayton Act* permits recovery of triple damages. There is a strong argument that these so-called triple damages do not overcompensate plaintiffs, but merely top up the award of uncompensated losses, such as prejudgment interest, plaintiff's expenses (e.g., experts' reports), and litigation costs (Lande 1993). The presence of a class action regime facilitates the compensation of affected parties, by facilitating the pursuit of smaller claims. In Europe, the pursuit of civil remedies is governed by national law which governs matters including standing, limitation periods, and damages. Opt-out (American-style) class actions are unknown in Europe. The few European jurisdictions which permit class actions (or "collective actions," as they are generally

referred to in Europe) typically do so on an opt-in basis. Given the effort (even if small) required to opt into a collective action, classes are typically small, thus leading to the compensation of few affected parties. In one instance in the UK, as a follow-on action to a finding of a Competition Tribunal, only 150 individuals opted into a class to obtain compensation for the overcharge on replica football kits (Which? 2011). This system clearly undercompensates victims of anti-competitive activity. To the extent that private damages supplement the deterrent effect of public enforcement, the lack of opt-out collective actions also likely under-deters anticompetitive conduct.

Types of Cartels

Cartel behavior typically manifests itself as agreements among competitors at the same level of the supply chain (horizontal cartels), as agreements among entities at different levels of the supply chain (vertical arrangements), or as arrangements where information is passed between members at the same level of the supply chain via an intermediary (usually a wholesaler or distributor) at a different level (so-called "A-B-C" or "hub-and-spoke" cartels).

Horizontal Cartels

Typically agreements among competitors who operate at the same level of the supply chain are regarded with the greatest suspicion by regulators, as these sorts of arrangements are most apt to harm consumers' interests through the ability of the participants' collusion to acquire some degree of monopoly power, and use this power to extract monopoly rents from their customers. Typical of such pernicious arrangements are:

- Price-fixing
- Bid rigging
- Restriction of output
- Allocation of geographic territory
- Allocation of customers

Such arrangements, referred to as "hard-core cartel activity" (e.g., OECD 2000, 2002; WTO

2002), are the subject of prohibition of every legal regime which seeks to control anticompetitive conduct. Indeed, in *Verizon* (at 408) Justice Scalia of the US Supreme Court referred to collusion as “the supreme evil of antitrust” on account of the economic harm these practices inflict.

However, not all forms of cooperation at the horizontal level are necessarily consumer welfare reducing. For instance, agreements on standards permit consumer gains from network effects. Likewise, agreements on research and development permit a joint venture between competitors to share the partners’ comparative advantages in skill, industrial property, and other resources in a symbiotic manner which could allow for the development of new products (and possibly share the financial risk associated with such a project) (Faull and Nikpay 2014, p. 891). In the EU context, the legality of such agreements is governed by TFEU Article 101(3); and to facilitate self-assessment, the European Commission has promulgated a number of regulations and guidelines on such agreements. These provide a safe harbor for proposed agreements and are predicated on the parties having a low market share in the relevant product markets. The requirement of a low market share ensures the inability of the parties to obtain significant market power and hence extract monopoly rents via their cooperation. In the USA, the Congress has provided research and development statutory exemptions from the Sherman Act (e.g., *National Cooperative Research Act of 1984* and the *National Cooperative Research and Production Act of 1993*). Other forms of agreements will be judicially scrutinized under either the rule of reason or quick look approaches discussed above. Further, to provide additional guidance, the Department of Justice and Federal Trade Commission have published joint guidelines on such arrangements (e.g., DOJ/FTC 1995; FTC/DOJ 2000).

Vertical Cartels/Agreements

These agreements operate at different levels of the distribution scheme; the most common of these sorts of arrangements is resale price maintenance (RPM), namely, a system where the resale price (usually minimum) of a certain good is set by

those higher in the chain (typically a manufacturer or distributor). The traditional view of these sorts of agreements is that they either fix prices or (if the set price is a “recommended” price) facilitate price-fixing. As such this practice was formerly viewed as per se illegal in the USA (*Dr Miles*). Since 2007, it is now examined under the rule of reason (*Leegin*). The economic reasoning (primarily starting with Bork 1993, pp. 280–298) in support of relaxing the per se restriction shows not only that there may be consumer welfare-enhancing effects of these products but also that “vertical restraints are not means of creating restriction of output” (Bork 1993, p. 290). These welfare-enhancing effects most prominently include encouraging pre- and post-sale service that would not be provided where retailers compete on price, due to the problem of free riders (Marvel and McCafferty 1984, pp. 347–349; Mathewson and Winter 1998, pp. 74–75; see also Guidelines on Vertical Restraints, para 224).

In Europe, however, the skepticism of the economic benefits of RPM remains. Although the EU’s Regulation 330/2010 on Vertical Agreements provides a safe harbor for vertical arrangements among firms with low market shares (Art 2(2)), it specifically excludes RPM practices from this exemption (Art 4(a)). In theory, as with any arrangement, a particular RPM practice could be justified under TFEU Art 101(3), but a close reading on the Commission’s guidance on point (Guidelines on Vertical Restraints, paras 223–229) seems to suggest that procompetitive effects are unlikely to be conclusively demonstrated to the satisfaction of the European enforcement authorities.

Other types of vertical arrangements take such forms as market partitioning (i.e., an exclusive grant of sales rights in a particular geographic area to a particular firm) or customer allocation. The EU regime views market partitioning as a threat to market integration, and since the first case, decided by the European Court (*Consten*), European law has attempted to balance integration concerns with efficiency (Jones and Sufrin 2014, p. 790). As Monti (2002, p. 1066) notes the absolutism of both the Chicago School (that geographic restrictions imposed by firms possessing

low market power are efficiency enhancing) and the Commission (that all territorial restrictions are inefficient and thus suspect) is likely misplaced, with a case-by-case analysis to be the most accurate means of assessing efficiencies. Regulation 330/2010 permits customer allocation in cases of small (less than 30%) market share in both seller's and buyer's market (Art 3(1); see Guidelines on Vertical Restraints, para 169). In cases like this the low market share precludes the sellers from extracting monopoly rents and may permit efficiencies when customer-specific investment is appropriate (Guidelines on Vertical Restraints, paras 172–173).

Hub-and-Spoke Cartels

These arrangements have both vertical and horizontal features, where the communication of information to members at the same level (the horizontal element) is made via a member at a different level (the vertical element). The vertical member is typically a wholesaler or distributor. In the UK, recent examples of such cartels include board games (*Argos*), replica football kits (*JJB Sports*), and the dairy industry (*Tesco*). In the USA, the recent e-books case (*Apple*) is an example. As hub-and-spoke cartels only differ from horizontal and vertical cartels by the means in which information is conveyed (via an intermediary at a different level, rather than directly among competitors), the economic analysis of their harms (or efficiencies) is identical to the above cases.

Industries Susceptible to Cartelization

It is an unfortunate fact that no industry appears to be immune from cartelization (see Jones and Sufrin 2014, p. 681, for a non-exhaustive list of European cartels). However, industries which exhibit certain characteristics tend to be more susceptible to collusion. Following Jones and Sufrin (2014, pp. 664–665; see also Veljanovski 2006, pp. 4–6 and Marshall and Marx 2012, pp. 211–237), we note that cartelization is easier (and hence more predominant) in industries in which:

- There are fewer firms.
- Where high entry barriers exist.
- Where cost structures are similar.
- Where the market is transparent.
- Where trade associations or other means of coordination exist.
- Where buyers have little countervailing purchasing power.
- Where the industry is operating in depressed conditions or below capacity.
- Where the good is an intermediate product.
- Where the good is homogeneous.

These are merely indicia of industries where such activity occurs; they are neither individually nor cumulatively necessary and sufficient conditions for cartelization. Bid rigging in public-sector construction projects may be an example of a form of collusion in an industry which does not exhibit many of the above features.

Collusion

There is a wide spectrum of activity which runs from explicit agreements to coordinate activity on the market to independent – but parallel – conduct in response to changes in market conditions (Kaplow 2013, pp. 21–49). EU law (TFEU Art 101(1)) prohibits “agreements between undertakings, decisions by associations of undertakings and concerted practices” which are anticompetitive. The concept of an agreement has been liberally interpreted by the European General Court, to mean the parties’ expression of “a joint intention to conduct themselves on the market in a specific way” (*Bayer*, para 67). The gravamen of the prohibited conduct is the making of an agreement: it need not be implemented. European law does not require that the agreement be formally accepted, tacit acceptance can suffice (*Ford*, *Sandoz*). On the other hand, a concerted practice exists where parties without taking their activity “to a stage where an agreement properly so-called has been concluded, knowingly substitute[s] for the risks of competition practical co-operation between [them]” (*Hüls*, para 158). This is a wider concept than an agreement, and for the arrangement to run

afoul of European law, it must be implemented on the market (*Hüls*, para 165).

The existence and proof of a concerted practice are two distinct legal issues. Although in *ICI* (at para 109), the ECJ held that a concerted practice could safely be inferred as it could “hardly be conceivable that the same action could be taken spontaneously at the same time, on the same national markets and for the same range of products,” subsequent decisions have admonished the Commission for being too ready to find such an infringement. In *Ahlström Osaakeyhtiö* (at para 71), the ECJ held that parallel conduct “cannot be regarded as furnishing proof of concentration unless concentration constitutes the only plausible explanation for such conduct.” In contrast, American law will seek to establish the existence of “plus factors” which are suggestive – or allow for the inference – of collusion (Kovacic et al. 2011; Kaplow 2013, pp. 109–114). Such plus factors include firms’ behavior, which appears not to be in their best interests, supernormal returns within an industry, and interfirm transfers of resources (Kovacic et al. 2011, pp. 415, 435–436).

Noneconomic Considerations

TFEU Article 101(3) permits some forms of collusive behavior if certain efficiencies are met and consumers obtain a “fair share” of the benefit of these efficiencies. The term “fair share” is not an economic one and raises normative considerations (See Witt 2012b). In the 1980s and 1990s, this paragraph was used to justify agreements in the *Synthetic Fibre* and *Dutch Brick Industries* (*Stichting Baksteen*) cases to permit an orderly restructuring of these industries. However, since 1999 the Commission has pursued a more “economic approach” to the enforcement of competition law (Witt 2012a, b, pp. 453–455); and as a consequence it is unlikely that such reasoning would be followed today. Nevertheless, it remains an open question as to whether this approach accurately reflects the wording of the European Treaties (Hodge 2012, pp. 104–114; Witt 2012b, p. 471).

Conclusion

The literature on cartels and collusion is voluminous, and space considerations preclude from considering more than an overview of the main issues associated with this sort of activity. What the literature does show, however, is that in certain forms, such activity is economically beneficial, allowing firms to cooperate to produce new or innovative products or processes, thereby permitting consumers to gain. However, these instances of beneficial cooperation are infrequent. The majority of instances of this sort of collusive activity take the form of conspiracies to fix prices or reduce output, to the disadvantage of the consumer. Given the economic harm these sorts of hard-core cartels inflict, control of them is a priority for any competition regime. Yet, there is no consensus on the means by which such collusion is controlled, with both criminal and administrative (both supplemented by ex post private lawsuits) being the chosen means.

References

Books, Articles, Speeches and Other Documents

- Beaton-Wells C, Ezrachi A (eds) (2011) *Criminalising cartels: critical studies of an international regulatory movement*. Hart, Oxford
- Becker G (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Bork RH (1993) *The antitrust paradox: a policy at war with itself*, rev edn. Basic Books, New York
- Brisimi V, Ioannidou M (2011) Criminalizing cartels in Greece: a tale of hasty developments and shaky grounds. *World Compet* 34:157–176
- Connor JM (2010) Recidivism revealed: private international cartels 1990–2009. *Compet Policy Int* 6:101–127
- Connor JM, Helmers CG (2007) Statistics on private international cartels. American Antitrust Institute AAI working paper no 07–01. Available at <http://ssrn.com/abstract=1103610>. Accessed 4 Oct 2014
- Cseres KJ, Schinkel MP, Vogelaar FOW (eds) (2006) *Criminalization of competition law enforcement: economic and legal implications for the EU member states*. Edward Elgar, Cheltenham
- Department of Justice (2014) Criminal enforcement fine and jail charts through fiscal year 2013. Available at <http://www.justice.gov/atr/public/criminal/264101.html>. Accessed 4 Oct 2014
- Department of Justice and Federal Trade Commission (1995) *Antitrust guidelines for the licensing of*

- intellectual property. Available at <http://www.justice.gov/atr/public/guidelines/0558.htm>. Accessed 4 Oct 2014
- EU Competition Commission (2013) EU competition law rules applicable to antitrust enforcement. General block exemption regulations and guidelines, vol II. Brussels, EU. Available at http://ec.europa.eu/competition/antitrust/legislation/handbook_vol_1_en.pdf. Accessed 4 Oct 2014
- EU Competition Commission (2014) Cartel statistics. Available at <http://ec.europa.eu/competition/cartels/statistics/statistics.pdf>. Accessed 4 Oct 2014
- Faull J, Nikpay A (2014) The EU law of competition, 5th edn. Oxford University Press, Oxford
- Federal Trade Commission and Department of Justice (2000) Antitrust guidelines for collaborations among competitors. Available at http://www.ftc.gov/sites/default/files/documents/public_events/joint-venture-hearings-antitrust-guidelines-collaboration-among-competitors/ftcdojguidelines-2.pdf. Accessed 4 Oct 2014
- Hodge TC (2012) Compatible or conflicting: the promotion of a high level of employment and the consumer welfare standard under article 101. *William and Mary Bus Law Rev* 3:59–138
- Jensen M, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. *J Financial Econ* 34:305–360
- Jones A, Sufrin B (2014) EU competition law: text, cases and materials, 5th edn. Oxford University Press, Oxford
- Kaplow L (2013) Competition policy and price fixing. Princeton University Press, Princeton/Oxford
- Kovacic WE, Marshall RC, Marx LM, White HL (2011) Plus factors and agreement in antitrust law. *Michigan Law Rev* 110:394–436
- Kroes N (2009) Tackling cartels – a never-ending task. Anticartel enforcement: criminal and administrative policy – panel session, Brasilia. Available at <http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/09/454%26format=HTML%26aged=0%26language=EN%26guiLanguage=en>. Accessed 4 Oct 2014
- Lande RH (1993) Are antitrust ‘treble’ damages really single damages. *Ohio State Law J* 54:115–174
- Leibenstein H (1966) Allocative efficiency vs. X-inefficiency. *Am Econ Rev* 56:392–415
- MacCulloch A (2012) The cartel offence: defining an appropriate ‘moral space’. *Eur Compet J* 8:73–93
- Marshall RC, Marx LM (2012) The economics of collusion: cartels and bidding rings. MIT Press, London/Cambridge, MA
- Marvel HP, McCafferty S (1984) Resale price maintenance and quality certification. *Rand J Econ* 15:346–359
- Mathewson F, Winter R (1998) The law and economics of resale price maintenance. *Rev Ind Organ* 13:57–84
- Monti G (2002) Article 81 EC and public policy. *Common Mark Law Rev* 39:1057–1099
- Motta M (2004) Competition policy: theory and practice. Cambridge University Press, Cambridge
- Neils G, Jenkins H, Kavanagh J (2011) Economics for competition lawyers. Oxford University Press, Oxford
- Organisation for Economic Co-operation and Development (OECD) (2000) Hard core cartels: 2000. OECD, Paris
- Organisation for Economic Co-operation and Development (OECD) Directorate for Financial, Fiscal and Enterprise Affairs Competition Committee (2002) Report on the nature and impact of hard core cartels and sanctions against cartels under national competition laws. DAFE/COMP(2002)7 OECD, Paris. Available at www.oecd.org/dataoecd/16/20/2081831.pdf. Accessed 4 Oct 2014
- Posner R (1975) The social cost of monopolies and regulation. *J Polit Econ* 83:807–827
- Smith A (1776) An inquiry into the nature and causes of the wealth of nations. London: printed for W. Strahan; and T. Cadell
- Stephan A (2008) Survey of public attitudes to price-fixing and cartel enforcement in Britain. *Compet Law Rev* 5:123–145
- Stephan A (2014) Four key challenges to the successful criminalization of cartel laws. *J Antitrust Enforcement* 2:333–362
- Stigler G (1964) A theory of oligopoly. *J Polit Econ* 72:44–59
- Veljanovski C (2006) The economics of cartels. *Finnish Competition Law Year Book*. Available at <http://ssrn.com/abstract=975612>. Accessed 4 Oct 2014
- Wardhaugh B (2012) A normative approach to the criminalisation of cartel activity. *Legal Stud* 32:369–395
- Wardhaugh B (2014) Cartels, markets and crime: a normative justification for the criminalisation of economic collusion. Cambridge University Press, Cambridge
- Werden GJ, Hammond SD, Barnett BA (2011) Recidivism eliminated: Cartel enforcement in the United States since 1999. Speech before the: Georgetown global antitrust enforcement symposium. Washington, DC. Available at <http://www.justice.gov/atr/public/speeches/275388.pdf>. Accessed 4 Oct 2014
- Whelan P (2007) A principled argument for personal criminal sanctions as punishment under EC cartel law. *Compet Law Rev* 4:7–40
- Whelan P (2014) The criminalization of European cartel enforcement theoretical, legal, and practical challenges. Oxford University Press, Oxford
- Which? (A Consumer Advocacy Group) (2011) JJB sports: a case study in collective action. Available at <http://www.which.co.uk/documents/pdf/collective-redress-case-study-which-briefing-258401.pdf>. Accessed 4 Oct 2014
- Whish RP (2000) Recent developments in community competition law 1998/99. *Eur Law Rev* 25:219–246
- Witt AC (2012a) From Airtours to Ryanair: is the more economic approach to EU merger law really about more economics? *Common Mark Law Rev* 49:217–246
- Witt AC (2012b) Public policy goals under EU competition law – now is the time to set the house in order. *Eur Compet J* 8:443–471
- World Trade Organization (WTO) (2002) Working group on the interaction between trade and competition policy, provisions on hardcore cartels: background note by the secretariat. WT/WGTCP/W/191. Geneva, WTO. 20 June 2002

Cases

Commission Decisions

Dutch Brick Industry (“Stichting Baksteen”), Commission Decision of 29 Apr 1994 (IV/34.456) [1994] OJ L-131/15

Synthetic Fibres, Commission Decision of 4 July 1984 (IV/30.810) [1984] OJ L-207/17

European Courts

Case 48/69 ICI v Commission [1972] ECR 619

Case 89/85 etc Ahlström Osakeyhtiö et al v Commission (“Wood Pulp II”) [1993] ECR I-1307

Case C-199/92 P Hüls AG v Commission [1999] ECR I-4287

Case C-277/87 Sandoz prodotti farmaceutici SpA v Commission [1990] ECR I-45

Case T-41/96 Bayer AG v Commission [2000] ECR II-3383, affirmed Cases C-2 and 2/01 P [2004] ECR I-23

Cases 56 and 58/64 Établissements Consten S.à.R.L. and Grundig-Verkaufs-GmbH v. Commission [1966] ECR 299

Cases 228 and 229/82 Forde Werke AG and Ford of Europe Inc v Commission [1984] ECR 1129

UK

Argos Limited and Littlewoods Limited v Office of Fair Trading [2004] CAT 24 (“Board Games”) (Judgment on Liability)

JJB Sports PLC v Office of Fair Trading and Allsports Limited v Office of Fair Trading [2004] CAT 17 (“Replica Football Kits”) (Judgment on Liability)

Tesco Stores Ltd, Tesco Holdings Ltd and Tesco Plc v Office of Fair Trading [2012] CAT 31 (“Dairy”) (Judgment on Liability)

US

Broadcast Music, Inc v CBS, 441 US 1 (USSC 1979)

California Dental Association v FTC, 526 US 756 (USSC 1999)

Continental TV, Inc v GTE Sylvania Inc, 433 US 36 (USSC 1977)

Dr Miles v John D. Park & Sons Co, 220 US 373 (USSC 1911)

Leegin Creative Leather Products, Inc v PSKS, Inc, 551 US 877 (USSC 2007)

National Collegiate Athletic Association v Board of Regents of University of Oklahoma, 468 US 85 (USSC 1984)

Northern Pacific Railway v United States, 356 US 1 (USSC 1958)

Standard Oil Co v United States, 221 US 1 (USSC 1911)

United States v Apple Inc, 952 F Supp 2d 638, 2013–2 Trade Cases P 78,447 (SDNY July 10, 2013)

Verizon Communications Inc v Law Offices of Curtis V Trinko, LLP, 540 US 398 (USSC 2004)

Statutes and Other Legal Instruments

UK

Enterprise Act 2002

Enterprise and Regulatory Reform Act 2013

European Union

Commission Regulation (EU) No 330/2010 of 20 April 2010 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to categories of vertical agreements and concerted practices [2010] OJ L-102/1

Council Regulation (EC) No. 1/2003 of 16 December 2002 on the implementation of the rules on competition laid down in Articles 81 and 82 of the Treaty (“Regulation 1/2003”) [2003] OJ L-1/1

Guidelines on Vertical Restraints [2010] OJ C-130/1

Regulation No 17: First Regulation implementing Articles 85 and 86 of the Treaty (“Regulation 17/62”) OJ 013, 21/02/1962 pp 0204–0211

Treaty for the Functioning of the European Union (Consolidated version) [2012] OJ C-326/47

US

Clayton Act, 15 USC §§ 12–27, 29, 52–53

National Cooperative Research Act of 1984 and the National Cooperative Research and Production Act of 1993, codified together at 15 USC §§ 4301–06

Sherman Act, 15 USC §§ 1–7

Cash Demand

Gerhard Graf

Johamer Qutenberg-Universität Maine, Nieder-Olm, Germany

Abstract

After a definition of cash demand and its peculiarities, cash demand is compared to the more encompassing money demand. Then, the main economic influences on cash demand are explained. A final point exhibits future developments of cash demand which is threatened by electronic inventions enhancing a cashless society.

Synonyms

[Currency demand](#)

Definition

Cash demand is a demand for the legacy currency of a country which is influenced by residents and nonresidents of a country.

The Meaning of Cash Demand

Cash demand is the demand for a currency, mostly banknotes. As coins constitute only a relatively small part of total cash demand, they are usually neglected in the analyses of overall cash demand. The demand for the legacy currency is, in general, influenced by the arguments which also drive macroeconomic money demand in a country. However, cash demand is specific first, because the amount demanded is normally fully accommodated by central banks, so demand growth or changes in the composition of banknotes according to their denomination are provided without macroeconomic or budget constraint consequences (Bartzsch et al. 2011). Second, cash demand underlies special microeconomic considerations which somehow balance in the analysis of total money demand, i.e., demand for cash and deposits. Third, cash demand for currencies of different legacies cannot be assumed to be identical. There are rather decisive distinctions between the demand motives for separate currencies (Fischer et al. 2004).

Cash Demand as a Component of Money Demand

Cash demand is a part of overall money demand, for instance, for M1 or M3. The analysis of total money demand should always take into account that it is not a simple sum of the demand for cash and for deposits since there are often substitution processes going on between the components. Moreover, money demand is confronted with macroeconomic money supply considerations and goals of the monetary policies which are not similarly influential for cash demand alone.

Main Influences on Cash Demand

Cash demand is mainly caused by the transaction motive. Households and businesses need cash in order to accomplish everyday consumption goods transactions. This medium-of-exchange function of cash goes together with consumption expenditures of households within a country. But not all

private consumption expenditures are paid in cash. A rising amount of cash transactions is substituted by card payments or other technical possibilities to directly withdraw money from personal deposits. In addition, cash is not used unanimously with all banknote denominations of a currency. There is evidence throughout all currencies that people concentrate on small- and medium-size denominations (Deutsche Bundesbank 2009). Thus, large banknotes are often not demanded for the payment of everyday consumption expenditures, but they serve the second demand motive as store of value. Cash as a store of value has opportunity costs. In quite a lot of countries, people rely on cash even if they could earn interest revenues and even if the opportunity cost is increased by inflationary developments of a currency. But, the use of cash as a store of value eases its future use in payments without relevant transactions costs. Additional national influences on cash demand are exerted by the fact that cash provides anonymity in the transactions which are executed with the help of banknotes. For some observers this peculiarity is decisive for the use of cash in the non-reported or illegal economy. Cash, at least some banknote denominations of a group of economies whose currencies are assumed to be more stable internally and more stable relative to other countries, does not only serve as a store of value within the countries themselves but also as a reliable store of value for people in other countries whose currencies are at the risk of inflation or devaluation. Therefore, cash demand for selected currencies is an international demand. This international demand is not only based on the function of international store of value but, to a certain extent, can also be related to the function of medium of exchange within the respective foreign countries (Deutsche Bundesbank 2012). The use of a currency beyond the national borders is a common phenomenon of the US Dollar, the Euro, and the former DM (European Central Bank 2011).

The Future of Cash Demand

During the last decades, cash is subdued to rising competitive assaults by new electronic

developments which tend to reduce cash payments, even for small transaction values, which is paramount to say that cash will lose its outstanding role as medium of exchange and as store of value. Changing payment habits may lead to a “cashless” society and in the end to a complete breakdown for cash demand (Lippi and Secchi 2009). Up to now, however, there are no signs of a definite abandonment of cash of the most important currencies. So cash demand will be a continuing phenomenon in the future. This tendency is enhanced by a supply side argument. As long as cash in the form of banknotes can carry a favorite symbol of a state and will contribute to government revenues via seigniorage, there will be an enduring cash provision easing further demand developments.

Cross-References

- ▶ [Central Bank](#)
- ▶ [Money Laundering](#)
- ▶ [Shadow Economy](#)
- ▶ [Tax Evasion by Individuals](#)
- ▶ [Underground Economy](#)

References

- Bartzsch N, Rösl G, Seitz F (2011) Foreign demand for euro banknotes issued in Germany: estimation using direct approaches. Deutsche Bundesbank discussion paper series 1: economic studies No. 20/2011
- Deutsche Bundesbank (2009) The development and determinants of euro currency in circulation in Germany. Monthly report. 45–58
- Deutsche Bundesbank (2012) The usage, costs and benefits of cash – theory and evidence from macro and micro data. Deutsche Bundesbank, Frankfurt
- European Central Bank (2011) The use of euro banknotes – results of two surveys among households and firms. Mo Bull 75–90
- Fischer B, Köhler P, Seitz F (2004) The demand for euro area currencies: past, present and future. in: European Central Bank working paper series, No. 330 April 2004
- Lippi F, Secchi A (2009) Technological change and the households’ demand for currency. J Monet Econ 56(2):222–230

Further Reading

- Ireland PN (2009) On the welfare cost of inflation and the recent behavior of money demand. Am Econ Rev 99:1040–1052
- Walsh CE (2010) Monetary theory and policy, 3rd edn. MIT Press, Cambridge, MA

Causation

Samuel Ferey

CNRS, BETA, University of Lorraine, Nancy, France

Abstract

Causation is often said to be one of the most intricate issues in private law. This entry deals with how causation is considered from a law and economics point of view and contributes to enhance our understanding of causation requirements in the law. First, it deals with the main distinction between *ex ante* and *ex post* approach of causation. This leads to study the relationships between causation requirements and efficiency, and the concept of scope of liability is discussed. Then, this economic approach is extended to more difficult cases implying uncertainty. Last, it provides some insights on how these findings remain relevant when considering bounded rationality.

What Is Causation?

“What then is time? If no one asks me, I know that it is. If I wish to explain it to him who asks, I do not know” said Saint Augustine about time. The same could be said about causation: the common sense intuitively knows what causation is, but neither science nor philosophy provides a unified concept of causation. The law does not escape the difficulty and legal causation is often said to be one of the most confused area in all the private law. As Coleman states: “No course in the first year curriculum is more baffling to the average law student than is torts, and for good reasons. In the first two weeks, the student learns that

causation is necessary for both fault and strict liability. Two weeks later the student learns that causation is meaningless, content-free, a mere buzzword. [...] Ordinary lawyers and law professors are as confused about causation and the role it plays in liability and recovery as are their students” (Coleman 1992: 270). This is all the more troublesome that causation requirements are one of the keystones of individual liability and justice: it is often said that it is because the injurer caused harm that he has to compensate the victim and that is why “we need an account of causation that can provide reasons of some weight for imposing liability” (Coleman 1992: 271; Cooter 1987).

In law, many concepts of causation are used – proximate cause, cause in fact, probabilistic causation – but the most common legal causation criterion is the “but for test.” The “but for test” has an intuitive content: “A causes B” means that B would not have occurred if A had not happened. The “but for test” relies on *necessity*, but beyond its simplicity, this criterion is challenged by many paradoxes and puzzles like overdetermined causes, concurrent causes, or uncertainty (Hart and Honoré 1985). Courts, legal theory, and legislators are aware of the difficulties (*Restatement (Third) on Torts*) and some new criteria are often discussed in the literature (Wright 1985a; Wright and Puppe 2016; Stapleton 2013).

Law and economics literature challenges these classical views on causation and wonders whether and how causation requirements are needed to optimally design tort law (Ben Shahr 2009). For law and economics scholars, the purpose of causation requirements is not so much to provide reasons for imposing liability, as to provide an optimal design of liability regimes. As Shavell states, “the basic function of causation requirement under strict liability, in others words, is that it furnishes socially appropriate incentives to reduce the risk of harm and to moderate the level of activity by imposing liability equal only to the increase in social costs due to a party’s action” (Shavell 2004: 251). From a positive point of view, economics may help to better understand legal doctrines and jurisprudence.

Torts Without Causation? From *Ex Post* to Forward Looking Causation

The paper on the *Problem of the Social Cost* published by Coase in 1960 is often said to be the birth of contemporary law and economics. At first glance, Coase’s paper is about externality. Coase criticizes classical pigouvian analysis on externality by demonstrating that, in case of zero transaction costs, efficiency could be reached by private negotiations among parties. The example of a cattle-raiser whose cows destroy crops grown by his neighbor is illuminating. Each unit of cow has a marginal cost (the harm suffered by the farmer measured by the value of the crops destroyed by this cow) and a marginal benefit (the additional profit for the cattle-raiser). First, if transaction costs are zero, people will exchange to reach the efficient level of the externality. The only condition is that people know precisely what the property rights are: if the farmer has the right not to be harmed, the cattle-raiser would buy the right to destroy his crops from him; if the cattle-raiser has the right to destroy crops, the farmer would buy the right from him. Second, if transaction costs are positive, efficiency cannot be restored by mutual advantageous exchanges: the initial property rights delineation and tort law play a crucial role to implement or not efficiency.

But the paper goes much further regarding causation. Coasean reasoning leads to consider that causation is reciprocal: “The question is commonly thought of one in which A inflicts harm on B and what has to be decided is: how should we restrain A? But this is wrong. We are dealing with a problem of reciprocal nature. To avoid the harm to B would inflict harm on A. The real question that has to be decided is: should A be allowed to harm B or should B be allowed to harm A? The problem is to avoid the more serious harm.” (Coase 1960: 2). The followers of Coase dug out deeper the idea that causation is reciprocal, and that causation should not play any role at all as soon as the aim of the tort system is to reach efficiency. In other words, causation is not a condition to hold a person liable anymore but the result of an economic reasoning (Landes and

Posner 1983). The same can be said about Calabresi “least cost avoider” principle according to which a party is liable for the harm if he was the person who would have avoided it at the least cost (Calabresi 1970; Calabresi 1975).

This first step has radical consequences: causation and efficiency are redundant inquiries as long as they are considered from an *ex post* perspective (Miceli 1997). But things are different as soon as causation is considered from an *ex ante* perspective. Following Calabresi, law and economics literature considers the injurer’s behavior may be viewed as a factor that influences the probability that harm occurs and a cause will be modeled as an increasing in the probability that harm comes about. This probabilistic approach has been criticized (Wright 1985b) but has made it possible to easily integrate causation into economics models. For example, the incidental accident illustrates this feature of causation. In the famous case *Berry v. Borough of Sugar Notch*, 43 A. 240 (1899), an accident occurs due to a falling tree on a streetcar (Shavell 2004: 253–254). An *ex post* causation perspective would consider the high speed of the bus as a but for cause for the accident. But an *ex ante* approach would lead to different conclusions: the probability that a bus be struck by a falling tree does not depend on its speed and would be avoided either the bus had been going slower or faster.

Causation, Scope of Liability, and Magnitude of Damage

Law and economics has added to the literature about causation by clarifying the role that causation requirements – both cause in fact and proximate cause – play in the design of liability regimes. A liability regime is defined by the level of care, the magnitude of compensation, and the causation requirements embedded in the scope of liability. We broadly consider causation issues here and we analyze, first, the role of the scope of liability and, second, the magnitude of the damages to be paid by the injurer due to harm he caused.

The scope of liability has been formally introduced by Shavell at the beginning of the 1980s. Shavell defines the scope of liability as the set of the states of the world where the party is held responsible for harm. Consider the following example: a firm builds a dam and takes a certain level of care below the efficient level. Consider now three possible states of the world: a low flood, a medium flood, and an overwhelming flood. Consequences may be associated to each state of the world: the first one would be contained by the dam and would cause no harm at all (event A), the second one would destroy the dam – because the dam has not been built with sufficient care – and would cause harm H (event B), the third one is so huge that it would destroy the dam even if the firm had chosen the required level of care (event C). In case of unrestricted scope of liability, a party is held responsible even if his behavior could not have prevented the loss (event C). In case of restricted scope of liability, the injurer is not held responsible for some states of the world where he would have been able to eliminate or mitigate the loss (event B). The optimal scope of liability is defined as the set of all the states of the world where the level of care is a necessary cause of the harm. In our case, C is not included in this set insofar as C leads to harm whatever the level of care be. A too restricted scope of liability leads to underdeterrence: some benefits of care are not taken into account by the injurer.

But what are the consequences of an unrestricted scope of liability? In other words, does it matter that the parties be held responsible for losses they did not caused? Suppose a firm pollutes the stream of a river and causes a loss of \$100, its benefits are \$150 and the cost of a device eliminating pollution is \$120. From an efficiency viewpoint, it is efficient for the firm to continue its activity, to pollute and to compensate victims for their losses. The strict liability rule implements efficiency: pollution occurs, victims are compensated, and the device is not bought. But suppose now that the pollution could occur due to external factors for an amount of \$100. If the parties are held responsible for the losses they did not cause, the firm is likely to pay up to \$200. It would prefer to stop its activity even if it is suboptimal.

In Shavell words, it is useful to distinguish the impact of the scope of liability on the care level on one side and on the activity level on the other side. In our example, a too broad delineation of the scope of liability would not change the level of care (in any case, the firm will not pay for the antipollution device insofar as this device has no effect on the probability that a pollution due to external factor comes out). However, it has a potentially “crushing out” effect on the activity level insofar as the firm would prefer to not produce.

In case of negligence, and supposing that the due level of care is the efficient level, an unrestricted scope of liability does not distort incentives. Even if the risk of liability is large due to this unrestricted scope of liability, the party has the incentive to take the due care level and escapes from any compensation to pay. There is no crushing effect of liability on the activity level for the same reason. However, if the scope of liability becomes too restricted, some of the states of the world where it would have been efficient for the society that precaution be taken, are not taken account by the party. This may lead to inefficient levels of care.

The second issue discussed by law and economics is about the magnitude of damage to be paid by the injurer. A subfield of the law and economics literature on causation has discussed the discontinuity of the individual’s cost function at the due care point (Grady 1983, 1989, 2014; Marks 1994; Hylon 2014; Kahan 1989). According to the literature, beyond this point, the injurer pays for the entire damage; below this point, he avoids any liability and pays nothing. Grady proposes to deeper consider causation and counterfactuals. Causation requirement is a counterfactual inquiry insofar as court wonders “what would have happen but for the injurer behavior”. But what is the relevant counterfactual? Is it the situation which would have come about without any behavior of the injurer or the situation which would have come about if the injurer had taken the level of due care? In the first case, the losses *caused* is said to be the entire loss *L*; in the second, the loss *caused* is

said to be the incremental part of damage caused by the incremental negligence of the party (compared to the level of care he was required to take). Under such a definition of “caused losses,” there is no discontinuity in the cost function anymore which leads to original consequences when the implementation of the care level is uncertain.

In most countries, tort law requires that an injurer cannot compensate the victim for more than the harm he actually caused. The statement seems to be obvious and dates back to the very beginning of the private law. Law and economics scholars have challenged this view. First, such an upper bound may lead to inefficiency. This is the case, for example, when the legal system as a whole badly works: if the probability for an injurer who caused a loss to be caught or suited is less than one, the expected cost of a wrongdoing behavior is decreasing and the level of care is likely to be inefficient. Opening the door to compensations above the loss amounts actually caused would help to reach efficiency. Punitive damages are the most obvious legal mechanisms to reach this objective. Other more complex situations involving several tortfeasors have also been studied. Efficiency arguments could be found in favor of original schemes of decoupled liability to increase the total cost paid by the injurer: the cost to be paid by the injurer is divided in two part, one is equal to the loss and is paid to the victim and the other is a fine to be paid to the State (Miceli 1997). Second, the reverse may occur. Recent researches have focused on the possibility, or not, for an injurer to mitigate the damages paid in case of offsetting benefits. Offsetting benefits come about when the injurer’s behavior caused both harms and benefits. Imagine the following example: “While driving his car, the defendant negligently hits the plaintiff, who was on her way to the airport to catch a flight, causing a minor injury to her leg. As a result of the accident, the plaintiff misses her flight, which later crashes” (Porat and Posner 2014: 1168). Should the court take into account the benefits caused by the injurer? Principles that courts could follow in these cases may be found in Porat and Posner (2014).

Causation and Uncertainty

As Shavell states, “in many situations, there is uncertainty about causation. For example, it may be not known which manufacturer out of many sold the product (a drug, a lead paint) that caused the injury, or whether an injury was caused by the defendant or by background factors” (Shavell 2004: 254). Such cases may have consequences on the design of optimal liability rule in terms of efficiency (inadequate or excessive incentives), legal errors, or, more broadly, consistency of the law.

Uncertainty over causation occurs when it is impossible to provide sufficient evidence to prove with certainty that a behavior has actually caused harms. Causation uncertainty has two sides: first, when injurers are unknown with certainty; second, when victims are not perfectly identified. The first case refers to the famous example of alternative causation when two hunters shoot and only one bullet hits the victim. The second case is illustrated in the following example. Imagine a neighborhood where each year a fixed proportion of its inhabitants get sick. Suppose the injurer spreads toxic in the air and therefore increases the risk of this disease. Under such circumstances, it is possible to prove that a proportion of the sick people is due to this negligent behavior but it is impossible to know with certainty which ones are the victims of the wrongdoing behavior and which one would have get sick but for it. The injurer is perfectly known but the victims are not.

Uncertain causation is another name for a lack of information due to insufficient evidence: causation is probabilistic not only *ex ante* but also *ex post*. Different rules may be used to determine whether or not the defendant be held liable. Some are implemented by torts law systems all around the world, some may be found only in certain countries, some are still speculative proposals, studied and discussed by scholars without any legal counterparts. These rules have different interesting properties in terms of efficiency, incentives, corrective justice and legal errors. An economic analysis of these rule leads to characterizing the trade-off between these competing objectives.

First is the rule based on the preponderance of the evidence. A party will be held responsible if it is more probable he caused the harm than not. Usually, the preponderance of the evidence is said to require a probability of 51%. According to Kaye, the first who considers this issue from a legal error perspective, this rule leads to minimizing legal error. Suppose the probability that A caused a \$1000 loss is 30%. Regarding the preponderance of the evidence, A should not be held liable. The legal error is low: in 30 cases out of 100, A should have paid \$1000 and pays zero and in 70 cases out of 100, A should pay zero and pays zero: the total error is \$300. Compare now to a proportional liability rule: A pays the amount of the expected damages caused. The error is now of \$700 in 30 cases out of 100 (A pays \$300 while he caused a loss of \$1000) and \$300 in 70 cases out of 100 (A pays \$300 while he caused zero). The total error is \$420.

Even though this minimizing errors, property depends on how legal error is defined (Ferey and G'Sell 2013), the fact that preponderance of the evidence be a threshold rule leads to unexpected consequences. First, it distorts the *ex ante* incentives and therefore does not lead to efficient deterrence. A party who would be aware that the probability of his behavior is below the threshold has no incentive to take the due level of care and to avoid liability while a party aware that its own probability is above the threshold is over deterred. All or nothing rules play as an unrestricted or a too restricted scope of liability. Inefficient activity levels may be expected. More recently, Levmore (2001) and Porat and Posner (2012) have raised paradoxes due to preponderance of the evidence. Suppose that a party involved in two separate and independent accidents and that the probability that he caused each accident is said to be 0.4. Following the preponderance of evidence, the injurer will be exonerated from any liability in both cases (it is more probable that he has not caused the damage, 0.6). This is the result of the black letter of the preponderance of evidence rule. The puzzle arises by considering that the probability that he caused *at least* one of them is above the threshold ($1 - 0.6^2 = 0.64$).

The second rule discussed by the law and economics literature is the proportional rule. The compensation paid by a party held responsible is proportional to the probability he had to cause the harm. The market share liability doctrine is one of the most famous examples of a proportional rule. Adopted by the Supreme Court of California (*Sindell v. Abbott Laboratories* 607 P.2d 924 (1980)), this rule is adopted, for example, by the European Group of Tort Law (European Group on Tort Law 2005). Under this rule, each potential injurer compensates the victims up to the probability he had to have caused the harm and pays for the expected damages he caused. Knowing whether such a rule leads to efficiency is still controversial. Some consider that the care level and the activity levels will be efficient under a proportional liability rules, others focus on the free rider problem as such as care is considered as a public good (Rose-Ackerman 1990).

Some have proposed to extend this logic in a more radical way. The risk-based liability regime is an example. Within such a regime, structured on the sole probabilities, an injurer is liable for the risks of losses suffered by the victim. From an *ex ante* perspective, a risk-based approach is compatible with efficient incentives, but many strange effects arise from an *ex post* perspective. Such a regime means that the causal link between the injuring behavior and the losses is broken: to be consistent, such an approach will give the right for a potential victim to get compensation even if he has suffered only the risk to be harmed and not the harm itself. And the actual victims of the loss will be undercompensated and will get back only the risk of harm.

Last, a particularly interesting rule is provided by Porat and Stein. For these authors, uncertainty has to be considered as a decision variable from the parties. If, the court or the victim has no sufficient information *ex post*, it is because the parties did not take sufficient precaution to provide the information. The principle to be applied is therefore the *least evidence providers*: liability has to be bear by the party which could have been the most efficient provider of information. In other words, it is efficient that the ones who could have

provided the evidence at the cheapest cost be the ones who should be declared liable (Porat and Stein 2001). Parties have incentives to avoid uncertainty and the efficiency liability regimes may be applied.

Causation Beyond Rationality

Recent literature in law and economics has challenged the mechanisms by which law induces people to behave in a certain way. Leading by Kahneman and Tversky, behavioral scientists raised important issues about bounded rationality in general and in the law in particular. Several findings of behavioral law and economics have wide impact and significance on causation. First Kahneman himself elaborates on the cognitive illusion of causation (Kahneman 2011). Biases about competence or about the causation links between a behavior and a result have been extensively studied. An interesting example of cognitive bias over causation is provided by the cases of hypothetical causal links between multiple sclerosis and vaccination against hepatitis B. The temporal succession of the events (vaccination following by the disease) seems sometimes to be sufficient to prove causation and courts seem to be victims of a representative bias (Borghetti 2016: 558).

Second, as probabilistic causation is one of the cornerstone of economic approach, many heuristics and biases leading to poorly assess probabilities could impact how people consider causal relationships. If courts, injured people, and injurers make systematic errors on probabilities, they will be mistaken on the true probabilistic links between an event and harm. The hindsight bias exemplifies how a behavioral approach renews law and economics of causation. The hindsight bias covers the fact that an event could be considered *ex ante* as unlikely and, at the same time, be considered as highly probable once it came about. In others words, people suffer inconsistencies in their assessment of the probability of an event depending on the fact that the event has occurred yet or not. In several papers, Rachlinski tests the hindsight bias in judging to see whether

people tend to overestimate the probability of an event once occurred. (Rachlinski 1998). What is at stake is the very ability of the legal system to implement efficient due care level that would be consistent through time: “The bias, in general, makes defendants appear more culpable than they really are. The bias can cause judges and juries to find liable even those defendants who attempted to avoid negligence by undertaking all reasonable precautions in foresight. Not only does this seem unjust, but it also might have adverse economic consequences. Any potential defendant who is aware of the implications of the hindsight bias might try to avoid liability by taking an excess of precautions. The hindsight bias thus suggests a problem with the law and economics of negligence” (Rachlinski 1998: 572). A lot has still to be done in this field but notice that law and economics converge to recent works in experimental philosophy questioning causation and intention.

Cross-References

- ▶ [Multiple Tortfeasors](#)
- ▶ [Nuisance](#)
- ▶ [Strict Liability Versus Negligence](#)

References

- American Law Institute (2010) Restatement (third) of the law of torts: liability for physical and emotional harm
- Ben Shahrar O (2009) Causation and foreseeability. In: Faure M (ed) Tort law and economics. Edward Elgar, Cheltenham, pp 83–108
- Borghetti JS (2016) Causation in hepatitis B vaccination litigation in France: breaking through scientific uncertainty. *Chicago Kent Law Rev* 91:543–568
- Calabresi G (1970) The costs of accidents. Yale University Press, New Haven
- Calabresi G (1975) Concerning cause and the law of torts: an essay for Harry Kalven. *Univ Chicago Law Rev* 43:69–100
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Coleman JL (1992) Risks and wrongs. Cambridge University Press, Cambridge
- Cooter RD (1987) Torts as the Union of Liberty and Efficiency: an essay on causation. *Chicago Kent Law Rev* 63:522–551
- European Group on Tort Law (2005) Principles of European tort law. Text and commentary. Springer, Vienna
- Ferey S, G’Sell F (2013) Pour une prise en compte des parts de marché dans la détermination de la contribution à la dette de réparation. *Recueil Dalloz* 41:2709
- Grady MF (1983) A new positive theory of negligence. *Yale Law J* 92:799–829
- Grady MF (1989) Untaken precautions. *J Leg Stud* 18:139–156
- Grady MF (2014) Marginal causation and injurer shirking. *J Tort Law* 7:1–34
- Hart HLA, Honore T (1985) Causation in the law, 2nd edn. Oxford University Press, Oxford
- Hylon KN (2014) Information and causation in tort law: generalizing the learned hand test for causation cases. *J Tort Law* 7:35–64
- Kahan M (1989) Causation and incentives to take care under the negligence rule. *J Leg Stud* 18:427–447
- Kahneman D (2011) Thinking, Fast and Slow. Farrar, Strauss & Giroux, New York
- Kaye D (1982) The limits of the preponderance of the evidence standard: justifiably naked statistical evidence and multiple causation. *Am Bar Found Res J* 7:487–451
- Landes WM, Posner RA (1983) Causation in tort law: an economic approach. *J Leg Stud* 12:109–134
- Levmore S (2001) Conjunction and aggregation. *Michigan Law Rev* 99:723–756
- Marks SV (1994) Discontinuities, causation, and Grady’s uncertainty theorem. *J Leg Stud* 23:287–301
- Miceli TJ (1997) Economics of the law. Oxford University Press, Oxford
- Porat A, Posner EA (2012) Aggregation and law. *Yale Law J* 122:2–69
- Porat A, Posner EA (2014) Offsetting Benefits. *Virginia Law Rev* 100:1165–1209
- Porat A, Stein A (2001) Tort liability under uncertainty. Oxford University Press, Oxford
- Rachlinski JJ (1998) A positive psychological theory of judging in hindsight. *Univ Chicago Law Rev* 65:571–625
- Rose-Ackerman S (1990) Market-share allocations in tort law: strengths and weaknesses. *J Leg Stud* 19:739–746
- Shavell S (2004) Foundations of economic analysis of law. Harvard University Press, Cambridge
- Stapleton J (2013) Unnecessary causes. *Law Q J* 129:39–65
- Wright RW (1985a) Causation in tort law. *California Law Rev* 73:1735–1828
- Wright RW (1985b) Actual causation vs. probabilistic linkage: the bane of economic analysis. *J Leg Stud* 14:435–456
- Wright RW, Puppe I (2016) Causation: linguistic, philosophical, legal and economic. *Chicago Kent Law Rev* 91:461–506

Central Bank

Karl-Friedrich Israel

Faculty of Economics and Business

Administration, Humboldt-Universität zu Berlin,
Berlin, Germany

Department of Law, Economics, and Business,
University of Angers, Angers, France

Abstract

Central banks evolved in Europe in the seventeenth and eighteenth centuries as centralized monetary authorities that often served the purpose of financing governments. They were instrumental in the transition from classical commodity-backed currencies to the global fiat money system of the present day. The role and functioning of central banks have changed substantially over time. During the classical gold standard, their main responsibility was to store and exchange gold in correspondence to the currencies in circulation. Today, they play a much more active political role in managing the money supply through various policy instruments. They follow different and sometimes conflicting goals, including price stability, stimulation of economic growth, and cutting unemployment rates. A lively debate has arisen on weighing out different policy goals and the proper role of central banking in the economy. This debate does not lack sharp criticisms, both on purely economical and ethical grounds.

Definition

A central bank is an institution that possesses a legal monopoly over the creation of money. It conducts monetary policies and manages the supply of credit and money in a given territory – a country or a union of countries.

The History, Role, and Critique of Central Banking

Central banks – although often considered to be politically independent – are the monetary

authorities of states or unions of states. In contrast to commercial banks, they possess legal monopolies over the issuance of money in given territories. They control and manage the supply of credit and money, which usually exhibits legal tender status, through monetary policy tools, such as open market operations, discount window lending, and changes in reserve requirements. Examples of central banks are the *Federal Reserve System* (FED) in the United States, the *Bank of England*, and the *European Central Bank* (ECB) in the eurozone.

According to Bordo (2007), there are three key goals of modern monetary policy. These are price stability or stability of the value of money; economic stability, that is, smoothing business cycles by offsetting shocks to the economy; and financial stability which essentially means granting credit to commercial banks suffering from liquidity shortages as a *lender of last resort*. Historically, the importance of these goals and the goals as such have changed substantially, along with the monetary regimes within which central banks were acting.

A Brief History of Central Banking

In the following section, a brief historical sketch of the development of central banks as monopolists over the supply of money and credit and the economic debates around the subject is presented.

Historically, moneys have been commodities, most often precious metals such as silver and gold. So it was under commodity money regimes that the first institutions resembling central banks were established. According to Glasner (1998, p.27), banking in general “evolved because it provided two services that proved to be strongly complementary: provision of a medium of exchange and intermediation between borrowers and lenders.” Instead of using precious metals directly as coins, people could use money substitutes that represented claims to fixed amounts of precious metals that were stored in the banks’ vaults.

Bank money was less costly than coins for several reasons. Holders of deposits often received interest and bore no losses from the wear and tear of coins.

They also bore no costs of transporting coins or of protecting them against theft or robbery. And, when making transactions, they could avoid the costs of counting, weighting, and inspecting coins. (Glasner 1998, p. 28)

Obviously, banks were put in a very powerful position. There are several historical examples in which these powers were misused. Italian bankers in England, for example, financed the military expenses of Edward I around 1300, when he faced social upheavals in Wales and Scotland. With the financial support of the Italian bankers, he could gather a much greater army than his predecessors were ever able to (Prestwich 1979).

The temptation for banks to issue more money substitutes than there were precious metals in the vaults was compelling. So the municipal bank of Barcelona, for example, had to suspend convertibility into specie during the Catalanian fight for independence in Spain in the fifteenth century (Usher 1943). In the sixteenth and seventeenth centuries, private but highly regulated banks in Venice financed war expenses by creating deposits against government debt. Due to high government demand for funds, the convertibility had to be suspended and the bank money depreciated against precious metals (de Roover 1976; Lane 1937).

In fact, on the way to complete monopolization of the money supply, military conflicts within and between states played a fundamental role. One of the first institutions that are commonly considered to be central banks is the Swedish *Riksbank* founded in the 1660s. It is also the oldest still existing central bank in the world. Another early example is the *Bank of England* founded in the 1690s. The Swedish authorities tried to prevent interference and misuse of the King, whereas the *Bank of England* was specifically founded with the objective of financing the ongoing conflicts with France. It was only through the establishment of this institution that William III was able to borrow £1,200,000, half of which was used to rebuild the navy. The *Banque de France* is another case in point. It was established after the French Revolution and just before the Napoleonic Wars (1803–1815) in 1800. Malpractices of private banks as well as increasing threats from

aggressive foreign countries set strong incentives for the establishment of centralized monetary authorities in other European countries and around the world. All that happened mostly under more or less binding silver, gold, or bimetallic standards, which restricted the powers of central banks as long as convertibility of bank notes into specie and the trust of the people in the currency were not to be jeopardized lightly. However, not before 1816, shortly after the Napoleonic Wars, during which convertibility was suspended, has the British pound for the first time been legally defined as a fixed weight of gold, although Britain had been on a de facto gold standard since 1717, due to an overvaluation of gold relative to silver by Sir Isaac Newton who was at that time Master of the Royal Mint.

In 1844, the *Bank Charter Act* reinforced the gold backing of the pound and marked the beginning of a period known as the classical gold standard that lasted until 1914. Also other currencies such as the franc, the mark (since 1871), and the US dollar were legally defined as fixed amounts of gold. Therefore, the classical gold standard was a period of fixed exchange rates between currencies, which lowered risks and fostered international trade, as has been suggested in various empirical studies (see, e.g., López-Córdova and Meissner 2003 or Flandreau and Maurel 2001).

What is today called the time-inconsistency problem of central banking might have played an important role in this development. On the one hand, the monetary authorities could have exploited their monopoly status on a slow but constant basis by inflating the money supply at a modest rate. On the other hand, this would lower the profit-earning potential under emergencies, such as military invasions or economic crises. In order to effectively exploit the monopoly status in emergencies, it is necessary to build up confidence in the stability and soundness of the currency among the general public during normal times. The authorities have to make a credible commitment to price stability (Kydland and Prescott 1977). The developments toward the universal gold standard can be interpreted as such a commitment (Bordo and Kydland 1995). On the other hand, it might be interpreted as a restriction on

governmental despotism and as a political effect of the spread of classical liberal philosophies and enlightenment ideas that generally endorsed limited state powers. The most important function of central banks in that era was to store and exchange gold reserves in correspondence to the national currencies in circulation.

Although there already existed institutions that resembled central banks in the United States in the nineteenth century, namely, the *First* (1790–1811) and the *Second Bank of the United States* (1816–1836), the *Federal Reserve System* as we know it today was only founded in 1913. Shortly after, with the beginning of the Great War (1914–1918), the era of the classical gold standard ended. The United States which entered the war in 1917 inflated their currency less than the European nations and was thus the only country that de jure remained on a gold standard throughout this period. The German, French, and British currencies depreciated substantially with respect to the US dollar and gold. They returned to gold in the 1920s at an artificially overvalued rate which led to constant complaints about gold or liquidity shortages, particularly in Britain, but also in other European countries (Rothbard 1998). The gold exchange standard that was established during the interwar period was essentially a pound sterling standard, since only the British pound sterling was redeemable in gold. Central banks of other European countries mostly held pounds as reserves. The United States was virtually the only country that remained on the gold standard in the classical sense. During the *Great Depression* in the 1930s, many countries suspended convertibility again. After the catastrophe of the Second World War, with the establishment of the Bretton Woods system in 1945, the British pound lost its status as a reserve currency to the US dollar. The *International Monetary Fund* (IMF) and the *International Bank for Reconstruction and Development* (IBRD), which today is part of the *World Bank*, were founded. The US dollar remained convertible into gold on a fixed rate of \$35 per fine ounce. Central banks of other countries were responsible for keeping their currencies in a fixed relationship to the US dollar. Only central banks, but no private citizens, were able to hand in US

dollars for redemption at the FED. The system lasted until the fifteenth of August 1971, when President Richard Nixon eventually suspended convertibility of the US dollar in a unilateral act. Since then, the global financial system is based on fiat money, that is, money which derives its value from government law and is independent of any commodity.

During the transition period from the classical gold standard to a fiat money system, several economic and political arguments against the gold standard and in favor of fiat currencies have been brought forward. First of all, it would have facilitated the financing of political projects, such as the *New Deal* under Franklin D. Roosevelt in response to the *Great Depression*. Furthermore, it would not have been necessary to increase interest rates in 1931 to maintain convertibility of the US dollar, after Britain was already forced to suspend convertibility (Romer 2003). A policy of low interest rates can only be maintained as long as needed under fiat currencies. Since the money production is not restricted by the production of any commodity, it renders monetary policy much more flexible. The Keynesian theory of business cycles suggests that economic downturns can be cured by stimulating aggregate demand through expansive monetary policy in combination with deficit spending (Keynes 1936; Krugman 2006). There have, of course, been critics of this view (see, e.g., Hazlitt 1959), but it is today accepted by the majority of economists that central banks should intervene more actively into the economy by adjusting the money supply appropriately. In doing so, they should consider and control, as far as possible, macroeconomic aggregates, such as price inflation, unemployment, and economic growth. Often, objectives of monetary policy are in conflict. Stimulating aggregate demand and economic growth through expansive monetary policies, for example, is in conflict to the general goal of price stability. Expanding the money supply can only be done at the risk of higher price inflation. Therefore, conflicting policy goals have to be weighed out. Still today, there is a lively debate on which policy instruments should be applied for which purposes and whether central bank policies should be conducted passively, based on rules, or

whether we should adopt a discretionary and more flexible central bank policy.

The Role and Functioning of Central Banks Today

Theoretically speaking, there exist two types of money in our financial system: base money and commercial bank money. Base money, the core of the money supply, is created by central banks whenever they buy assets or extend credit to commercial banks. It can be destroyed when central banks sell assets or when credit is repaid. According to Belke and Polleit (2009, p. 29), “[c]ommercial banks need base money for at least three reasons: (i) making inter-bank payments; (ii) meeting any *cash drain* as non-banks want to keep a portion of their deposits in cash (notes and coins); and (iii) holding a certain portion of their liabilities in the form of base money with the central bank (*minimum reserves*).” The minimum reserve requirement is one instrument of monetary policy to regulate the overall money supply in the economy. Minimum reserves are held in cash directly at the commercial bank or as deposits with the corresponding central bank. Given a reserve requirement of 1% as currently in the eurozone, for any unit of base money, a commercial bank can create commercial bank money by extending loans of up to 99 units to the private sector (*excess reserves*). Hence, the maximum money creation potential of the banking sector is determined by the central bank through minimum reserve requirements.

The primary tools of central banks to manage the supply of base money are open market operations (Belke and Polleit 2009, p. 33). Their implementations work similarly in the United States and in the eurozone. However, minor differences can be observed. In the United States, the *Domestic Trading Desk* (Desk) arranges open market operations on a daily basis. It has to figure out whether there are imbalances between supply and demand for base money and react accordingly. Imbalances are indicated by differentials between the effective interest rate at which base money is borrowed and lent on the interbank

market and the target interest rate set by the *Federal Open Market Committee*. Usually, short-lived imbalances are corrected through *temporary* operations. In special cases, when imbalances turn out to be more persistent than expected, the Desk may perform *outright* operations. These operations involve the buying and selling of government bonds on the *secondary market*, that is, the market in which formerly issued government bonds are traded. This means that, under normal circumstances, the FED does not buy bonds directly from the government. Outright open market operations affect the base money supply permanently, whereas the much more common temporary open market operations will unwind after a specified number of days. The latter are combined with *repurchase agreements*. For example, the Desk may decide to increase the base money supply and buys government securities from commercial banks. In order to keep this increase temporary, it agrees to resell those securities to its counterparties on a future date. *Matched sale-purchase transactions* are the tool with which the base money supply is temporarily decreased. Securities are first sold and will then be bought back in the future. The Desk may also redeem maturing securities, rather than replacing them with new ones, and can thereby reduce the portfolio without entering the market directly (Edwards 1997).

Similar to the Desk, the ECB mostly uses reverse transactions, that is, buying or selling eligible assets under repurchase agreements or lending money against eligible assets provided as collateral (Belke and Polleit 2009, p. 43). These transactions are used for *main refinancing operations* with a maturity of usually one week as well as *longer-term refinancing operations* with a maturity of usually three months. The ECB may use *fine-tuning operations* in reaction to unexpected liquidity fluctuations to steer interest rates in the form of *outright transactions* and *foreign exchange swaps*. The latter are spot and forward transactions in euro against foreign currencies. Furthermore, the ECB manages its structural position vis-à-vis the financial sector by issuing *ECB debt certificates* (ECB 2006). One of the main policy objectives of the ECB is to control short-

term interest rates and reduce their volatilities. The *marginal lending facility* and the *deposit facility* are always available for credit institutions on their own initiative, whenever there is a lack of trading partners on the money market. Interest rates at the lending facility are usually higher than on the money market. The deposit facility usually offers lower interest rates. Those two institutions therefore provide boundaries within which interest rates on the overnight money market fluctuate. There is no limit on the access to these facilities other than collateral requirements at the lending facility.

Central banks generally have to decide on which policy instruments to use. In a simplified version, the problem can be seen as a choice between setting interest rates and letting the money supply be determined endogenously or the other way around. In practice, as Blinder (1998) points out, there are many more choices to be made, including various definitions of the money supply, several different choices for interest rates, bank reserves, and exchange rates. The practical problems involved might be more complicated than the theory suggests.

The intellectual problem is straightforward in principle. For any choice of instrument, you can write down and solve an appropriately complex dynamic optimization problem, compute the minimized value of the loss function, and then select the *minimum minimorum* to determine the optimal policy instrument. In practice, this is a prodigious technical feat that is rarely carried out. And I am pretty sure that no central bank has ever selected its instrument this way. But, then again, billiards players may practice physics only intuitively. (Blinder 1998, pp. 26–27)

As the above quote suggests, there is usually a clear discrepancy between the theoretically optimal and the practically possible. One of the various controversial subjects, when it comes to monetary policy in practice, is the question whether political actions should be rule based or discretionary. It has been argued that central banks left with discretion tend to err systematically in the direction of too much inflation. In order to correct this bias, one needs more or less strict rules (Kydland and Prescott 1977). Simons (1936) favored a strict commitment of the FED to price

stability, that is, zero inflation, rather than pursuing any other possible policy goals. Friedman (1959) and other economists in the monetarist tradition advocated a low but constant growth rate of the money stock. Yet another, even more restrictive, rule was proposed by Wallace (1977). He argued that the FED should consider holding the money stock constant. Such a rule would essentially end monetary policy altogether. This view has not found many adherents. Instead, a modern trend in central banking is inflation rate targeting that only evolved after the end of the Bretton Woods system, in which exchange rates were targeted. Usually, inflation targets are given in a more or less narrow range around 2%. Allowing some inflation to take place within a certain range provides more flexibility to pursue other policy goals. The target can be met through adjusting interest rates appropriately. Whenever inflation rates are below the target range, interest rates can be lowered and vice versa. A declared inflation target also makes central bank policy more transparent. Investors can more easily anticipate possible changes in interest rates. This may lead to an overall stabilization of the economy. However, in the course of the current financial crisis (since 2008), inflation targeting has been abandoned in order to intervene more actively into the economy by means of more expansionary monetary policy. Interest rates are lower than ever before, providing liquidity for credit institutions on the financial markets at the risk of higher inflation rates. The reactions on the current crisis have been criticized on very different grounds. For some economists, they are still too conservative. For others, they constitute only a treatment of the symptoms, rather than the causes. In their view, they will prolong the problem instead of solving it. Central banking practices in general have been criticized, both on purely economical and ethical grounds.

Critique of Central Banking

A first very general point of criticism lies in the state-granted monopoly status of central banks. What justifies a legal monopoly over the supply

of money? There has been a lively debate over the question whether money is a natural monopoly. For this to be the case, according to the traditional definition, the production of money should exhibit economies of scale, which means that the average costs of production for one firm producing the whole output are always lower than if two or more firms with access to the same technology divide the output (Glasner 1998, p. 23). However, the fact that central banks under today's fiat money regime produce money at almost zero production costs does not imply economies of scale. Anyone could produce an unbacked or digital money at almost zero production costs.

The demand side characteristics of money also fall short of justifying its legal monopoly. Even if it is theoretically beneficial and cost reducing to use only one universally accepted medium of exchange within a given area and even if in fact only one universally accepted medium of exchange would emerge among freely cooperating individuals, for example, gold, there is no justification for legally restricting the production of that medium to one authorized institution. Evidently, for the functioning of a fiat money regime, it is necessary to restrict production to one institution, since otherwise competition would lead to an excessive expansion of the money supply and a rapid devaluation until its value reaches production costs – essentially zero (Hoppe 2006). But as mentioned above, the fiat money regimes that govern today's global economy are themselves the result of state interventions into traditional commodity standards. Therefore, the mere existence of a fiat money regime cannot justify the legal monopoly per se. Furthermore, to consider money to be a public good is false, since it lacks the criteria of non-rivalness and non-excludability (Vaubel 1984). If, however, the money production is legally monopolized by establishing central banks, then the general economic analysis of monopolies should be applicable to it. In general, monopolies are considered to be economically inefficient and costly. There is an omnipresent danger that the monopoly status is irresponsibly exploited at the expense of the public.

Central banks in a fiat money regime are able to create as much money as they please and can

serve as a *lender of last resort* for banks that lack liquidity. Within such an environment, an incentive for commercial banks is set to operate under a lower equity ratio, to hold less money, and to engage in riskier projects, as they can always borrow money at relatively low interest rates from the central bank. Hence, a *moral hazard* problem arises. As a result, the financial system as a whole becomes more fragile and susceptible to crises. The same analysis holds for governments. The European debt and financial crisis is a dramatic case in point (Bagus 2012).

Opposed to the Keynesian theory that ultimately monetary spending determines economic progress, the *Austrian Theory of the Business Cycle* advises against artificial credit expansion through lowering interest rates, which is easier to accomplish than ever before under a fiat money regime controlled by central banks. This theory, introduced by Ludwig von Mises (Mises 1912) and further developed by Friedrich August von Hayek (Hayek 1935), sees the root cause of economic depressions in an imbalance between investments and real savings brought about through this very process of credit expansion. In the Austrian view, the interest rate is not an arbitrary number that should be interfered with. Instead, it is the price that tends to accommodate the roundaboutness of production processes or investment projects to the available subsistence fund in the economy. Interest rates tend to fall when consumers save more and thereby increase the subsistence fund. In these situations, more roundabout investment projects can be sustained. If, however, interest rates are lowered artificially, the subsequent excess investments are not covered by real savings. At least some of the *malinvestments* have to be liquidated when the imbalance becomes evident. This situation constitutes the economic bust.

Another point of criticism lies in the inflationary tendencies of the modern monetary system under central bank control. Hyperinflation rates of more than 100%, as in the Weimar Republic or more recently in Zimbabwe, have obviously devastating effects. But also, moderate inflation rates are not neutral. They can be interpreted as a tax

that enables governments to pursue policy goals that lack democratic legitimization (Hülsmann 2008, p. 191). State control over money was the result of the “characteristic quest by the state for sources of revenue” (Glasner 1998, p. 24). Fiat inflation therefore leads to an excessive growth of the state for which the citizens do not pay directly through taxes but rather indirectly through a devaluation of the currency they use. The redistributive effects of inflation are known as Cantillon effects (Cantillon 2010). Since inflation does not take place uniformly, but rather gradually ripples through the economy, the first receivers of the newly created money benefit on the expense of all others, since they can buy goods on the market for still relatively low prices. As they spend the money, prices tend to rise. Late receivers and people on fixed incomes face price increases before or entirely without increases in their incomes. They suffer a loss in real terms. Usually, the first receivers and beneficiaries of newly created money are commercial banks, other financial institutions, governments, and closely related industries. It is argued that the general public carries the burden. Although some groups doubtlessly benefit from the inflationary tendencies of the fiat money system, some economists argue that it cannot benefit society as a whole. The mere possibility to position oneself on the winner side leads to some kind of “collective corruption” and the maintenance of the system (Polleit 2011).

References

- Bagus P (2012) The tragedy of the Euro, 2nd edn. Ludwig von Mises Institute. Available Online: <http://mises.org/document/6045/The-Tragedy-of-the-Euro>
- Belke A, Polleit T (2009) Monetary economics in globalised financial markets. Springer, Berlin/Heidelberg
- Blinder AS (1998) Central banking in theory and practice (the Lionel Robbins lectures). Massachusetts Institute of Technology, Cambridge
- Bordo MD (2007) A brief history of central banks, Federal Reserve Bank of Cleveland – economic commentary. Available Online: <http://www.clevelandfed.org/research/commentary/2007/12.cfm>
- Bordo MD, Kydland FE (1995) The gold standard as a rule: an essay in exploration. *Explor Econ Hist* 32:423–464
- Cantillon R (2010) An essay in economic theory. Ludwig von Mises Institute. Available Online: <http://mises.org/document/5773/An-Essay-on-Economic-Theory>
- de Roover R (1976) Chapter 5: New interpretations in banking history. In: Business, banking, and economic thought in the Middle Ages and early modern Europe. University of Chicago Press
- ECB (2006) The implementation of monetary policy in the Euro area, general document of eurosystem monetary policy instruments and procedures. Available Online: <http://www.ecb.europa.eu/pub/pdf/other/gendoc2006en.pdf>
- Edwards CL (1997) Open market operations in the 1990s. Federal Reserve Bulletin, pp 859–874. Available Online: <http://www.federalreserve.gov/pubs/bulletin/1997/199711lead.pdf>
- Flandreau M, Maurel M (2001) Monetary union, trade integration and business fluctuations in 19th century Europe: just do it. Centre for Economic Policy Research working paper no. 3087
- Friedman M (1959) A program for monetary stability. Fordham University Press, New York
- Glasner D (1998) An evolutionary theory of the state monopoly over money, in money and the nation state – the financial revolution, government and the world monetary system. The Independent Institute, Oakland
- Hazlitt H (1959) The failure of the “New Economics” – an analysis of the Keynesian Fallacies. D. van Nostrand. Available Online: <http://mises.org/document/3655/Failure-of-the-New-Economics>
- Hoppe H-H (2006) How is fiat money possible? – or, the devolution of money and credit, published in the economics and ethics of private property – studies in political economy and philosophy, 2nd edn. Ludwig von Mises Institute. Available Online: <http://mises.org/document/860/Economics-and-Ethics-of-Private-Property-Studies-in-Political-Economy-and-Philosophy-The>
- Hülsmann JG (2008) The ethics of money production. Ludwig von Mises Institute. Available Online: <http://mises.org/document/3747/The-Ethics-of-Money-Production>
- Keynes JM (1936) The general theory of employment, interest and money, Macmillan Cambridge University Press. A revised version is Available Online: <http://www.marxists.org/reference/subject/economics/keynes/general-theory/>
- Krugman P (2006) Introduction to the general theory of employment, interest, and money, by John Maynard Keynes. Available Online: <http://www.pkarchive.org/economy/GeneralTheoryKeynesIntro.html>
- Kydland FE, Prescott EC (1977) Rules rather than discretion: the inconsistency of optimal plans. *J Polit Econ* 85(3):473–492
- Lane FC (1937) Venetian bankers, 1496–1533: a study in the early stages of deposit banking. *J Polit Econ* 45(2):187–206

- López-Córdova JE, Meissner CM (2003) Exchange-rate regimes and international trade: evidence from the classical gold standard era. *Am Econ Rev* 93(1): 344–353
- von Mises L (1912) *Theorie des Geldes und der Umlaufsmittel*. Verlag von Duncker und Humblot, München und Leipzig. Available Online: <http://mises.org/document/3298/Theorie-des-geldes-und-der-Umlaufsmittel>; for the English edition see: <http://mises.org/document/194/The-Theory-of-Money-and-Credit>
- Polleit T (2011) Fiat money and collective corruption the quarterly. *J Aust Econ* 14(4): 297–415. Available Online: https://mises.org/journals/qjae/pdf/qjae14_4_1.pdf
- Prestwich M (1979) *Italian merchants in late thirteenth and fourteenth century england, in the dawn of modern banking*. Yale University Press, New Haven
- Romer CD (2003) Great depression, forthcoming in the *encyclopædia britannica* Available Online: http://eml.berkeley.edu/~cromer/great_depression.pdf
- Rothbard MN (1998) *The gold exchange standard in the interwar years, in money and the Nation State – The Financial Revolution, Government and the World Monetary System*, The Independent Institute, Oakland. Also published in *A history of money and banking in the United States – the Colonial Era to World War II* and Available Online: <http://mises.org/document/1022/History-of-Money-and-Banking-in-the-United-States-The-Colonial-Era-to-World-War-II>
- Simons HC (1936) Rules versus authorities in monetary policy. *J Polit Econ* 44(1):1–30
- von Hayek FA (1935) *Prices and production*, 2nd edn. Augustus M. Kelley, New York Available Online: <http://mises.org/document/681/Prices-and-Production>
- Usher AP (1943) *The early history of deposit banking in mediterranean Europe*. Harvard University Press, Cambridge
- Vaubel R (1984) The Government's money monopoly: externalities or natural monopoly. *Kyklos* 37(1): 27–58
- Wallace N (1977) Why the fed should consider holding M0 constant. *Q Rev Federal Reserve Bank of Minneapolis, Minneapolis, MN* 1(1):2–10

Certification Labeling

► Labeling

Change of Circumstances

► Impracticability

Child Maltreatment, The Economic Determinants of

Jason M. Lindo^{1,2,3} and Jessamyn Schaller⁴

¹Texas A&M University, College Station, TX, USA

²NBER, Cambridge, MA, USA

³IZA, Bonn, Germany

⁴The University of Arizona, Tucson, AZ, USA

Abstract

This entry examines the economic determinants of child maltreatment. We first discuss potential mechanisms through which economic factors, including income, employment, aggregate economic conditions, and welfare receipt, might have causal effects on the rates of child abuse and neglect. We then outline the main challenges faced by researchers attempting to identify these causal effects, emphasizing the importance of data limitations and potential confounding factors at both the individual and aggregate levels. We describe two approaches used in the existing literature to address these challenges – the use of experimental variation to identify the effects of changes in family income on individual likelihood of maltreatment and the use of area studies to identify the effects of changes in local economic conditions on aggregate rates of maltreatment.

Definition

The economic determinants of child maltreatment refer to the broad set of economic factors that have causal effects on the rates of child abuse and neglect, either directly or indirectly, potentially including income, employment, aggregate economic conditions, and welfare receipt.

Introduction

Child maltreatment, including physical abuse, sexual abuse, emotional abuse, and neglect, is a

prevalent and serious problem. In the United States alone, more than six million children are involved in reports to Child Protective Services (CPS) annually, while countless more are subject to unreported maltreatment (Petersen et al. 2014). Child maltreatment has severe and lasting consequences for victims, injuring physical and mental health and affecting interpersonal relationships, educational achievement, labor force outcomes, and criminal behavior (see, e.g., Gilbert et al. 2009; Berger and Waldfogel 2011). Child maltreatment is costly to society as well, generating productivity losses, increased burdens on criminal justice systems and special education programs, and substantial costs for child welfare services and health care (Fang et al. 2012; Gelles and Perlman 2012).

Given the pervasive and damaging nature of the problem, it is not surprising that a substantial literature spanning many disciplines and several decades is devoted to identifying the causes of child maltreatment. (For a summary of this literature, see Petersen et al. (2014).) Within this literature, a variety of economic factors, including family income, parental employment, macroeconomic conditions, and welfare receipt, have been identified as predictors of child abuse and neglect (Pelton 1994; Stith et al. 2009; Berger and Waldfogel 2011). Yet, due to data limitations and identification challenges, researchers have only recently begun to make progress isolating the causal effects of these variables on maltreatment.

This entry is devoted to the economic determinants of child maltreatment. We begin with etiological theories of child maltreatment from the fields of psychology and economics, outlining the potential mechanisms by which different economic factors might be correlated with child abuse and neglect at the individual and aggregate levels. Next, we describe different types of data used in the study of child maltreatment and discuss their limitations. We then discuss the additional challenges that maltreatment researchers face in estimating the causal effects of economic conditions, the empirical approaches that researchers have taken to try to overcome these challenges, and the lessons learned from these studies before concluding.

Theory and Mechanisms

The most commonly cited etiological models of child maltreatment are the developmental-ecological and ecological-transactional models originating in psychology (Garbarino 1977; Belsky 1980; Cicchetti and Lynch 1993). These models posit that maltreatment results from complex interactions between individual, familial, environmental, and societal risk factors. Among the risk factors for maltreatment in these models, economic variables, such as family income and parental employment status, have garnered particular attention in the literature, both because they are robust, easily measured predictors of maltreatment and because they can be manipulated through policy intervention. However, as ecological models posit that maltreatment results from *interactions* between economic variables and characteristics of individuals, families, and communities, these models do not generate clear predictions about how economic factors should be correlated with maltreatment. (For example, the effect of a stressful life event such as a reduction in family income on the likelihood of maltreatment may be exacerbated by individual characteristics such as depression while also being mitigated by social support and other buffering factors (National Research Council 1993).)

Economists have approached theoretical modeling of child maltreatment from a different perspective, seeking to understand child maltreatment within a framework of budget constraints and utility functions. Several empirical investigations of child maltreatment, including those of Paxson and Waldfogel (2002), Seiglie (2004), Berger (2004, 2005), and Lindo et al. (2013), have been motivated by theoretical models of investments in child quality, sometimes in combination with altruistic, cooperative bargaining, and noncooperative bargaining models used in economic studies of marriage and divorce, family labor supply, and domestic partner violence. There is also overlap between theoretical models of child maltreatment and economic models of criminal behavior. (Berger (2004, 2005) provides a nice summary of several theoretical economic models relevant to the analysis of child abuse and

neglect. To our knowledge, the only study with formal model of child maltreatment is Seiglie (2004), which builds on economic models of investment in child quality.)

In developing a theoretical framework for understanding the oft-observed link between poverty and maltreatment, it is important to distinguish between reasons child maltreatment might be associated with poverty and causal pathways through which economic variables might affect the incidence of abuse and neglect. For example, parental education, community norms with regard to parenting behaviors, parental history of abuse, and innate personality characteristics of parents have all been cited as important factors that could explain some (or potentially all) of the association between poverty and child maltreatment. In thinking about the *causal pathways* through which economic factors may affect child maltreatment, it may be useful to imagine a hypothetical experiment in which a household is randomly selected to receive an intervention such as a cash transfer, an unanticipated job displacement, or a change in aggregate economic conditions and to consider the effects of this treatment on the likelihood that the children in that household will experience abuse or neglect. With these types of experiments in mind, researchers have identified a number of potential pathways through which these economic “treatments” might influence the likelihood of child abuse and neglect. (In this section we focus on the relationship between economic factors and the likelihood of committing maltreatment rather than the likelihood of being reported, investigated, or punished for abuse. We discuss issues related to reporting and data quality in the next section.)

First, income may have direct effects on the likelihood of maltreatment if parents are constrained in their ability to provide sufficient care for their children (Berger and Waldfogel 2011). This mechanism is particularly relevant to the study of child neglect, which is commonly defined as the failure of a caregiver to provide for a child’s basic physical, medical, educational, or emotional needs, and thus is often considered to be “underinvestment” in children within the context of economic models (see, e.g., Seiglie 2004).

(Weinberg (2001) notes that family income may be directly associated with abuse as well, as it relates to the availability of resources that can be used to elicit desired behavior from children.)

Changes in the amount and sources of family income may also affect child maltreatment by altering the distribution of bargaining power within households and changing the expected cost of abuse. Building on bargaining models used in economic studies of domestic violence, Berger (2005) posits that, in two-parent households, shifts in the distribution of family income away from the perpetrator of abuse and toward a non-abusing partner can result in a shift in the balance of power within the relationship, which can in turn affect the incidence of maltreatment. Additionally, as in economic models of criminal behavior, income shocks can affect the cost that the perpetrator of maltreatment expects to incur if he/she is caught. Specifically, the perpetrator’s access to income is jeopardized if maltreatment leads to dissolution of a relationship and loss of access to a partner’s income. The removal of a child can also lead to the loss of child-conditioned transfers such as welfare payments and child support.

Economic shocks may also affect rates of child abuse and neglect through their impacts on mental health. At the aggregate level, research has shown that economic downturns are associated with deterioration of population mental health, as measured by the incidence of mental disorders, admissions to mental health facilities, and suicide (Zivin et al. 2011). Job displacement has also been linked to a number of mental-health-related outcomes, including psychological distress (Mendolia 2014), depression (Brand et al. 2008; Schaller and Stevens 2014), psychiatric hospitalization (Eliason and Storrie 2010), and suicide (Eliason and Storrie 2009; Browning and Heinesen 2012). Meanwhile at the individual level, a large literature documents a correlation between poverty and mental health in the cross-section. However, empirical evidence on the causal effects of individual and family income on mental health is sparse and inconclusive. (Several papers have examined mental health outcomes of lottery winners, with mixed results (e.g., Kuhn et al. 2011; Apouey and Clark 2014).)

Substance abuse and partnership dissolution may also mediate the relationship between economic shocks and child maltreatment. Alcohol and drug use and single parenthood are both correlated with socioeconomic status and are also well-known risk factors for child abuse and neglect. However, the causal links between economic shocks and these variables are not well understood. (For example, Deb et al. (2011) identify heterogeneity in the response of drinking behavior to job displacement and the empirical evidence on the effects of aggregate economic downturns on alcohol consumption is mixed (Ruhm and Black 2002; Dávalos et al. 2012). Meanwhile, while layoffs lead to increased divorce rates in survey data (Charles and Stephens Jr 2004; Doiron and Mendolia 2012), aggregate divorce rates are found to decrease in recessions (Schaller 2013).)

Finally, parental time use is a rarely mentioned mechanism by which economic shocks can affect maltreatment. In particular, involuntary changes in employment and work hours have the potential to affect the incidence of maltreatment through their effects on the amount of time children spend with parents, other family members, childcare providers, and others (Lindo et al. 2013). This mechanism may work in different directions depending which parent experiences the employment shock and on the type of maltreatment considered. (A shock that shifts the distribution of childcare from the mother to the father may increase the incidence of abuse since males tend to have more violent tendencies than females. As another example, additional time at home with a parent may reduce the likelihood of child neglect but increase the likelihood of physical, sexual, and emotional abuse.)

Identifying Causal Effects

Identifying the causal effects of economic factors on child maltreatment requires (i) child maltreatment data linked to measures of economic conditions and (ii) empirical strategies that can isolate the effects of economic factors despite the fact that these factors tend to be correlated with other

determinants of maltreatment. Both of these issues present challenges for researchers that are difficult – though not impossible – to overcome.

Data

Data Based on Maltreatment Reports

Child abuse reports have historically been the primary source of data for researchers interested in studying child maltreatment on a large scale. While these data are attractive because they often span large areas and many time periods, a natural concern is that maltreatment report data may not accurately reflect the true incidence of maltreatment. While there is no doubt that false reports are sometimes made, the consensus view is that statistics tend to understate the true prevalence of child abuse because underreporting is such a serious issue (Waldfoegel 2000; Sedlak et al. 2010). In fact, the Fourth National Incidence Study of Child Abuse and Neglect (NIS-4), which identifies maltreated children outside of the United States Child Protective Services (CPS) system, found that CPS investigated the maltreatment of only 32% of children identified in the study as having experienced observable harm from maltreatment. Applying CPS screening criteria to the maltreatment cases that were not investigated by CPS, the researchers concluded that underreporting was the primary reason for this low rate of investigation: three quarters of the cases would have been investigated if they had been reported to CPS (Sedlak et al. 2010).

Nonetheless, reports are likely to be strongly related to the true incidence of maltreatment and thus may serve as a useful proxy. The key consideration with the use of any proxy variable is the degree to which the measurement error is the same across comparison groups. If a comparison is made across groups that have the same degree of measurement error (or across time periods that have the same degree of measurement error), then the percent difference in the proxy will be identical to the percent difference in the variable of interest. For example, if State A has 1,200 maltreatment reports and State B has 800 maltreatment reports and the true incidence of maltreatment is understated in both states by 20%, then the

percent difference in reports $((1,200-800)/800 \times 100\% = 50\%)$ will be equal to the percent difference in the true incidence of maltreatment $(1,200 \times 1.2 - 800 \times 1.2)/800 \times 1.2 \times 100\% = 50\%)$.

Given that estimating the causal effects of economic factors on child maltreatment will inevitably entail comparisons across groups and/or time periods, this discussion naturally raises the question: is it generally safe to assume that the measurement error in abuse reports is the same across groups and across time? Unfortunately for researchers, while this assumption may hold in certain circumstances, it is unlikely to hold in most instances. When making comparisons across states, we must address the fact that states differ in how they define abuse, who is required to report abuse, and in how they record and respond to reports of abuse. When making comparisons across time, we must acknowledge that children's exposure to potential reporters and individual propensities to report maltreatment may be changing over time and that the rate of reporting may in fact even be correlated with economic factors. Moreover, states have periodically changed their official definitions of abuse, reporting expectations, and standards for screening allegations. As such, comparisons of abuse reports across states and time have the potential to reflect differences in measurement error in addition to differences in the incidence of maltreatment. Comparisons across groups defined in other ways will be susceptible to similar issues. For example, the maltreatment of infants and toddlers may be less likely to be detected than the maltreatment of school-aged children who spend more time in the presence of mandatory reporters.

It is also important to note that focusing on substantiated reports does not necessarily improve our ability to make valid comparisons – and could actually make things worse – even in a scenario in which agencies are perfectly able to discern true and false reports. Comparisons of substantiated reports (in percent terms) will do better than comparisons of all reports if and only if the *difference* in the measurement error in substantiated reports across groups is less than the *difference* in the measurement error in overall reports across groups, which may not be the case. (Here the

measurement error we refer to is the degree to which the variable differs from what we would like to measure: true incidents. As an example in which we would do worse by focusing on substantiated reports, suppose State C has 2,500 true incidents, 40% of which are reported, and 5 false reports per 100 true incidents, while State D has 2,000 incidents, 35% of which are reported, and 10 false reports per 100 true incidents. Then, assuming true reports are substantiated and false reports are not substantiated, the percent difference in reports would correctly identify the true percent difference in incidents, whereas the percent difference in substantiated reports would not, as the true percent difference $= (2,500-2,000)/2,000 \times 100\% = 25\%$, the percent difference in reports $= [2,500 \times (40\% + 5\%) - 2,000 \times (35\% + 10\%)]/2,000 \times (35\% + 10\%) \times 100\% = 25\%$, and the percent difference in substantiated reports $= (2,500 \times 40\% - 2,000 \times 35\%)/2,000 \times 35\% \times 100\% = 43\%$.)

The major take-away from this discussion is that we must take into consideration the process by which maltreatment that occurs becomes observable to the researcher. In particular, when a researcher estimates the causal effect of an economic factor on the observed incidence of maltreatment, we must consider the degree to which the effects are driven by actual changes in maltreatment and/or by changes in the rate at which occurrences of maltreatment are detected and reported.

Alternative Sources of Data

Survey data, hospital data, and internet search data have also been used to gain insights into the prevalence of maltreatment and the way it varies with economic factors. Cross-sectional surveys include retrospective questionnaires that solicit information on occurrences of maltreatment over one's childhood or within a specific time window, while panel surveys solicit information on a year-to-year basis. Hospital data can be used to measure maltreatment using diagnosis codes that explicitly indicate maltreatment or by considering outcomes that are expected to be highly correlated with maltreatment (e.g., accidents, shaken-baby syndrome, etc.), as in Wood et al. (2012). And

internet search data can be used to measure the frequency with individuals are searching for phrases that are expected to be highly correlated with maltreatment (e.g., child protective services, dad hit me, etc.), as in Stephens-Davidowitz (2013).

While all of these sources of data have the potential to shed new light on maltreatment in ways that administrative reports data cannot, they are also susceptible selection bias. Just as economic factors may affect both the incidence of maltreatment and the likelihood that maltreatment cases are reported to officials, economic factors may affect the likelihood that an individual reports being abused in a questionnaire, the likelihood that a doctor's diagnosis involves maltreatment, the likelihood that a maltreated child is taken to the hospital, or the likelihood that individuals suspecting or experiencing maltreatment search the internet for information. As such, they do not lessen the importance of considering the process by which maltreatment that occurs becomes observable to the researcher.

Links to Measures of Economic Conditions

Because of the sensitive nature of the subject, most maltreatment data are only available as aggregates (e.g., counts for states and years). Where micro data is available, it often does not include information on families' economic circumstances. As such, it is often only possible to consider links between maltreatment and the economic conditions of an area, which introduces the possibility that estimated relationships may be subject to the ecological fallacy. That is, a relationship between economic conditions and maltreatment that is observed in the aggregate may not reflect the relationship that exists for individuals. For example, it is possible for unemployment at the local level to increase child maltreatment while an individual being unemployed may have the opposite effect. Nonetheless, while it is important to acknowledge the limitations of what can be learned from estimates based on aggregate data, it is also important to note that there is value to understanding the links between economic conditions and child maltreatment in the aggregate.

With that said, some data on child maltreatment *do* provide information on the economic conditions of the household that the child lives in. It is from these data that we know that maltreated children tend to come from households that are economically disadvantaged relative to the average household. While these sorts of data are useful for providing descriptive statistics for children who are (observed) maltreated, data that has been selected on the outcome of interest cannot be used estimate causal links in any straightforward manner. Using microlevel data to estimate the degree to which various factors affect the probability of maltreatment requires data on individuals who are *not* maltreated in addition to those who are maltreated. Toward this end, researchers have used survey data including the National Family Violence Survey, the Fragile Families and Child Wellbeing Study, the National Longitudinal Survey of Youth, and by linking data sets with information on economic conditions to child abuse report data.

Empirical Strategies

As discussed in the “[Theory and Mechanisms](#)” section above, child maltreatment can be thought of as resulting from complex interactions between individual, familial, environmental, and societal risk factors. Given the large number of factors that may contribute to maltreatment and the interrelatedness of these factors, researchers face a major challenge in trying to identify the causal effects of economic conditions on maltreatment. In this section we highlight two approaches to overcoming this challenge, one that is best suited for estimating the effects of household economic factors and one that is best suited for estimating the effects of broader economic conditions.

Estimating the Effects of Household Economic Factors

Acknowledging that household economic conditions are generally *not* random, quantifying their causal effects requires researchers to consider circumstances in which they can measure the effects of random shocks to these conditions. Because it is difficult to identify these circumstances and to collect the maltreatment data necessary to

examine these circumstances, only a handful of such studies exist.

Fein and Lee (2003) take this approach in an experimental evaluation of a welfare reform program in Delaware. In particular, they compare outcomes for households subject to welfare reform to outcomes for those who were not subject to welfare reform, which was determined by random assignment. They find that the reform increased the incidence of reports of neglect but had no significant effect on reports of abuse or foster care placement. While this study represents some of the most convincing evidence to date that household economic factors have a causal effect on child maltreatment, it also underscores the difficulty of teasing out the causal effects of different interrelated economic factors. In particular, Delaware's welfare reform involved changes to benefit levels and work incentives in addition to other factors, any of which may have contributed to the increase in reports of neglect.

Cancian et al. (2013) also exploit evidence based on an experiment among welfare recipients to learn about the causal effect of household income on child maltreatment. In particular, they evaluate the effect of Wisconsin's reform that allowed a full pass-through of child support to welfare recipients (as opposed to the prior policy in which the government retained a fraction of child support payments to offset costs). Because the experimental intervention only changed child support pass-through – and no other aspect of child support or welfare receipt – the design allows for a straightforward interpretation of the results: that increasing income through this mechanism reduces maltreatment reports. The authors are careful to note, however, that increasing income through other mechanisms may have different effects on maltreatment. For example, an increase in income that is generated by an increase in maternal labor supply could very well increase the incidence of maltreatment.

Berger et al. (2014) take a different approach to identifying the causal effect of household economic conditions, exploiting naturally occurring variation in income (as opposed to experimentally manipulated variation) that they argue can be thought of as random. In particular, their strategy

utilizes variation in the generosity of the Earned Income Tax Credit (EITC) across states and over time. While this approach allows for a study that is broader in scope than the aforementioned experiments, a disadvantage of this approach is that changes in EITC rules can affect levels of income, work activity, and the broader social economic climate, which again highlights the challenge in the identification and interpretation of causal effects.

Estimating the Effects of Broader Economic Conditions

Another strand of the literature on the causal effects of economic conditions on child maltreatment abstracts from the household to consider the effects of changes in local economic conditions on rates of maltreatment in the aggregate. Acknowledging that local economic conditions tend to be correlated with many socioeconomic factors that predict maltreatment, several studies have taken an “area approach” that considers how rates of maltreatment in an area change *over and above changes occurring across all areas* when its economic conditions change *over and above changes occurring across all areas*. As such, estimates based on this approach are identified using variation across areas in the timing and severity of changing economic conditions. This approach is operationalized via regression models that include time-fixed effects to capture changes occurring across all areas at the same time, area-fixed effects to capture time-invariant area characteristics, and (sometimes) area-specific trends. The validity of this approach rests on the assumption that unobservable variables related to the outcome variable do not deviate from an area's trend when its economic conditions deviate from trend.

Studies taking this approach vary considerably in their measures of maltreatment, their measures of economic conditions, and the way they define areas. Paxson and Waldfogel (1999, 2002, 2003), Seiglie (2004), and Bitler and Zavodny (2002, 2004) use state-level panel data to estimate the effects of a variety of economic indicators on maltreatment reports, finding mixed results. Lindo et al. (2013) and Frioux et al. (2014) use county-level data from California and

Pennsylvania, respectively, also finding mixed results. Wood et al. (2012) focus on hospital admissions for abuse-related injuries using panel data from 38 hospitals from 2000 to 2009 along with a variety of economic indicators and find evidence that local economic downturns significantly increase the incidence of severe physical abuse; however, they do not account for the likely autocorrelation in the error terms within hospitals over time, which would serve to widen their confidence intervals.

Conclusion

Child maltreatment is an important topic that has received relatively little attention in the field of economics, despite generating large financial costs for society and significant consequences for the health, human capital accumulation, and eventual labor market outcomes of its victims. The scarcity of economic research on the topic is especially unfortunate given that a literature spanning many disciplines and several decades has found economic factors, including local economic conditions, family income, neighborhood poverty, employment status, and receipt of public assistance, to be robust predictors of child abuse. We suspect that this scarcity is driven by economists' strong emphasis on the identification of causal effects, which is particularly challenging for research on the economic determinants of child maltreatment. In some sense, identifying causal effects in this area requires a perfect storm in which there is random variation in economic conditions, the researcher has access to maltreatment data that allows for comparisons utilizing this random variation, and the researcher can be confident that the way in which maltreatment becomes observed in these data does not vary across the groups of individuals and/or time periods he/she intends to compare. Moreover, even when this perfect storm occurs such that a causal estimate can be obtained, the interrelatedness of economic factors can make it difficult to interpret such estimates. For example, the causal effect of a parent's job displacement could reflect the effects of income or time use (or other factors).

Despite these challenges, recent progress has been made in identifying the causal effects of economic factors on child maltreatment through the use of experimental (natural and true) variation and area studies. These studies indicate that changes in economic conditions can have meaningful impacts on maltreatment. However, there is still much work to be done in identifying exactly which economic factors matter and in characterizing the nature of these relationships.

References

- Apouey B, Clark AE (2014) Winning big but feeling no better? The effect of lottery prizes on physical and mental health. *Health Econ*, <http://onlinelibrary.wiley.com/doi/10.1002/hec.3035/full>
- Belsky J (1980) Child maltreatment: an ecological integration. *Am Psychol* 35(4):320–335
- Berger LM (2004) Income, family structure, and child maltreatment risk. *Child Youth Serv Rev* 26(8): 725–748
- Berger LM (2005) Income, family characteristics, and physical violence toward children. *Child Abuse Negl* 29(2):107–133
- Berger LM, Waldfogel J (2011) Economic determinants and consequences of child maltreatment. OECD social, Employment and migration working papers, No. 111, <http://storage.globalcitizen.net/data/topic/knowledge/uploads/20120229105523533.pdf>
- Berger Lawrence M, Sarah A Font, Kristen S Slack, Jane Waldfogel (2014) Income and child maltreatment: evidence from the earned income tax credit Mimeo
- Bitler M, Zavodny M (2002) Child abuse and abortion availability. *Am Econ Rev* 92(2):363–367
- Bitler M, Zavodny M (2004) Child maltreatment, abortion availability, and economic conditions. *Rev Econ Househ* 2(2):119–141
- Brand JE, Levy BR, Gallo WT (2008) Effects of layoffs and plant closings on subsequent depression among older workers. *Res Aging* 30(6):701–721
- Browning M, Heinesen E (2012) Effect of job loss due to plant closure on mortality and hospitalization. *J Health Econ* 31(4):599–616
- Cancian M, Yang M-Y, Slack KS (2013) The effect of additional child support income on the risk of child maltreatment. *Soc Serv Rev* 87(3):417–437
- Charles KK, Stephens M Jr (2004) Job displacement, disability, and divorce. *J Labor Econ* 22(2):489–522
- Cicchetti D, Lynch M (1993) Toward an ecological/transactional model of community violence and child maltreatment: consequences for children's development. *Psychiatry* 56(1):96–118
- Dávalos ME, Fang H, French MT (2012) Easing the pain of an economic downturn: macroeconomic conditions and

- excessive alcohol consumption. *Health Econ* 21(11): 1318–1335
- Deb P, William T, Gallo PA, Fletcher JM, Sindelar JL (2011) The effect of job loss on overweight and drinking. *J Health Econ* 30(2):317–327
- Doiron D, Mendolia S (2012) The impact of job loss on family dissolution. *J Popul Econ* 25(1):367–398
- Eliason M, Storrie D (2009) Does job loss shorten life? *J Hum Resour* 44(2):277–302
- Eliason M, Storrie D (2010) Inpatient psychiatric hospitalization following involuntary job loss. *Int J Ment Health* 39(2):32–55
- Fang X, Brown DS, Florence CS, Mercy JA (2012) The economic burden of child maltreatment in the United States and implications for prevention. *Child Abuse Negl* 36(2):156–165
- Fein DJ, Lee WS (2003) The impacts of welfare reform on child maltreatment in Delaware. *Child Youth Ser Rev* 25(1):83–111
- Frioux S, Wood JN, Oludolapo F, Luan X, Localio R, Rubin DM (2014) Longitudinal association of county-level economic indicators and child maltreatment incidents. *Matern Child Health J*, <http://link.springer.com/article/10.1007/s10995-014-1469-0>
- Garbarino J (1977) The human ecology of child maltreatment: a conceptual model for research. *J Marriage Fam* 39(4):721–735
- Gelles RJ, Staci P (2012) Estimated annual cost of child abuse and neglect. Prevent Child Abuse America, Chicago
- Gilbert R, Widom CS, Browne K, Fergusson D, Webb E, Janson S (2009) Burden and consequences of child maltreatment in high-income countries. *Lancet* 373(9657):68–81
- Kuhn P, Kooreman P, Soetevent A, Kapteyn A (2011) The effects of lottery prizes on winners and their neighbors: evidence from the Dutch postcode lottery. *Am Econ Rev* 101(5):2226–2247
- Lindo JM, Schaller J, Hansen B (2013) Economic conditions and child abuse. National bureau of economic research working paper w18994, Cambridge, MA
- Mendolia S (2014) The impact of husbands job loss on partners mental health. *Rev Econ Househ* 12(2): 277–294
- National Research Council (1993) Understanding child abuse and neglect. The National Academies Press, Washington, DC
- Paxson C, Waldfogel J (1999) Parental resources and child abuse and neglect. *Am Econ Rev* 89(2):239–244
- Paxson C, Waldfogel J (2002) Work, welfare, and child maltreatment. *J Labor Econ* 20(3):435–474
- Paxson C, Waldfogel J (2003) Welfare reforms, family resources, and child maltreatment. *J Policy Anal Manage* 22(1):85–113
- Pelton LH (1994) The role of material factors in child abuse and neglect. In: Melton GB, Barry FD (eds) *Protecting children from abuse and neglect: foundations for a new national strategy*. Guilford Press, New York, pp 131–181
- Petersen A, Joshua J, Monica F (eds) (2014) *New directions in child abuse and neglect research*. The National Academies Press, Washington, DC
- Ruhm CJ, Black WE (2002) Does drinking really decrease in bad times? *J Health Econ* 21(4):659–678
- Schaller J (2013) For richer, if not for poorer? Marriage and divorce over the business cycle. *J Popul Econ* 26(3):1007–1033
- Schaller J, Stevens AH (2014) Short-run effects of job loss on health conditions, health insurance, and health care utilization. National bureau of economic research working paper w19884, Cambridge, MA
- Sedlak AJ, Mettenburg J, Basena M, Petta I, McPherson K, Greene A, Li S (2010) Fourth national incidence study of child abuse and neglect (NIS-4): report to congress. US Department of Health and Human Services, Administration for Children and Families, Washington D.C.
- Seiglie C (2004) Understanding child outcomes: an application to child abuse and neglect. *Rev Econ Househ* 2(2):143–160
- Stephens-Davidowitz S (2013) Unreported victims of an economic downturn. Mimeo, <http://static.squarespace.com/static/51d894bee4b01caf88ccb4f3/t/51e22f38e4b0502fe211fab7/137377720363/childabusepaper13.pdf>
- Stith SM, Ting Liu L, Davies C, Boykin EL, Alder MC, Harris JM, Som A, McPherson M, Dees JEMEG (2009) Risk factors in child maltreatment: a meta-analytic review of the literature. *Aggress Violent Beh* 14(1):13–29
- Waldfogel J (2000) Child welfare research: how adequate are the data? *Child Youth Serv Rev* 22(9): 705–741
- Weinberg BA (2001) An incentive model of the effect of parental income on children. *J Polit Econ* 109(2): 266–280
- Wood JN, Sheyla P, Medina CF, Luan X, Localio R, Fieldston ES, Rubin DM (2012) Local macroeconomic trends and hospital admissions for child abuse, 2000–2009. *Pediatrics* 130(2):e358–e364
- Zivin K, Paczkowski M, Galea S (2011) Economic downturns and population mental health: research findings, gaps, challenges and priorities. *Psychol Med* 41(07): 1343–1348

Child Molestation

► Sex Offenses

Child Pornography

► Sex Offenses

Children

► Adoption

Choice Under Risk and Uncertainty

Gianna Lotito¹ and Anna Maffioletti²

¹DiGSPES, University of Eastern Piedmont, Alessandria, Italy

²ESOMAS, University of Turin, Turin, Italy

Definition

This entry outlines what is meant by decision-making under risk and uncertainty. It illustrates the model of expected utility, its properties, and the Allais paradox as the main violation of the model. It describes the subjective expected utility model of decision under uncertainty, and the Ellsberg paradox as an example of the Knight's approach to uncertainty.

Introduction

In real economic life, many decisions are taken under risk and uncertainty, for example, investment decisions, decisions about consumption through time, buying and selling insurance, investment in new industries and countries, choosing new technologies, stock market purchases, and sales.

The literature on decision-making under risk and uncertainty can be divided in: (1) the literature concerning decision-making under risk, which includes: (a) the expected utility model (EU) and its axiomatizations (Bernoulli 1954; von Neumann and Morgenstern 1947); (b) the criticisms to the EU model (Allais 1953) and its alternatives (Kahneman and Tversky 1979; Quiggin 1982, 1993; Loomes and Sugden 1982); (2) the literature concerning decision-making under uncertainty, which can be divided into two different approaches: (a) the Bayesian approach

(De Finetti 1937; Ramsey 1931; Savage 1954), according to which people assign subjective probabilities to uncertain events and these probabilities follow the rules of mathematical probability theory and (b) the approach by Knight (1921) and Keynes (1921), according to which people *cannot* assign subjective probabilities to uncertain events that follow the mathematical rules of probability theory (Ellsberg 1961; Schmeidler 1989; Tversky and Kahneman 1992).

In a situation of *risk*, one does not know with certainty the outcomes of one's choice, and the uncertainty related to a decision can be represented by an objective probability distribution over the events or states of the world which relate actions to outcomes, for example, with a gamble based on the roll of a die or a roulette wheel.

In a situation of *uncertainty*, the uncertainty relating a decision *cannot* be represented by an objective probability distribution. In this case, the individual is either considered able to attach a subjective esteem of the probability to each event (Savage 1954) – in which case decision under uncertainty reduces into decision under risk – or probabilities are not known and cannot be assigned (Knight 1921).

Choice Under Risk

Let us define now a prospect or lottery like the combination of all the possible outcomes with the probabilities of the events under which these outcomes occur.

Consider, for example, having a ticket for the following lottery $L = (\$100, \frac{1}{2}; \$0, \frac{1}{2})$ – where the outcomes are monetary prizes – in which a coin is tossed: in the event the coin falls Head (which occurs with probability $\frac{1}{2}$), a prize of \$100 is won; in the event the coin falls Tail (which also occurs with probability $\frac{1}{2}$), nothing is won.

Thus, an action in a risky situation corresponds to playing a lottery or prospect, which associates each outcome to the probability of its occurrence. As a consequence, in such a situation, choice can be viewed as a choice of the preferred lottery or prospect. But how can the individual value the different prospects? One possibility is that the

individual calculates the *expected value* (EV) of the different prospects, and chooses the prospect with the highest of these values. The concept of expected value was first developed by seventeenth century mathematicians, for example, Pascal. The expected value of a lottery is the average of the monetary prizes associated with the different outcomes, weighted by their respective probabilities. The expected value of the above lottery L would be $(\frac{1}{2} * \$100 + \frac{1}{2} * \$0) = \$50$. However, maximizing the expected monetary value does not always appear to be a satisfactory criterion to choose among different risky situations. Let us consider the two lotteries $L_1 = (\$60, \frac{1}{2}; \$0, \frac{1}{2})$ and $L_2 = (\$40, \frac{1}{2}; \$20, \frac{1}{2})$. Despite both have the same expected value of \$30, some people might prefer L_2 , where there always is a probability of gaining a positive outcome, to L_1 , where there is a 50% chance of winning nothing. As a further example, consider the following lotteries: $L_3 = (\$100, \frac{1}{2}; \$-1, \frac{1}{2})$ and $L_4 = (\$100000, \frac{1}{2}; \$-1000, \frac{1}{2})$. At a first sight, many people would be willing to participate to the first lottery, but would they accept so easily to play the second? It does not seem so: L_4 has a 50% probability of winning an outcome 1000 times higher than L_3 , but also a 50% probability of losing an outcome 1000 times higher. However, if one calculates the expected value of the lotteries, it turns out that the EV of L_3 (\$49) is much lower than the EV of L_4 (\$49500). Clearly, expected value is not always a sufficient criterion to make a lottery attractive. An alternative is needed.

History

The first important argument on the subject was that of the mathematician Daniel Bernoulli (1738), who (independently from Gabriel Cramer) developed a new hypothesis based on the solution of a problem posed by his cousin Nicholas Bernoulli, the “St Petersburg Paradox.” Suppose you have to toss a coin till “Head” comes out; the first time this happens, you stop tossing the coin and prizes are determined in the following way: if “Head” comes out at the first toss, you earn \$2; if “Head” comes out at the second toss, you earn \$2²; if “Head” comes out at the third toss, you

earn \$2³, and so on. The probability that at every toss “Head” comes out is equal to $\frac{1}{2}$, and tosses are independent from each other: then, the probability that “Head” comes out at the first toss is $\frac{1}{2}$, the probability that “Head” comes out at the second toss is $(\frac{1}{2})^2$ (that is, the probability that “Tail” comes out at the first toss times the probability that “Head” comes out at the second toss, that is, $\frac{1}{2} * \frac{1}{2}$), the probability that “Head” comes out at the third toss is $(\frac{1}{2})^3$, and so on. Thus, the expected value of this lottery is infinite: $2(\frac{1}{2}) + 4(\frac{1}{2})^2 + 8(\frac{1}{2})^3 + \dots + 2n(\frac{1}{2})^n + \dots = 1 + 1 + 1 + 1 + \dots = \infty$, as the coin is thrown, if necessary, an infinite number of times and the expected prize from each toss is equal to 1. An individual who looks at the highest expected value would be willing to pay an infinite amount of money to play this lottery. But this is very unrealistic: nobody would be willing to pay more than a modest amount for it. In fact, this is a very risky lottery: it gives the opportunity to gain an increasingly bigger prize with a decreasingly smaller probability – people would not be willing to undertake this risk.

Bernoulli argued that this and similar choices could be explained by assuming that individuals do not choose the lottery with the highest *expected value*, but that with the highest *expected utility*, where the utility is represented by a function like the square root of wealth, where utility increases with wealth, but at a decreasing rate. Bernoulli introduced both the concepts of *expected utility* maximization and of decreasing marginal utility of wealth. However, Bernoulli’s idea of *expected utility* maximization had little effect on the theory of decision-making under risk till the work “Theory of Games and Economic Behaviour” by von Neumann and Morgenstern was published in 1947. After that, it had soon to become the normative and positive theory of decision-making under risk.

The Expected Utility Model

The expected utility model relies on the hypothesis that the individual possesses – or acts *as if* he possesses a von Neumann-Morgenstern utility function over a set of outcomes, and when he faces alternative lotteries over these outcomes he

chooses that lottery which maximizes the expected value of this utility. In particular, von Neumann and Morgenstern start by assuming specific conditions on the preference relations between lotteries; these are necessary and sufficient to show the existence of a utility function that assigns a numerical value to the “satisfaction” of the different outcomes of the lotteries. The individual thus chooses the lottery for which the expected utility is the highest, where the expected utility of a lottery $L = (x_1, p_1; \dots; x_n, p_n)$ is given by the sum of the products of the probability and utility over all possible outcomes, that is,

$$EU(L) = \sum_{i=1}^n U(x_i)p_i \quad i = 1, \dots, n.$$

Given p_i the probability of the outcome x_i and $U(x_i)$ its utility, the expected utility of a lottery $L = (x_1, p_1; x_2, p_2)$ will be equal to $EU(L) = p_1U(x_1) + p_2U(x_2)$. The expected utility of a lottery is then the expected value of the utilities of the possible outcomes. As an example, consider the two lotteries L_1 and L_2 illustrated above. Assume that the utility function is of the form $U(x) = \sqrt{x}$, where x is the monetary value of the outcome. The expected utility of L_1 is then equal to $EU(L_1) = \frac{1}{2}$

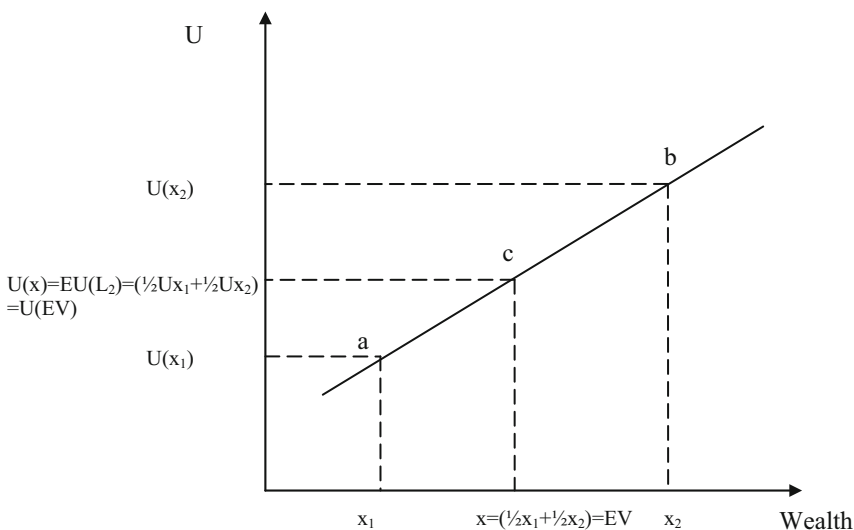
$\cdot \sqrt{60} = \$3.87$ and $EU(L_2) = \frac{1}{2} \cdot \sqrt{40} + \frac{1}{2} \cdot \sqrt{20} = \$3.16 + \$2.23 = \5.39 .

We can obtain a very important information by observing the expected utilities of these lotteries. If we compare them with the expected value (equal to \$30 for both), we notice that the expected utilities are not only different from the expected value, but are different from each other. In particular, the expected utility of the “riskier” lottery L_1 (the one which gives a 50% chance of winning nothing) is lower than that of L_2 .

The von Neumann-Morgenstern utility function gives us information about the individual’s attitudes toward risk.

Attitudes Toward Risk

The possible attitudes to risk of an individual can be represented graphically as follows. Consider Fig. 1 with prizes in monetary value (wealth) on the horizontal axis and their utility on the vertical axis. The straight line represents the von Neumann-Morgenstern utility function of the individual and tells us the utility he or she assigns to a given sum of money. In this case in which the shape of the utility function is a straight line, the



Choice Under Risk and Uncertainty, Fig. 1 von Neumann-Morgenstern utility function of a risk-neutral individual

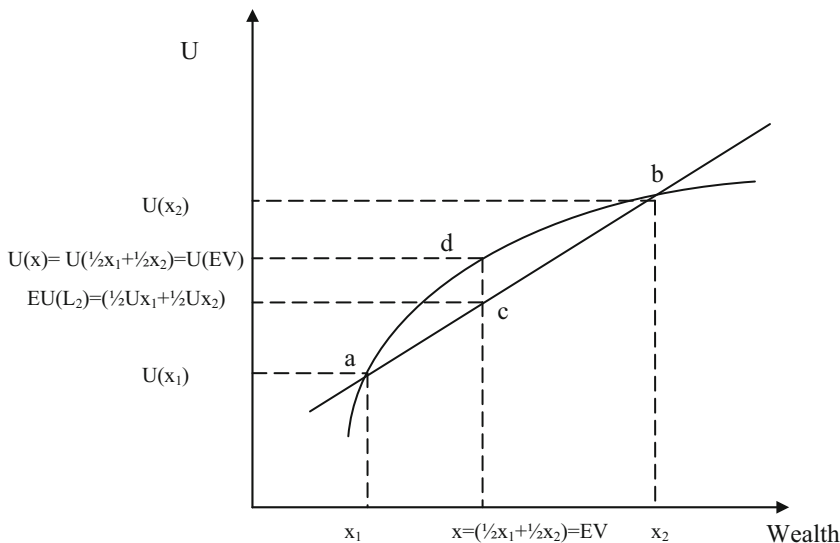
individual is neutral to risk: when facing different lotteries he or she chooses on the basis of their expected monetary value.

Consider the two lotteries $L_1 = x$ for certain and $L_2 = (x_1, \frac{1}{2}; x_2, \frac{1}{2})$, where outcome x is also equal to the expected value of L_2 . L_2 is riskier than L_1 , as its prizes have a higher variability. The risk-neutral individual will not take into account this risk and will value the lotteries only on the basis of their expected value: he or she will be indifferent between them. In Fig. 1, point a represents the utility of the lowest outcome x_1 , $U(x_1)$, and point b the utility of the highest outcome x_2 , $U(x_2)$; point c is, therefore, the expected utility of lottery L_2 , $EU(L_2) = \frac{1}{2} * U(x_1) + \frac{1}{2} * U(x_2)$, and is exactly midway between a and b . The expected utility of the sure outcome x , $U(x)$ is given by point d . Then, an individual with such a utility function will be indifferent between having x for certain or a lottery with expected value equal to x : they have the same expected utility. The fact that L_2 is riskier than L_1 does not influence choice. For this individual thus the expected utility of a lottery is equal to the utility of a sure outcome that in turn is equal to the expected value of the lottery

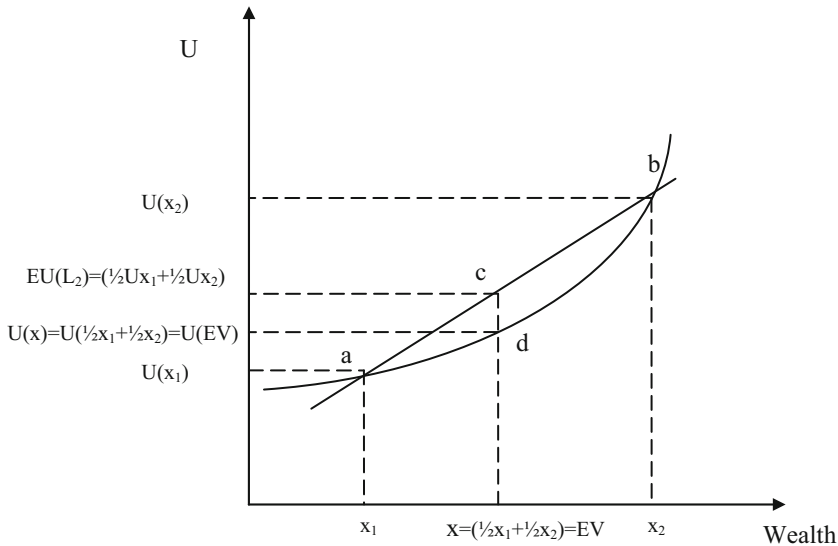
Consider now Fig. 2 which represents an attitude of aversion to risk: the utility function represented here is a curve, where the utility

increases with wealth, but at a decreasing rate – an individual with this utility function will not be indifferent between a sure outcome and a lottery. Suppose the individual is facing the same alternatives as before. In the graph, point a represents $U(x_1)$ and point b $U(x_2)$; the expected utility of lottery L_2 will be the weighted average of these two utilities and will be given by point c , in the middle of the segment unifying a and b , $\frac{1}{2} U(x_1) + \frac{1}{2} U(x_2)$. However, we see now that the expected utility of the sure prize x is given by point d , which is above c : $U(x) > \frac{1}{2} U(x_1) + \frac{1}{2} U(x_2)$. For this individual thus the expected utility of a lottery is lower than the utility of a sure outcome that is equal to the expected value of the lottery. The individual who is averse to risk will choose the sure alternative to a lottery which gives an expected prize equal to the sure one – he or she prefers the expected value of the lottery to the lottery itself. The individual is averse to the risk connected to the lottery.

We should note that this risk aversion is implicit in the shape of the utility function, which is concave. As an illustration of this, consider lotteries $(\$60, \frac{1}{2}; \$0, \frac{1}{2})$ and $(\$40, \frac{1}{2}; \$20, \frac{1}{2})$ above. In calculating their expected utilities, we have assumed a form for the utility function equal to $U(x) = \sqrt{x}$, which has a shape like in



Choice Under Risk and Uncertainty, Fig. 2 von Neumann-Morgenstern utility function of a risk-averse individual



Choice Under Risk and Uncertainty, Fig. 3 von Neumann-Morgenstern utility function of a risk-loving individual

Fig. 2. This is a utility function implying risk aversion, as shown by the fact that when calculating the expected utilities of the above lotteries, we find that the riskier lottery has a lower expected utility than the “safer” lottery.

The case of the utility function of an individual with a proneness to risk is given in Fig. 3. It can be seen here that for such an individual the expected utility of a lottery is higher than the utility of a sure outcome that is equal to the expected value of the lottery. The individual who is risk loving will choose the lottery which gives an expected prize equal to the sure outcome instead of the sure alternative – he or she prefers the lottery to the expected value of the lottery itself. This individual loves the risk connected to the lottery.

The individual preferences toward risk can also be expressed using the concepts of (1) *certainty equivalent* (CE) and (2) *risk premium* π .

1. As an example, consider the lottery $L = (\$100, \frac{1}{2}; \$0, \frac{1}{2})$. The expected value of this lottery is 50. Consider asking this individual what is the amount of money which makes her indifferent to the lottery. This is defined as the *certainty equivalent* of the lottery. Therefore, the utility of the CE of the lottery is defined as equal to its expected utility – $U(CE) = EU(L)$.

Suppose the individual is indifferent between the lottery and \$40 for certain, that is, his CE for the lottery is equal to 40, less than the expected value of the lottery. It follows that the subject would accept any sum >40 instead of the lottery and the lottery to each sum <40 . In this case, we say that the subject is averse to risk – $U(CE) < EU(L)$. This is represented in Fig. 4. The reverse inequality defines risk proneness, and the equality defines risk neutrality.

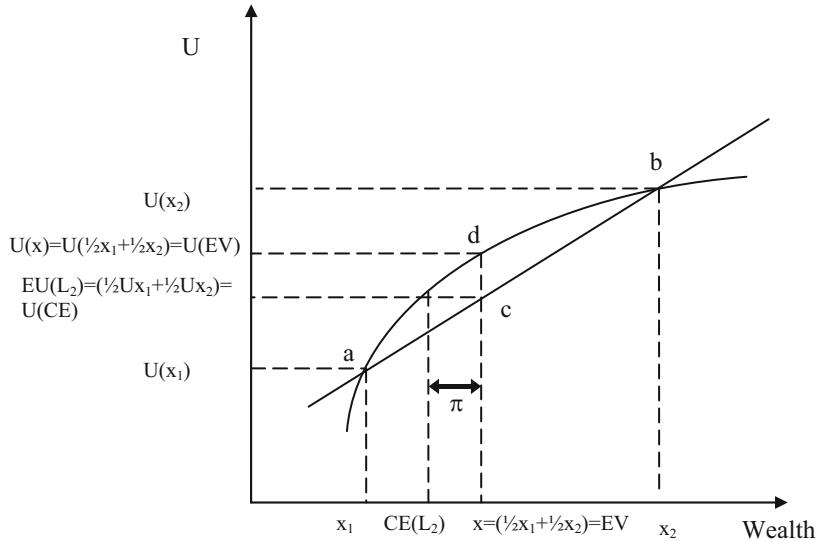
2. The risk premium π is defined as *the maximum part of the expected value that the subject is willing to give up to avoid the risk associated to the lottery*, $\pi = EV(L) - CE$ (see Fig. 4).

Consider the example of a CE for the above lottery equal to 40; this means that the individual is willing to give up at most \$10 of the expected monetary value of the lottery, that is, the risk premium is 10. The sign of the risk premium gives the attitude to risk for that lottery. In general, when for any lottery π is positive, the individual is averse to risk; when π is negative, he is risk lover; when $\pi = 0$, the individual is risk neutral.

The most common analytical measures of risk aversion have been introduced by Arrow (1964) and Pratt (1964) and extensively used in finance, insurance markets, health, and game theory.

Choice Under Risk and Uncertainty,

Fig. 4 Utility function of a risk-averse individual with certainty equivalent and risk premium measures



The Allais Paradox

The most important implication of the specific form of the expected utility preference function $EU(L)$

$$= \sum_{i=1}^n U(x_i)p_i \text{ is linearity in the probabilities}$$

(Marschak 1950; Machina 1987; Raiffa 1968). For this reason, most of the empirical investigation of the expected utility hypothesis has focused on this property (MacCrimmon and Larsson 1979; Allais and Hagen 1979; Starmer and Sugden 1989; Starmer 2000), revealing widespread systematic violations. The best-known violation of linearity in the probability is the Allais paradox (Allais 1953). Consider that an individual is asked to make a pairwise choice between the two following pairs of lotteries (outcomes in French francs as in Allais 1953, p. 527):

- A 100% chance of 100 millions
- B $\left\{ \begin{array}{l} 10\% \text{ chance of } \$500 \text{ millions} \\ 89\% \text{ chance of } \$100 \text{ millions} \\ 1\% \text{ chance of } 0 \end{array} \right.$
- C $\left\{ \begin{array}{l} 11\% \text{ chance of } \$100 \text{ millions} \\ 89\% \text{ chance of } 0 \end{array} \right.$
- or
- D $\left\{ \begin{array}{l} 10\% \text{ chance of } 500 \text{ millions} \\ 90\% \text{ chance of } 0 \end{array} \right.$

If the individual has expected utility preferences, a choice of A over B in the first pair

implies a choice of C over D in the second pair. However, empirical findings show that the modal choice is for A over B in the first couple of lotteries, but D over C in the second couple. This pattern of choice violates expected utility. In fact, A preferred to B implies $U(100 \text{ millions}) > .10U(500 \text{ millions}) + .89U(100 \text{ millions}) + .01U(0)$, that is, $.11U(100 \text{ millions}) + .89U(0) > .10U(500 \text{ millions}) + .90U(0)$, while D preferred to C implies $.10U(500 \text{ millions}) + .90U(0) > .11U(100 \text{ millions}) + .89U(0)$, which is a contradiction.

The widespread and systematic empirical violations of the linearity property has put under discussion the descriptive validity of the theory, giving rise to a growing body of literature of new theoretical models of choice under risk as, for instance, Prospect Theory (Kahneman and Tversky 1979; Tversky and Kahneman 1992), Rank Dependent Utility Theory (Quiggin 1982, 1993), Regret Theory (Loomes and Sugden 1982).

Choice Under Uncertainty

The first distinction between choice under risk and choice under uncertainty was made by Knight (1921) and by Keynes (1921). They define a choice under uncertainty when we deal with a choice in which the probability of the occurrence of an outcome is not known. Savage (1954)

extended the theory of Expected Utility to uncertainty and called it Subjective Expected Utility. According to Savage, all individuals can have a subjective esteem of the probability attached to an event. If this is true, then uncertainty is reduced to risk. Moreover, according to Savage, this individual estimate of probability follows the mathematical rules of probability. Consider the following lottery: “you will receive \$100 if tomorrow at noon the temperature in Rome is higher than 30 degree Celsius and \$0 otherwise.” The probability that we assign to the event “at noon the temperature in Rome is higher than 30 degree Celsius” and the probability that we assign to the opposite event “at noon the temperature in Rome is 30 degree Celsius or less” follow the rules of probability, that is, they sum to one. In addition, our probability estimate should be inferred from our choice between lotteries.

Ellsberg (1961) pointed out that this is not true, and hence uncertainty cannot be reduced to risk. Consider the following example given by Ellsberg. You have in front of you two urns: Urn I and Urn II. Both urns are opaque so you cannot see inside. You are told that in Urn I there are 100 balls: 50 are black and 50 are red. Instead, Urn II contains 100 balls that can be either black or red but you do not know in which proportion.

You are asked to bet on a color and choose the urn from which to draw the ball simultaneously. If the ball that you have drawn is of the same color you have chosen, then you win \$100, otherwise you will get nothing.

According to Ellsberg, most of the people will decide to draw the ball from Urn I, which contains 50 red and 50 black balls, independently from which color they have chosen to bet on. In this case, in fact the probabilities are known to be 50/100 for Red and 50/100 for Black.

The proportion of the red and black balls in the second urn is not known and so we do not know which probability to assign to red and black. Avoiding to choose to draw a ball from Urn II is called by Ellsberg *uncertainty aversion*.

According to the theory of Savage, we should be indifferent between the two urns since the probability of getting red can be represented by a second order probability distribution

(a probability distribution over probability), whose expected probability is 50/100 as in the case of Urn I. However, people prefer to bet on the first urn showing that they prefer to draw a ball from an urn in which the probabilities of the two color balls are precisely defined.

Aversion to uncertainty creates a problem to Savage theory also for another reason. Let's consider again the two urns. If you always choose to draw a ball from Urn I, whatever color you prefer, this implies that you consider the probability of getting red plus the probability of getting black in Urn I different from the probability of getting red plus the probability of getting black in Urn II. If this is the case, then, the sum of your probability estimates of the two color balls for the two urns is going to be different from 1. In particular, ambiguity aversion implies that your estimate of the probability of red plus your estimate of the probability of black in Urn II is less than 1. Your probability estimates thus are not following the mathematical rules of Savage theory. Intuitively, the difference of the sum of the two probabilities from one is the room that we leave to the existence of uncertainty, that is to say, the possibility of occurrence of some unexpected situation. As a consequence, it is very difficult to deduce our probability estimate from our choice as pointed out by Ellsberg. Ellsberg's work has been confirmed by a substantial number of empirical and theoretical research contributions (Camerer and Weber 1992; Trautmann et al. 2008; Machina and Siniscalchi 2014; Gilboa and Marinacci 2016).

Cross-References

► [Risk Management, Optimal](#)

References

- Allais M (1953) Fondements d'une théorie positive des choix comportant un risque et critique des postulats et axiomes de l'école Américaine. *Econometrica* 21(4):503–546
- Allais M, Hagen O (eds) (1979) Expected utility hypothesis and the Allais paradox. D. Reidel, Dordrecht
- Arrow K (1964) The role of securities in the optimal allocation of risk-bearing. *Rev Econ Stud* 31:91–96

- Bernoulli D (1954) Specimen theoriae novae de mensura sortis. (Commentarii Academiae Scientiarum Imperialis Petropolitanae, 1738). *Econometrica* 22:23–26. (Trans. as Exposition of a new theory on the measurement of risk)
- Camerer CE, Weber M (1992) Recent developments in modeling preferences: uncertainty and ambiguity. *J Risk Uncertain* 5(4):325–370
- de Finetti B (1937) La prévision: ses lois logiques, ses sources subjectives. *Ann I H Poincaré* 7(1):1–68
- Ellsberg D (1961) Risk ambiguity and the Savage axioms. *Q J Econ* 75:643–669
- Gilboa I, Marinacci M (2016) Ambiguity and the Bayesian paradigm. In: Arló-Costa H, Hendricks VF, van Benthem J (eds) *Readings in formal epistemology*. Springer, New York
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47:273–291
- Keynes J,M (1921) *A treatise on probability*. McMillan, London
- Knight F (1921) *Risk, uncertainty and profit*. Houghton-Mifflin, Boston
- Loomes G, Sugden R (1982) Regret theory: an alternative theory of rational choice under uncertainty. *Econ J* 92:805–824
- MacCrimmon KR, Larsson S (1979) Utility theory: axioms versus ‘paradoxes’. In: Allais M, Hagen O (eds) *Expected utility hypothesis and the Allais paradox*. D. Reidel, Dordrecht
- Machina M (1987) Choice under uncertainty: problems solved and unsolved. *J Econ Perspect* 1(1):121–154
- Machina M, Siniscalchi M (2014) Ambiguity and ambiguity aversion. In: Machina M, Viscusi K (eds) *The handbook of the economics of risk and uncertainty*. North-Holland, Oxford
- Marschak J (1950) Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18:111–141
- Pratt JW (1964) Risk aversion in the small and in the large. *Econometrica* 32:122–136
- Quiggin J (1982) A theory of anticipated utility. *J Econ Behav Organ* 3:323–343
- Quiggin J (1993) *Generalized expected utility theory*. Kluwer, Dordrecht
- Raiffa H (1968) *Decision analysis: introductory lectures on choices under uncertainty*. Addison-Wesley, Reading
- Ramsey FP (1931) Truth and probability. In: *The foundations of mathematics and other logical essays*. Harcourt, Brace, New York
- Savage L (1954) *The foundations of Statistics*. Wiley, New York
- Schmeidler D (1989) Subjective probability and expected utility without additivity. *Econometrica* 57(3):571–587
- Starmer C (2000) Developments in non-expected utility theory. *J Econ Lit* 38:332–382
- Starmer C, Sugden R (1989) Violations of the independence Axiom in common ratio problems: an experimental test of some competing hypotheses. *Ann Oper Res* 19(1):79–102
- Trautmann ST, Vieider FM, Wakker PP (2008) Causes of ambiguity aversion: known versus unknown preferences. *J Risk Uncertain* 36(3):225–243

- Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 5:297–323
- von Neumann J, Morgenstern O (1947) *Theory of games and economic behaviour*, 2nd edn. Princeton University Press, Princeton

Civil Law System

Pier Giuseppe Monateri
SciencesPo, Ecole de droit, University of Torino,
Torino, Italy

Definition

This entry defines what is meant today for civil law; sketches an outline of its historical background; individualizes the two main models of it, German and French; and points to the difficulties of comparison with the common law, as to the hardships of a work of harmonization.

Introduction: What Is Civil Law as a Legal Family?

What we call “civil law system” is indeed a family of different legal systems tracing their historical roots to the Roman law.

As such, this family of legal systems is differentiated, today, especially in regard to the other two major legal families existing in contemporary world legal landscape: the common law legal family and the Sharia of the Islamic legal model. As well as the civil law, the “common law” is a set of highly differentiated systems of law sharing the same origin to be found in the history and development of the English common law. Differently the Muslim Sharia is supposed to be a unique system of principles and rules, based on the divine revelation contained in the Koran, even if its interpretation may vary very greatly in different jurisdictions, cohabiting, also, with European-like codes and modern constitutions, and today is, on the average, applied only to the *status personae*, the personal condition of the subject, as marriage, divorce, inheritance, and other related matters (Samuel 2014).

This given, it is manifest that when we speak of common and civil law, as the two major variants of the Western legal tradition, we make reference to the different *legal origins* of modern systems, implying that these differences are still molding the actual structure of our laws (World Bank 2003).

Historical Background of the Civil Law Origins

The term “civil law” is an English term used to translate the *jus civile* or the proper Roman law as it evolved from classical times to the end of the empire when it became codified by Justinian, from 529 to 534 AD, in his codes, constituting an ordered collection of a mass of writing known as the *Corpus Juris Civilis* or *The Body of Civil Law*. The work was planned to be divided into three parts: the *Code* as a compilation of imperial enactments, the *Digest* or *Pandects* composed of *advices* given by older Roman jurists on different points of the law and deemed to have authority for their learned character, and finally the *Institutes* conceived as a textbook for law students at the newly established law school of the empire in Beyrouth.

Tribonian has been the editor in chief of this massive work, thought to represent the whole of the jurisprudential tradition evolved from early Roman times up to the date of the compilation.

It is important to note two main facts:

First, it is the fact that the Roman Empire at that time was split into two parts and that this compilation was enacted, having force of law, only in the eastern part of the empire speaking Greek. In this way, the Justinian compilation, quite exotically, has been written in Latin for an empire speaking Greek and was never enacted as such in the West, but influenced its legal progress in the strongest possible way, something which defeats any of our actual understandings of the working of law.

Second, this enterprise has marked a total revolution of Roman law, changing completely its style and its structure. Roughly speaking,

classical Roman law was an *oral* law, without codes, but only with pieces of legislation passed by the various political assemblies. There was *not* a formal system of legal education, each one having to learn the law from a practicing lawyer, and especially there were not regular courts of law (Glenn 2000).

The Roman magistrate directing the trial, the praetor, was a politician, appointed for 1 year, controlling only the *form* of actions pleaded before him by the parties. Then, to afford the trial, he had to nominate a *judex*, a “judge,” a layman, to be agreed by the parties. In this way he was more an *arbitrator* than a *judge*. Just for this reason, the learned opinion of jurists of great reputation played such an important role: they had to advice the *praetor* and the *judex*, as laymen, upon difficult and disputed points of the law. Moreover, classical Roman law was ruling *only* Roman citizens, namely, only male adults, whose father was already dead and belonging to Roman families, a very small proportion of the inhabitants of the empire. Roman law has never been the clue of the empire: Egypt was ruled by Egyptian law, Greek cities by their own laws, and so on. Only in 212 AD emperor Caracalla extended, for fiscal reasons, the citizenship to all the inhabitants of the empire.

This “classical model” evolved, then, slightly over time into the opposite one, which was finally molded by Justinian, having a central court of justice at the imperial chancellery, a formal legal education at the law school in Beyrouth, and a fixed system of written sources collected into the *Corpus Juris* and universally applicable to the whole of the empire. By this fact, we can say that the final shape of Roman law, left in inheritance to the Middle Ages, was exactly the opposite of its beginnings: from an *oral* law, administered by laymen, valid only for the very few, to a written law, administered by professionals, universally valid. It is important to remember that all this happened in the East and *not* in the western part of the empire which remained a patchwork of different laws: old Roman law, canon law, and the various laws of

the “German” nations, Goths, Franks, and others, which occupied the West (Berman 1983).

This *eastern* legacy became, anyway, extremely important in the West for theological political reasons linked to the birth and development of a renewed Western Sacred Empire from Charlemagne, 800 AD, to the establishment of the first modern university in Bologna (1174 AD) and on.

The “great space” of continental Europe became to be shaped in “catholic” terms: the Sacred Empire was to be thought as a single “body,” because eating the same Holy Communion, all his inhabitants shared the same flesh. The compilation made by Justinian became to be regarded as a real “revelation” of the law for all mundane affairs not strictly confined to the church or to be left to morality. Indeed, this compilation was the only extant remain of the law, because it was written in bounded volumes of parchments, made to last, whereas all the previous scripts were on papyrus paper, necessitating to be regularly copied to be preserved, and so went quite completely lost in the barbarian west. Besides, it was much more comprehensive and well ordered than any existing barbarian compilation of laws.

In this way, nobody really enacted the Justinian compilation as positive law in the West, but it was thought to be the *ratio scripta*, the codified reason, of the law of a sacred unitary political body ontologically grounded on the Holy Communion of all its inhabitants.

This sacred, and universal, as well as *rational* character of the compilation explains why it became the basis of the university teaching of the law at Bologna, the first university established in the West, from which sprang Padua, Paris, Oxford, and Cambridge, where indeed Roman and *not* English law was taught. But the English Kingdom always refused to become a *terra imperialis* and so always refused to give any practical application to Roman Justinian law. On the continent, this common teaching shaped similarly, all over the places, the legal mind of professionals, and it was deemed applicable, as a law of reason and last resort, in all cases not patently covered by local legislation.

This legal landscape formed the era of *jus commune* in continental Europe to be broken only by the advent of modern codifications at the end of the eighteenth century. This also explains, in comparison with the English legal system, the highly intellectual character of civil law: it was a university scholarly law. Besides, on the continent, the use of writing never went completely abandoned as it almost happened in England. English *jury* trial, as an *oral* pleading, was quite a necessity given the incapacity of the jurors to read documents, whereas the continent could adopt a more sophisticated system of trial, based on documents and administered by clerks (Watson 2001).

Law and Modern Codifications: The French Model

As we have seen in the previous paragraph, continental law evolved as a *jus commune* of a common empire, based on a theory of the Justinian compilation both as sacred and as rational. Of course, the destiny of this political theological complex was to come to an end with the growing antagonism of France, Spain, and Germany, and especially with the 30-year war (1618–1648) of religion following the protestant reform.

It is out from this war that emerged on the continent the idea of the modern sovereign state.

The inter-Christian war was not terminable *but* in pure political terms: a sovereign absolute on his territories deciding also the faith of his subjects. This rising of the local princes to the status of absolute independent rulers fractured the catholic space of the empire into different territories with different jurisdictions giving rise, with the peace treaty of Westphalia (1648), to the modern system of interstate relationships known as international law. Each new sovereign became like a local, territorial bound, piece of the fractured mirror of the global universal authority of the empire, which was reflecting God’s government of the world.

It is quite natural, then, that from a concept of the sovereign, as an absolute concentration of local political power, emerged the idea that it was in the hand of this sovereign to ordain and establish the laws of his realm; and since the imagery linked to Justinian was still that of him

as the template of the lawgiver, the various monarchies tried to follow his model in projecting codes of a comprehensive, universal, and rational character for their own domains (David and Brierly 1968).

The first project was that of Frederick I of Prussia, then performed by Frederick II, leading to a *Project eines Corporis Juris Fridericiani* (1749–1751), drafted by Samuel von Cocceji. The same name of the project is displaying the Justinian ambitions of these modern sovereigns. This project led to the so-called Allgemeines Landrecht or the general laws for the Prussian states finally codified in 1794 under the supervision of Svarez and Klein, who were under the orders of Frederick the Great. This project is of extreme importance since it represents the idea that the sovereign state can shape society at it wishes and that he has not only the political power of war and peace but also that of ordering society by legislation. In this way Justinian law which was really a universal legislation served as a template for local legislations of the modern states, breaking the previously prevailing universal conception of space.

Following this German example, Maria Theresa, Empress of Austria, decided, about 1770, to charge a committee with the task of preparing a code of all her lands. After 40 years of preparatory works directed by Karl Anton Freiherr von Martini and Franz von Zeiller, this project was enacted in 1811 as the *Allgemeines bürgerliches Gesetzbuch* (ABGB), the Civil Code of the Austrian-Hungarian Empire.

What happened in between was one of the real major breaks in all European political history: the French Revolution. From a legal point of view, the revolution captured the sovereign within the state, making him no longer the possessor of the state but one of his constitutional organs, and finally sentenced the king to death for high treason, conferring an all-mighty power to the popularly elected legislative assembly. The revolutionary government went on performing a complete subversion of the existing law, hooting almost the 75% of judges, dissolving the Bar, and closing all the law schools. The new faculties of law were founded, the legal profession was

completely reorganized, and a new judiciary was established inventing the modern pyramid of courts we can find in every civil law jurisdiction. It is made of many tribunals, in quite every district, to judge on cases of first instance, then of fewer appellate courts to review their judgments, and finally of one *Cour de Cassation* established to grant a uniform application of the law.

Meanwhile, many measures were adopted to grant a legislative unity of the state, and at the end of the revolution, when Napoléon I became emperor, on 21 March 1804, he installed a commission to draft a code and, on the same year, he enacted the French *Code Civil* or *Code Napoléon*, officially the *Code civil des Français*, as a real *liberal constitution* of the civil society.

The whole apparatus to reach this goal was once again derived from Roman templates. After all the revolution was conceived to reestablish a kind of “Roman Republic,” giving back to the people all the powers and prerogatives usurped by the kings and the church; and the first title assumed by Napoléon himself was that of First Consul of this polity.

He participated to the most of the discussions in the committee and imposed a literary style to “his” code inspired by the principles of brevity and clarity, as it was thought to be a code for the commons and not for the specialists. This same code became to be surrounded by a constellation of other codes: the penal code, the code of civil procedure, the code of commerce, and the code of criminal instruction. The civil code was divided into three parts: persons, property, and “the different ways of acquiring and transferring property,” a section mainly devoted to contracts, torts, and unjust enrichment. The code is very liberal considering marriage as a contract, defending property as an absolute right, shaping contract as an agreement based on the free choices of the parties, and considering negligence as the basis of any liability.

In this way, France became the real model of any modern codified system, and her codes had an immense impact on the other countries from Italy, Poland, Spain, and Greece, to Latin American legal systems, then to Egypt, Syria, and many other systems in Africa and in Asia.

So to speak, France *is* what we have in mind today when we speak of a civil law jurisdiction.

Its main features are codes covering the whole of the legal field and a judiciary diffused all over the country and organized on the three levels of tribunals, courts of appeal, and a central court of cassation.

It is important, here, to underline the pivotal role assumed by legislation confiding to it the power to order society in all its details, because of its revolutionary political role. The center of gravity of the revolution has been the legislative assembly, and the revolution was mainly a revolution of laws, collapsing all the structures of the Ancien Régime, something which never happened in England, where this ideology of legislation was rejected also by liberals like Edmund Burke, in favor of a “sublime” conception of an oral law and an unwritten constitution as instantiated in judicial decisions, remembering, anyway, the extremely *elite* nature of the English judiciary having only *one* High Court in London, with an appellate division, submitted to the nine justices of the House of Lords (now called the Supreme Court of the United Kingdom). The French arrangement of the judiciary is extremely more diffused: the English Law Lords are nine deciding approximately 60 cases per year; at the Court of Cassation, we find more than 150 judges deciding quite 7,000 cases a year (Milo et al. 2014).

The most important point is anyway that legislation, and the rational constructivist idea of the possibility for it to *design* society, lies at the basis of the French legal system molding also the French legal style. Courts are rendering very brief decisions adopting the same style of the code, almost one page long only, whereas an English or American decision can be also 40 or 50 pages long, reporting not only the impersonal view of the court, as a unanimous organ, but all the opinions of minority and majority justices.

There is finally another factor to be remembered which is normally underscored. Parallel to the general jurisdiction, the French system adopted a special administrative jurisdiction, confined to cases involving the public administration, having its *apex* in a peculiar French institution: the *Conseil d’État*. The very existence of this

institution was singled out by authors like Dicey as the major difference between the English and the French system. In this way, the common law idea of judicial review of administrative acts is not followed in France. Normal judges have no jurisdiction over state acts: these can be questioned only behind the administrative jurisdiction and the Conseil d’État, an organ which is not only working as a court but also as a counselor of the administration in producing by-rules and acts. Under this respect, no two other systems could be more different.

Civil Law and Modern Codifications: The Rise of the German Model

As we have seen, Prussia had a code before France, but then the Napoleonic Empire extended French domination all over Europe, transplanting French patterns and methods all across the continent, up to when the French Army was defeated in Russia in 1812. The Germans lived the time between 1812 and the final defeat of Napoléon at Waterloo as an era of *national wars of liberation* against the French. After the Vienna Congress of 1815, Germany was restored but as a constellation of 39 different sovereign states: Prussia, Schleswig-Holstein, Bavaria, and so on. Anyway, its “space” (Reich) was deemed unitary from the standpoint of sharing a common culture, a common language, *and* a common university teaching. So attempts were made for having also a common legislation overpassing the differences between the various states notwithstanding the lack of a political unity (Wieacker 2003).

Thibaut was an author who sponsored the theory of adopting a German version of the French code. His idea was rejected by the most prominent German law scholar of all times Friedrich von Savigny. In an outstanding article (*Vom Beruf unserer Zeit fuer die Gesetzgebung und Rechtswissenschaft*), he traced a parallel between law and language (likely to be derived from the Scottish Enlightenment) in order to block the adoption of a foreign legislation. As the language is a complex spontaneous order, so it is the law.

Law and language are evolving orders that no single group of human minds have consciously designed nor can control. They are decentrated orders, like markets (Hayek). So it is impossible and hazardous for legislation, as a consciously designed order, to try to mold the whole of society. Society is different from the state, which is one of the many purposive organizations pursuing their goals within society. It follows that the overall order of society cannot be designed, but can only evolve piecemeal.

This theory is rather understandable if we remind that there wasn't a unitary state in Germany, so that effectively there was no possibility for a central authority to mold the law, nor there was any unitary judiciary to promote it. What was unitary in the various German states was the university system. A student could also spend a term in Munich and the next term in Berlin; and what Savigny proposed, after the feelings raised by the very conception of the wars of liberation to build up a newer Germany, was to entrust the development of the law to the legal science (*Rechtswissenschaft*) as practiced by the German *professoriat*.

If law is like language and language is a depositary of culture, it makes no sense to adopt a foreign law and destroy our culture while engaging in liberation and the making of renewed Germany. Law and language lie in the *spirit of the people* (*Volksgeist*). Only a scholar can have a good insight over it, because of his learning, to be able to produce a well-conceived framework of concepts to give it voice, creating a kind of scholarly made law (*Juristenrecht*) different from both judicial-made law and from legislation. And, after all, Germany was to be considered as the real heir of the "space" of the empire (*Reich*), and as such went on, and was going on, elaborating the *jus commune*, the actualized version of the Roman law. This law was not a piece of ancient history in Germany but an *actual* system of living law. In this way, Roman law was no more an alien system, but it really became, in many centuries, part of the national spirit. Indeed, Savigny's major work was entitled *Der System des heutigen Roemischen Rechts*, "The System of the Actual Roman Law."

Here, we may find a version of the civil law totally opposite to that of the French. Where France claims to be "republican," but she is indeed the continuation of the imperial model of Justinian, entrusting law to legislation, with the possibility of a political design of society, here, Germany is representing the ideal of the "classical" Roman law as a law practically *without* legislation, and certainly without codes, slightly evolving through learning, as the great jurists of Rome did *before* Justinian and as the great lawyers of the *jus commune* did after Bologna. France is claiming a continuity with Roman templates of codification, but Savigny is claiming a deeper and strong continuity where legislation is but an episode of a much more complex story of the civil law tradition.

If we perceive this, we can easily spot how codes are an unnecessary feature of a civil law system and maybe are contrary to its original nature.

Savigny prevailed against Thibaut and Germany went on developing "scientifically" the Roman law. But when, with the war of 1866 against Austria and of 1870 against France, Germany was unified in the form of the Second Empire, the pressure for having a common legislation became too strong. This pressure could anyway be filtered by the already established institution of the *professoriat* as a real factor of the legal progress. So professors started to work on the idea of making a new code different from the French one and based on the "concepts" used to elaborate their own actualized version of Roman law (*Begriffsjurisprudenz*), especially Windscheid, a well-known author of one of the major textbooks on the *Pandects* paved by his scholarship, the way to a first draft of the code in 1888. A committee of 22 members, comprising not only jurists but also representatives of financial interests and of the various ideological currents of the time, compiled a second draft. After significant revisions, the BGB (*Burgerliches Gesetzbuch*, Civil Code) was passed by the Reichstag in 1896. Political authorities gave 4 years to the legal profession to study and learn the new legislation, which was put into effect on 1 January 1900 and has been the central codification of Germany's civil law ever since.

The BGB served as a template for several other civil law jurisdictions, including Portugal, Estonia, Latvia, Japan, Brazil, and Greece. It never had, anyway, the same world impact as the French code. What had a tremendous impact all over the civil law countries were German scholarship and the German method strongly influencing Italy, Spain, Latin America, and quite all the jurisdictions that maintained a French-like legislation (Merryman and Pérez-Perdomo 2007).

So, after all, also Germany became a codified system, and quite all civil law jurisdictions can be deemed to be a “hybrid” of French legislation and German scholarship.

What is peculiar is that the two codes, French and German, are really very different. The German code, especially, possesses a General Part (*Allgemeiner Teil*), which does not exist in the French code. In this General Part, we can find all the general concepts to be adopted to grasp the specific parts devoted to contracts, torts, and property. This different approach is obviously indebted to the fact that this code has been elaborated by professors and that they have been able to act as a unitary factor to reach a national goal.

Anyway, the Germans structured the judiciary in quite the same French way and maintained a separate administrative jurisdiction as in France.

Conclusion: The Problems of Harmonization and of Comparison Between Common and Civil Law Jurisdictions

All this, the mixing of the French and German patterns, is giving to civil law, considered as a general tradition, her intellectualistic flavor as well as her pro-legislation biased aspect.

When we speak of civil law jurisdictions, we mean systems that (1) have codes, (2) have a similar and diffused judiciary handling many more cases than a common law jurisdiction, (3) possess a separate – seemingly pro-state biased – administrative jurisdiction, and (4) know a much stronger and active role in the legal development of scholars and universities (van Caenegem 2001).

Notwithstanding this general image of the civil law, there are some myths to deconstruct about the comparison of civil and common law systems. First of all one is the myth that civil law *is* legislation and common law *is* a judge-made law.

Today, the most of legal matters in common law countries are covered by statutory law. Corporate governance, for instance, is always legislative also in these countries, as it is sale of goods or secured transactions. On the other side, it is true that the legislation of the continental codes is very broadly conceived, so that the role of judges in developing the sense of the codes cannot be underestimated. Case law is as important to understand a provision of a civil code as it is to know what the common law is on a certain point.

Second, it is not true that legislation is a permanent and overwhelming factor in civil law countries. They lived for centuries without codification, and we may find, as in the case of Savigny, theories of the essence of the civil law which are directly antagonistic to the role of legislation.

Third, it is true that the civil law appears more “conceptualized,” for the role always played by universities in her elaboration, but we cannot overpass the role of theory in the United States. It would be hard to consider American law without considering that each case is based upon a *doctrine* and that it is much more American scholarship, than state case law, to give a picture and a frame of what this law is and to influence the rest of the world, and we cannot bypass the role of great law schools in the practical organization of the elite of the legal profession, their ways of thinking, of elaborating solutions, and so on. From a civilian perspective, an American piece of legal scholarship is much more based on theory than it is, today, an average civil law writing displaying more erudition and knowledge than intellectual claims.

It is rather to be accepted that both families are a different compound of different factors always acting, sometimes in competitive ways, in the legal history: legislation, judicial decisions, and scholarly writings. The different mixtures of these elements are marking the difference between France and England, but it is marking the difference between England and the United States, also,

as it marks a difference between France and Germany (Ginsburg et al. 2014).

What is really different in common and civil law is the figure of the judge and the fact of having a separate administrative jurisdiction.

Judges in common law are fewer and decide a much lesser number of cases. This is something in search for an explanation. There are approximately 6,000 judges in France and 600 judges in England. Besides, a common law judge is an old member of the Bar (the United Kingdom), or she is directly appointed by the political power at state or federal level (the United States). A civil law judge is the winner of a public competition for recruitment. It means that you become judge when you are young, just maybe practicing the law for a few years, and then you make a *judicial career* from the last of tribunals to the chair of the president of the Court of Cassation, whereas there is scarcely something as a judicial career in the United States, so few being the case of persons appointed as state or federal circuit judges then becoming appointed at the Supreme Court. Under this respect, the two systems cannot be more divergent. This factor depends heavily on the costs of justice. Civil law is cheaper, and that's also why it is normally longer; but no serious attempt has been made to understand precisely why, and this certainly does not depend on Roman origins.

The fact of having a separate administrative jurisdiction is also of extreme relevance. This fact, again, cannot be traced back to the Roman origins of the civil law systems; rather, it is a by-product of political modernity: the rise of an absolute state on the continent and the absence of a political upheaval similar to French revolution in the common law world.

It is strange to note the following paradox: in common law, ordinary jurisdiction is much more politicized in the sense that the judge can be appointed directly by the political power, but the civil law is granting more room for state action by creating an administrative compartment separated from ordinary jurisdiction.

But is the separation of ordinary and administrative jurisdictions connaturate to a civil law tradition? One could really wonder. For centuries,

again, there was not such a separation, and it is much more likely to be due to the *form* assumed by political power on the continent of Europe than to deep legal structures linked with distant origins.

Finally, what is certainly absolutely distant, even today, is the *style* of these two families of laws. There is scarcely any similitude between a French and an American judicial decision, as there is not a common way to handle precedents, and also the modes of interpreting statutes are rather distant. In a sentence we could say that the apparently politically flat world of globalization is still *striped*, fractured, and discontinued by the *legal styles* (Zweigert and Koetz 1998).

To what extent, if any, these legal styles have an economic impact is a question open to investigation. What it certainly represents is a legal *duality* of the West, and especially of Europe, displaying two different appearances of what we call justice, rendering any work for harmonization harder than expected.

Summary/Conclusion/Future Directions

There are two different main versions of the civil law, German and French, as there are similarities and differences between civil and common law which are hard to grasp. The major difference lying in the different *styles* of these legal traditions hampering any actual conscious work of harmonization as they represent complex spontaneous orders which can but imperfectly been managed by purposive design.

Cross-References

- [Administrative Law](#)
- [Globalization](#)
- [Law and Economics, History of](#)

References

- Berman HJ (1983) Law and revolution: the formation of the western legal tradition. Harvard University Press, Cambridge, MA

- David R, Brierly JE (1968) Major legal systems in the world: an introduction to the comparative study of law. The Free-Press/Collier Macmillian, London
- Ginsburg T, Monateri PG, Parisi F (2014) Classics in comparative law. Edward Elgar, Cheltenham/Northampton
- Glenn HP (2000) Legal traditions of the world: sustainable diversity in law. Oxford University Press, Oxford
- Merryman J, Pérez-Perdomo R (2007) The civil law tradition: an introduction to the legal systems of Europe and Latin America. Stanford University Press, Palo Alto
- Milo JM, Lokin JHA, Smits JM (2014) Tradition, codification and unification: comparative historical essays on developments in civil law. Intersentia Ltd, Cambridge, UK
- Samuel G (2014) An introduction to comparative law theory and method. Hart Publishing, Oxford/Portland
- van Caenegem RC (2001) European law in the past and the future: unity and diversity over two millennia. Cambridge University Press, Cambridge, UK
- Watson A (2001) The evolution of Western private law. John Hopkins University Press, Baltimore
- Wieacker F (2003) A history of private law in Europe with particular reference to Germany. Clarendon, Oxford
- World Bank (2003) Doing business in 2004. Understanding regulation. The World Bank, Washington, DC
- Zweigert K, Kötz H (1998) Introduction to comparative law. Oxford University Press, Oxford

Class Action and Aggregate Litigations

Giovanni Battista Ramello
DiGSPES, University of Eastern Piedmont,
Alessandria, Italy
IEL, Torino, Italy

Synonyms

[Collective redress](#); [Mass tort litigation](#); [Small claims litigation](#)

Definition

Class action and other forms of aggregate litigation introduce in the legal procedure a powerful means for gathering dispersed interests and channeling them into a type of action in which the different parties concur to promote individual

and social interest. They can restore the full working of the legal system, and in addition they can be a powerful device for promoting social welfare when other institutional arrangements seem to be ineffective or inefficient.

Introduction

Class action and other forms of aggregate litigation are the answer to an organizational puzzle in civil procedure dealing with reconciling enforcement of the dispersed victims' rights, the lack of proper incentive for promoting a legal action, and the social interest of producing the public goods of deterrence and, possibly, regulatory change (Ramello 2012). The underlying problem is the twofold partial or total failure of individual litigation and regulation which is essentially explained by the fact that neither institution is able to produce the appropriate incentives for obtaining an appropriate outcome (Dam 1975). In other words, these two production "institutional technologies" are unfitted for the context in which they operate, so that the solution must go by some alternative route.

The economic argument here mirrors that used for explaining the emergence of hierarchies when there is a need to internalize externalities, for example, in the well-known problem in economics of indivisibility in production, which arises in the case of economies of scale (or scope), and makes it impossible to rely on the competitive market for optimal allocation of resources (Edwards and Starr 1984). Indivisibility plays a prominent part in the understanding of industrial organizations and of course likewise affects the market structure. In consequence, the different organizations and multiple forms of enterprises in the market, and of aggregate ventures in the judicial market, should be regarded as institutional solutions designed to achieve adequate productive configurations for specific contexts. The same applies to dispute resolution industry.

It is worth noting that the creation of a hierarchy defines an exclusive right over the specific productive activity. Such a right, in the judicial market, corresponds to a specific legal action and

thus in practice means creating a local monopoly on a particular litigation. This aspect is by no means peripheral to the incentive system in the case of collective redress: it is a prerequisite for being able to assign a property right over the potential rewards of the legal action. Such a right, in its turn, becomes the central element (i.e., the price) for achieving transfer of risk through a contingent fee reward scheme (Eisenberg and Miller 2004, 2013; Sacconi 2011). The party financing the legal action – often the attorney – thus obtains the right to extract a portion of the awarded proceeds as a remuneration for the risk.

Class Versus Aggregate litigation

The currently dominant reference models for aggregate litigation, including class action, are those of the US legal system, whose Rule 20, Rule 23, and Rule 42 of the Federal Rules of Civil Procedure and Section 1407 of Title 28 of the US Code, taken together, introduce various ways of pursuing aggregate litigation in the form of class action, multi-district litigation, formal consolidation, and other solutions, thereby redrawing the boundaries of litigation (ALI 2010).

Rule 23 is the most well known, in that it introduces class action, which has the role of exhausting in a single litigation all possible claims of a predefined population of victims. Among the technicalities of class action, there is also the indirect representation of victims who are unable to join the legal action on their own account (so-called absent parties). The other solutions, in a more fragmentary way, promote collective or coordinated legal actions which, for example, “involve a common question of law or fact [and in which] the court may: (1) join for hearing or trial any or all matters at issue in the actions; (2) consolidate the actions; or (3) issue any other orders to avoid unnecessary cost or delay” (Rule 42a, 2009 edition).

While the specific technical features of each procedural solution are discussed elsewhere (Hensler 2001, 2011; Calabresi and Schwartz 2011), in all cases one of the key criteria for

choosing between them is efficiency – meaning the extent to which the aggregation is able to pursue expedition and economy.

Hence, the different forms of aggregation can be compared to the different types of business entities (e.g., public company, joint venture, etc.), whose function is to best exploit the advantages of the hierarchy in different situations. Under this analogy, in the productive organization of the judicial market, class action lies at one extreme, since it exhausts in a single litigation the claims of a broad population of victims who become shareholders in the legal action (essentially a sort of public company). The other solutions occupy intermediate positions, making it possible to exploit some benefits of aggregate litigation even in situations where all the victims cannot join in a single lawsuit, so that a class action is not practicable (and might in fact even be invalidated).

Sketches of History

Although Yeazell (1987) has detected a precursor to class action (and aggregate litigation) in the medieval group litigation of England, class action which is somewhat the benchmark for aggregate litigation was first introduced in the US legal system in 1938, through Rule 23 of the Federal Rules of Civil Procedure. It then took nearly three decades for class action to be fully implemented into the US civil procedure, with the 1966 issuing of the new version of Rule 23 by the Supreme Court. Since then, class action has been fiercely criticized by a number of opponents (Hensler et al. 2000; Klement and Neeman 2004). Despite the negative stances, it has over the years become “one of the most ubiquitous topics in modern civil law” in the USA and nowadays one of “[t]he reason for the omnipresence of class actions lies in [its] versatility” (Epstein 2003, p. 1) which, according to a great many commentators, can make it an effective means for serving justice and efficiency in a broad sense.

The collective litigation system thus continues to operate and to develop, and its utility remains undisputed in the North American judicial system.

The most recent amendment, brought by the Class Action Fairness Act (CAFA 2005; Pub. L. No. 109–2, 119 Stat. 4, 2005), though aimed according to some authors at curbing some of its pernicious features (Lee and Willging 2008), carefully avoided criticizing collective litigation as a whole and in fact reaffirmed its substantive validity, strongly asserting that “class-action law suits are an important and valuable part of the legal system when they permit the fair and efficient resolution of legitimate claims of numerous parties by allowing the claims to be aggregated into a single action against a defendant that has allegedly caused harm” (CAFA 2005, Sect. 2).

In other countries and especially in Europe, aggregate litigations and collective redress systems have recently been introduced. However, local constraints especially derived from the specific legal culture – such as, e.g., the revulsion at accepting the role of the entrepreneurial activity sometime needed in order to trigger the legal action – or simply the political aversions have produced outcomes very distant from the American model, sometime raising substantial concerns about the real effectiveness (Hilgard and Kraayvanger 2007; Baumgartner 2007; Issacharoff and Miller 2012; for a perspective on distinct European countries, see Backhaus et al. (2012) and the contributions therein).

Procedural Features

The first effect of class action and with some variance also of other forms of aggregate litigation is to permit the adjudication of meritorious claims that would otherwise not be litigated due to imperfections in the legal systems (Rodhe 2004). In fact, class action is a legal device employed today for tackling torts in a wide array of cases, including insurance, financial market, and securities fraud in recent times (Pace et al. 2007; Helland and Klick 2007; Porrini and Ramello 2011; Ulen 2011). However, from its inception, class action was infused with a broader political agenda, extending beyond the tort domain to embrace matters such as civil rights (in particular segregation), health protection, consumer

protection, environmental questions, and many others (Hensler et al. 2000). This legacy is sometime emerging in other aggregate litigations.

As a whole, collective actions have the effect of altering the balance of power and the distribution of wealth among the various social actors – e.g., firms versus consumers – thereby extending their scope in terms of overall impact on society. All the above elements, taken together, thus play an important role in guiding the legislator’s decision of whether (or not) to adopt aggregate litigations, and, ultimately, the battle in favor of or against the introduction of these procedural devices into the different legal systems is played out on a purely political terrain (Porrini and Ramello 2011).

Indeed, it is the procedural technicalities that have for the most part given skeptics grounds for criticizing class action and questioning its ability to be implemented in legal systems different from those where it arose. These are often specious arguments which disregard the simple fact that any “juridical technology” intended to achieve certain outcomes must be adapted, in its design, to the constraints of the target legal system, if it is to provide regulatory solutions that are effective and compatible with its context. The heart of the problem, therefore, consists in opportunely adapting the “legal machinery” to each jurisdictional setting in a manner that obtains the desired results without prejudicing its essential features. These characterizing features can be:

- (i) The aggregation of separate but essentially cognate claims, united by design and not by substantive theory
- (ii) The indirect representation of absent parties (in the case of class action)
- (iii) The provision of entrepreneurial opportunity to an attorney, who thus becomes the main engine of the civil action

Despite different forms of aggregate litigations rely upon distinct features, the common element is that they all try to a great extent to eliminate duplications in related claims, by aggregating in some way the potential claimants into a group. The obvious main consequence is that, by

aggregating in some way similar claims, aggregate litigations increase the possibility to vindicate a tort or however they redress the imbalance which exists between plaintiffs and defendants in several areas of litigation.

The indirect representation, essentially characterizing class action, stems from the fact that the attorney is not appointed directly by each individual claimant, but rather through a specific set of procedures established by law, which essentially rely upon the initiative of a minority among them, and the subsequent acceptance by the judge, to start the trial (Hensler et al. 2000). In fact, the civil action is filed by an individual or a small group of victims assisted by an attorney. The class is then certified by the judge who consequently also “appoints” the attorney as a representative of all the class members (Dam 1975).

It is worth noting that the mere appointment of an attorney does not, of course, per se assure attainment of any efficient outcome, nor does it rule out opportunistic behaviors (Harnay and Marciano 2011). It is only a first step for making the desired outcomes possible and, as usual in tort litigation, demands a well-designed set of incentives for the lawyer in order to work properly (Klement and Neeman 2004; Sacconi 2011).

It is worth reminding that in a sense collective actions bear some similarities – albeit limited to the civil procedure domain – to regulation: in fact, where the judge determines that individual actions may not be sufficiently effective, yet the litigation is in the collective interest, on request of a representation of victims, he or she reallocates the individual rights over that particular prospective litigation. Thus, also in this case, an agent is nominated to represent the interests of a group, but with a narrower scope compared to fully fledged regulation. Here, the indirect representation serves merely to exploit the possibility of aggregating related claims without bearing the costs of searching for and coordinating a huge number – often a “mass” – of potential plaintiffs that would otherwise make bringing the lawsuit unaffordable (Cassone and Ramello 2011).

Finally, there is one last feature that makes collective action possible: it is the creation of a specific entrepreneurial space for the class

counsel, who undertakes to identify an unmet demand for justice and, acting self-interestedly, restores access to legal action for the victims. The class counsel is generally driven by the purely utilitarian motives of a “bounty hunter,” who offers a service in exchange for recompense (Macey and Miller 1991; Issacharoff and Miller 2012). It is thus a behavior consistent with the paradigm of methodological individualism and which is sometimes regarded with suspicion by those who consider private interests unsuitable for representing the collective interest.

Aggregate litigation thus has the particular merit of aligning the private interest of the case attorney, who seeks to obtain a profit, with that of the victims, who seek redress of the harm and promotion of justice, and with that of society which instead benefits from a system that internalizes the externality. This in fact creates a deterrent to wrongdoing and ultimately may work to minimize the social costs of accidents, in accordance with the Hand’s rule (Calabresi 1970). In this light, therefore, the miracle of the invisible hand is again renewed, and the self-interest of the victims and class counsel can play a role of public relevance.

Law and Economics Features

Law and economics is further brought into play when we consider the wider effects of aggregate litigations on the judicial system and on the economic system. In particular, economic science offers two complementary routes for conducting the analysis. The first concerns the manner in which these litigations can serve efficiency and collective welfare; the second provides the analytical framework for representing the legal machinery and studying its workings, thus determining under what conditions and in what way they promote social welfare.

However, for the investigation to be fruitful, we have to specify the initial conditions, i.e., the circumstances under which regulation and individual action are not effective. In other words, we must define the context that gives rise to some shortcomings justifying the introduction of new

legal devices. The conditions may be the following:

- Existence of *fragmented claims*, very often worth less to each plaintiff than the individual litigation cost or which in any case entail a prohibitively costly individual litigation
- Sufficient *homogeneity of claims* for the court to issue a “one size fits all” decision and for the victims to be able to adhere to the collective action
- A *judicial market failure*, as a result of which some claims, no matter how meritorious, are either not brought, so that certain individuals are unable to exercise their rights, or are imperfectly exercised
- A *failure of regulation* which thus does not offer a practicable alternative for resolving the preceding issues

A condition under which aggregate litigation is potentially useful is that where certain rights established by law are not exercised or only imperfectly exercised, due to a misalignment between what is theoretically asserted by the law and the concrete incentives provided to individuals. The solution involves an institutional reorganization to produce a lowering of these costs and/or promote the – sometimes forced – reallocation of the rights.

By thus regarding victims as owners of “property rights” over a specific litigation, whose enforcement may incur costs exceeding the expected individual benefits, we can interpret class action as a system that follows a comparable judicial path to that described for property, aggregating the individuals’ rights when their exercise on the judicial market is precluded (or limited) by contingencies which make the net benefit of the action negative (Ramello 2012).

In general, these contingencies arise from the aforesaid fragmentation and its attendant coordination costs, from the limited size of the individual damages (so-called small claims), and also from the existence of asymmetries between the would-be plaintiffs and defendant (i.e., availability of information, capacity to manage the litigation risk, access to financial resources, and more).

Creating a pool of rights thus enables victims to access a less costly litigation technology and thereby pursue justice. The productive efficiency of a static character concerns the overall production of “justice,” on the demand and supply sides, since on the judicial market, both jointly concur to its production, albeit for different reasons. Aggregate litigations in fact allow a so-called judicial economy to emerge, which on the demand side, e.g., through aggregation of small claims, produces economies of scale in litigations that cause individual costs to decrease with increasing number of plaintiffs (Bernstein 1977). On the supply side, there is likewise a reduction in costs if the aggregation permits overall savings in resources compared to multiple individual actions, provided though that the savings afforded by aggregation are not offset by an increase in the number of lawsuits.

There is, then, a third level of efficiency connected with the economic nature of aggregate litigation and which has the purpose of aligning different interests to achieve the previously stated goal. In effect, the system, if properly applied, has to introduce a set of distinct incentives which together concur to produce three different outputs: a profit for the attorney, redress of the harm for the victims, and deterrence of wrongdoing (thereby minimizing the social cost) for society.

Using the traditional categories of economic analysis and with special reference to class action, Cassone and Ramello (2011) disentangle the “productive” roles of the various actors taking part in the litigation. In other words, the role of these legal procedural devices reconciles the conjoined individual interests of victims with the collective interest of society, by passing through the private interest of the attorneys. It thus has the nature of a *private good* for the attorney, who takes on the entrepreneurial role of setting in motion or managing the collective action, which is in its turn aimed at obtaining redress of the harm (Dam 1975). Though this ultimately has an effect on each victim, it can only be produced as a local public good for the cohort of all victims and thus takes the form of a *club good* (Cassone and Ramello 2011). Finally, the transfer of the cost of the wrongdoing from the victims to the injurer

has the consequence of reestablishing a higher level of deterrence, thereby resulting in production of a *public good* (Eisenberg and Engel 2014). This deterrence, it is worth noting, pertains to what is generally termed dynamic efficiency, since its production in a given time frame is also instrumental to the intertemporal optimal production of other goods. In fact, besides the usual production of deterrence, as in general done by tort law, there can also be the production of inputs for regulation, thus establishing a causal relation between litigation and regulatory rule-making. In this respect, aggregate litigations have the additional feature of producing information externalities and consensus among a broad panel of individuals acting as a proxy for the society that serve as direct inputs to regulation. Then the litigation becomes a sort of R&D laboratory, in which plaintiffs act as a proxy for society and the judicial solution serves as a prototype for regulatory change (Arlen 2010; Ramello 2012).

Cross-References

► [Mass Tort Litigation: Asbestos](#)

References

- ALI (2010) Principle of the law of aggregate litigation. American Law Institute, Philadelphia
- Arlen J (2010) Contracting over liability: medical malpractice and the cost of choice. *Univ Penn Law Rev* 158:957–1023
- Backhaus JG, Cassone A, Ramello GB (2012) The law and economics of class actions in Europe. Lessons from America. Edward Elgar, Cheltenham
- Baumgartner SP (2007) Class actions and group litigation in Switzerland. *Northwest J Intl Law Bus* 27(2):301–349
- Bernstein R (1977) Judicial economy and class action. *J Legal Stud* 7(2):349–370
- CAFA 2005 is an acronym for Class Action Fairness Act written in the text the first time. Hence it does not need reference
- Calabresi G (1970) The cost of accident. Yale University Press, New Haven
- Calabresi G, Schwartz KS (2011) The costs of class actions: allocation and collective redress in the US experience. *Eur J Law Econ* 32:169–183
- Cassone A, Ramello GB (2011) The simple economics of class action: private provision of club and public goods. *Eur J Law Econ* 32:205–224
- Dam KW (1975) Class actions: efficiency, compensation, and conflict of interest. *J Legal Stud* 4:47–73
- Edwards BK, Starr RM (1984) A note on indivisibilities, specialization, and economies of scale. *Am Econom Rev* 77:192–194
- Eisenberg T, Engel C (2014) Assuring civil damages adequately deter: a public good experiment. *J Empir Legal Stud* 11:301–349
- Eisenberg T, Miller GP (2004) Attorney fees in class action settlements: an empirical study. *J Empir Legal Stud* 1:27–78
- Eisenberg T, Miller G (2013) The english vs. the American rule on attorneys fees: an empirical study of attorney fee clauses in publicly-held companies' contracts'. *Cornell Law Rev* 98:327–382
- Epstein RA (2003) Class action: aggregation, amplification and distortion. University of Chicago, Legal Forum, pp 475–518
- Harnay S, Marciano A (2011) Seeking rents through class actions and legislative lobbying: a comparison. *Eur J Law Econ* 32:293–304
- Hensler DR, Dombey-Moore B, Giddens E, Gross J, Moller E, Pace M (eds) (2000) Class action dilemmas. Pursuing public goals for private gain. Rand Publishing, Santa Monica/Arlington
- Helland E, Klick J (2007) The trade-off between regulation and litigation: evidence from insurance class actions. *J Tort Law* 1:1–24
- Hensler DR (2001) Revisiting the monster: new myths and realities of class action and other large scale litigation. *Duke J Comp Int Law* 11:179–213
- Hensler DR (2011) The future of mass litigation: global class actions and third-party litigation funding. *George Wash Law Rev* 79:306–323
- Hilgard MC, Kraayvanger J (2007) Class action and mass action in Germany. International Bar Association Litigation Committee Newsletter, September, pp 40–41
- Issacharoff S, Miller GP (2012) Will aggregate litigation come to Europe? In: Backhaus JG, Cassone A, Ramello GB (eds) The law and economics of class actions in Europe. Lessons from America. Edward Elgar, Cheltenham
- Klement A, Neeman Z (2004) Incentive structures for class action lawyers. *J Law Econ Org* 20:102–124
- Lee EG, Willging TE (2008) The impact of the class action fairness act on the federal courts: an empirical analysis of filings and removals. *Univ Penn Law Rev* 156:1723–1764
- Macey JR, Miller GP (1991) The plaintiffs' Attorney's role in class action and derivative litigation: economic analysis and recommendations for reform. *Univ Chicago Law Rev* 58:1–118
- Pace NM, Carroll SJ, Vogelsang I, Zakaras L (2007) Insurance class actions in the United States. RAND, Santa Monica
- Porrini D, Ramello GB (2011) Class action and financial markets: insights from law and economics. *J Fin Econ Policy* 3:140–160
- Ramello GB (2012) Aggregate litigation and regulatory innovation: another view of judicial efficiency. *Int Rev Law Econ* 32(1):63–71

- Rodhe DL (2004) *Access to justice*. Oxford University Press, Oxford/New York
- Sacconi L (2011) The case against lawyers' contingent fees and the misapplication of principal-agent models. *Eur J Law Econ* 32:263–292
- Ulen TS (2011) An introduction to the law and economics of class action litigation. *Eur J Law Econ* 32:185–203
- Yeazell SC (1987) *From medieval group litigation to the modern class action*. Yale University Press, New Haven

Climate Change Remedies

Donatella Porrini
Department of Management, Economics,
Mathematics and Statistics, University of Salento,
Lecce, Italy

Abstract

The law and economics analysis of the climate change remedies has been focused on the question of which would be the policy instrument most suited to provide incentives to reduce greenhouse gas emissions. The literature focuses mainly on the comparison of carbon taxes and emission trading scheme. But a relevant role can be played by financial and insurance instruments, especially considering the adaptation and mitigation strategies. Finally, another instrument is considered, largely used to internalize other environmental externalities but still not so much analysed for climate change, the liability system.

Definition

Over the last centuries, climate change has become a very important issue all over the world. The change in climate corresponds to an increase in the earth's average atmospheric temperature, which is usually referred to as global warming.

In response to scientific evidence that human activities are contributing significantly to global climate change, and particularly the emissions of greenhouse gas (GHG) emissions, decision-makers are devoting considerable attention to

find remedies to reduce the consequences in terms of climate change.

The Concept of “Economic Global Public Goods”

Dealing with climate change implies the concept of “economic global public goods” that can be defined as goods with economic benefits that extend to all countries, people, and generations (Kaul et al. 2003).

First of all, the emissions of GHG have effects on global warming independently of their location, and local climatic changes are completely linked with the world climate system.

In addition, the effects of GHG concentration in the atmosphere on climate are intergenerational and persistent across time.

The fact that climate change is clearly “global” in both causes and consequences implies that, on one side, we cannot determine with certainty both the dimension and the timing of climate change and the costs of the abatement of emissions, on the other side, it emerges a relevant equity issue among countries because industrialized countries have produced the majority of GHG emissions, but the effects of global warming will be much more severe on developing countries.

About this last point, the countries that have more responsibilities will face less consequence in the future and vice versa. So it is a global issue to decide the distribution of emission reductions among countries and how the costs should be allocated, taking into account the differences among countries characterized by high- or low-income, high- or low- emissions level, and high and low vulnerability.

Climate change is going to generate natural disasters, meaning events caused by natural forces that become “man-made” disasters, meaning events associated with human activities, given the role of greenhouse gases emitters. More precisely, we can speak of “unintended man-made” disasters originated by global warming (Posner 2004, p. 43).

The rising costs associated with climate change effects pose serious challenges to governments to adopt efficient strategies to manage the increasing

economic consequences, and governments are facing the issue to introduce policies to tackle the causes and combat all the effects of greenhouse gas emissions.

Dealing with global public goods, the choice of environmental policies requires a global coordination (Nordhaus 2007). But, in any case, it is difficult to determine and reach agreement on efficient policies because economic public goods involve estimating and balancing costs and benefits where neither is easy to measure and both involve major distributional concerns. As a consequence, it is necessary to reach through governments to the multitude of firms and consumers who make the vast number of decisions that affect the ultimate outcomes.

Carbon Tax and Emission Trading Scheme

The policy instruments that are mainly implemented as remedies against climate change are carbon tax and emission trading scheme (ETS).

A carbon tax is a particular levy on GHG emissions generated by burning fuels and biofuels, such as coal, oil, and natural gas. It is generally introduced with the main goal to level the gap between carbon-intensive (i.e., firms based on fossil fuels) and low carbon-intensive (i.e., firms that adopt renewable energies) sectors.

A carbon tax provides a strong incentive for individuals and firms to adjust their conduct, resulting in a reduction of the emissions themselves because the relative price of goods and services changes. Hence, by decreasing fuel emissions and adopting new technologies, both consumers and businesses can reduce the entire amount they pay in carbon tax.

An emission trading scheme (ETS) is an instrument based on an agreement that sets quantitative limits of emissions and the allocation of emission, allowing the trade in order to minimize abatement costs. At the beginning the allocation of permits can be set through either an auction or a grandfather allocation. Under an auction, government sells the emission permits, whereas under the

grandfather rule, the allocation of emission permits is based on historical records.

An ETS is defined as a quantity-based environmental policy instrument. It is also called cap and trade because it is characterized by the allowable total amount of emissions (cap) and the right to emit that becomes a tradable commodity. Under an ETS system, prices are allowed to fluctuate according to market forces.

On the other hand, carbon taxes are defined as price-based policy instruments for the correlated effects to increase the price of certain goods and services, thereby decreasing the quantity demanded.

An emission trading system may efficiently give the incentive to decrease the emissions wherever abatement costs are lowest with positive effects beyond the national borders. As costs associated with climate change have no correlation with the origin of carbon emissions, the rationale for this policy approach is that an emission trading system allows to fix a certain environmental outcome and the companies are called to pay a market price for the rights to pollute regardless of where there will be the benefits. This is the reason why an emission trading system is suitable for international environmental agreements, such as the Kyoto Protocol, and also for the characteristic that a defined emission reduction level can be easily agreed between states.

Emission trading can be an advantage for private industry because, by decreasing emissions, firms can actually profit by selling their excess GHG allowances, in a way that such a market of permits could potentially drive emission reductions below targets.

A system of ETS entails significant transaction costs, which include search costs, such as fees paid to brokers or exchange institutions to find trading partners, negotiating costs, approval costs, and insurance costs. Conversely, taxes involve little transaction cost over all stages of their lifetime.

Carbon taxes are economic instruments that works dynamically offering a continuum incentive to reduce emissions. In fact, technological and procedural improvements and their subsequent efficient diffusion lead to decreasing tax payment. On the other hand, trading systems will adjust

when the emission goals are easier to meet, so that in this case a decreasing demand of permits causes a reduction in their price but not as rapidly as taxes.

The law and economics literature describes as alternative instruments carbon tax and tradable permit system, the former as a price control instrument and the latter as a quantity control one.

Many contributions compare the relative performance of price and quantity instruments under uncertainty, starting with the seminal contribution of Weitzman (1974). For example, Kaplow and Shavell (2002) deal with the standard context of a single firm producing externality; moreover, they consider the case of multiple firms that jointly create an externality, concluding with the superiority of taxes to permits.

In the case of climate change, there are arguments for price controls. The first point is that climate change consequences are uncertain because it is not the level of annual emissions that matters, but rather the total amount of GHG that have accumulated in the atmosphere. The second point is that “while scientists continue to argue over a wide range of climate change consequences, few advocate an immediate halt to further emission” (Pizer 1999, p. 7).

Even if a carbon tax is preferable to an ETS scheme in terms of social costs and benefits, this policy obviously faces political opposition. Private industry opposes carbon taxes because of the transfer of revenue to the government; environmental groups oppose carbon taxes for an entirely different reason: they are unsatisfied with the prospect that a carbon tax, unlike a permit system, fails to guarantee a particular emission level.

The Role of Financial and Insurance Instruments

To face climate change economic consequences, a role can be assigned also to private sector to stimulate the reduction of the probability of catastrophic losses and to manage economically large-scale disaster risks. In this sense a relevant part can be played by the financial and insurance products that are based on mechanisms to manage the

economic consequences of risk, including the threat posed by natural hazards.

With the typical insurance contract, for example, individuals and companies protect themselves against an uncertain loss by paying an annual premium toward the pool’s expected losses. The insurer holds premiums in a fund that, along with investment income and supplementary capital (where necessary), compensates those that experience losses.

First of all, climate change consequences are insured through the coverage of the risks that insurance companies accept from their customers, since policies already include the provision of the economic consequences of changes in the intensity and distribution of extreme weather events and of the resulting risk of catastrophic property claims (Porrini 2011).

The supply of this kind of products, that are the core business of the insurance industry, experiences some problems.

First, climate change’s relationship to global weather patterns increases the potential for losses so large that they threaten the solvency of the insurance companies.

Second, uncertainties in assessing climate change’s impacts are high, affecting property and casualty, business interruption, health, and liability insurance, among others. As a result, insurers could charge a significantly higher premium or, in certain cases, avoid to supply this kind of policies.

Third, many climate change-related risks may be correlated, creating a skewed risk pool and exacerbating the risk of extremely large losses, and that some of these risks are not well distributed across existing insureds.

Beyond the problems of insurability, financial and insurance market provide for other kind of products. Examples are “compensation funds,” such as special government disaster funds, to promote framework of contingency measures to tackle climate change consequences. These funds, created in connection with a regulatory system mainly to cover environmental damage and victims’ compensation, can be financed by a taxation system or by a firm’s contribution system. The main example is the Superfund in the

United States, connected with the regulatory system by Environmental Protection Agency (EPA).

Other examples are products characterized by ex ante commitment of financial resources, such as the so-called “financial responsibility” instruments. This term defines all the tools that require polluters to demonstrate ex ante sufficient financial resources to correct and compensate for environmental damage that may arise through their activities.

In its common application, financial responsibility implies that the operation of hazardous plants and other business is authorized only if companies can prove that future claims will be financially covered, for example, through letters of credit and surety bonds, cash accounts and certificates of deposit, self-insurance, and corporate guarantees.

Generally, financial responsibility may be complementary, sometimes mandatory, to the legislation on environmental accidents. In its different applications, it has a common motivation: to ensure the future internalization of the costs in order to indemnify the victims and discourage different forms of environmental deterioration.

On a law and economics point of view, financial responsibility can be defined as (potentially) efficient instruments to correct the asymmetric information issue. First of all, there is an incentive for the financial institutions to check that the companies are taking adequate preventive measures. Secondly, the companies are motivated to take precautions because financial responsibility guarantees that the expected costs of environmental risks appear on their balance sheet and business calculation (Feess and Hege 2000).

There are also alternative risk transfer products. A first kind of products is catastrophe bonds, consisting in securitizing some of the risk in bonds, which could be sold to high-yield investors. The cat bonds transfer risk to investors that receive coupons that are normally a reference rate plus an appropriate risk premium. By these products, financial institutions may limit risk exposure transferring natural catastrophe risk into the capital markets.

Weather derivatives are another kind of financial instrument used to hedge against the risk of

weather-related losses. Weather derivatives pay out on a specific trigger, e.g., temperature over a determined period rather than proof of loss. The investor providing a weather derivative charges the buyer a premium for access to capital, but if nothing happens, then the investor makes a profit.

With all this kind of insurance and financial products, it is possible to reach some efficiency goals. First of all, they give the possibility to stimulate ex ante preventive measure and to economically compensate ex post the victims. The second goal is the availability of extra capital for recovery that comes from financial markets. Finally, the accuracy and the resolution of hazard data and the likely impacts on climate change may improve with the involvement of financial market forecast ability.

The Mitigation and Adaptation Strategies

The challenge of reducing in the future the consequences of climate change is often framed in terms of two potential strategies: adaptation and mitigation. Mitigation involves lessening the magnitude of climate change itself; adaptation, by contrast, involves efforts to limit the vulnerability to climate change impacts through various measures, while not necessarily dealing with the underlying cause of those impacts.

“Mitigation” indicates any action taken to permanently eliminate or reduce the long-term risk and hazards of climate change to human life. A definition can be “An anthropogenic intervention to reduce the sources or enhance the sinks of greenhouse gases” (IPCC 2001).

“Adaptation” refers to the ability of a system to adjust to climate change to moderate potential damage, to take advantage of opportunities, or to cope with the consequences. A definition can be “Adjustment in natural or human systems to a new or changing environment” (IPCC 2001).

Mitchell and Tanner (2006) defined adaptation as an understanding of how individuals, groups, and natural systems can prepare for and respond to changes in climate. According to them, it is crucial to reduce the vulnerability to climate change.

While mitigation tackles the causes of climate change, adaptation tackles the effects of the phenomenon. The potential to adjust in order to minimize negative impact and maximize any benefits from changes in climate is known as adaptive capacity. A successful adaptation can reduce vulnerability by building on and strengthening existing coping strategies.

In general, the more mitigation there is, the less will be the impacts to which we will have to adjust and the less the risks for which we will have to prepare. Conversely, the greater the degree of preparatory adaptation, the less may be the impacts associated with any given degree of climate change.

The idea is that less mitigation means greater climate change effects, and consequently more adaptation is the basis for the urgency surrounding reductions in greenhouse gases. The two strategies are implemented on the same local or regional scale and may be motivated by local and regional priorities and interests, as well as global concerns. Mitigation has global benefits, although effective mitigation needs to involve a sufficient number of major GHG emitters to foreclose leakage. Adaptation typically works on the scale of an impacted system, which is regional at best, but mostly local, although some adaptation might result in spillovers across national boundaries.

Climate mitigation and adaptation should not be seen as alternatives to each other, as they are not discrete activities but rather a combined set of actions in an overall strategy to reduce GHG emissions. The challenge is to define an efficient mix of government policy interventions to provide the right incentives to invest in cost-effective preventive measures to reduce the final cost of disasters. The target is to tackle the consequences of climate change by mitigation, through the promotion of ways to reduce greenhouse gas emissions and make society to adapt to the impacts of climate change, by promoting the effective limitation and management of risks from extreme weather-related hazards.

On a law and economics perspective, generally, private contracting has been recognized as a

significant and potentially effective means of influencing private actors' behavior and even as a form of environmental policy instrument. So the financial and insurance products, that we have above analyzed, have significant potential to influence the behavior of individuals through its contracting contents, and this implies that the financial markets can play a role within the mitigation and adaptation policies.

For example, insurance companies may offer differential premiums to customers depending on the customers' level of protection from loss caused by weather-related disasters with an opportunity for insurers to reduce their own overall and maximum possible loss exposure while promoting communities' overall resilience in the face of climate change's impacts. Moreover, financial products can include arrangements intended to bring needed capital that will reduce the risk posed by future climate-related hazards to those who are most likely to be in peril.

Financial and insurance products could affect incentives for individuals to address climate change seeking mechanism to facilitate mitigation of GHG and adaptation to the inevitable impacts of climate change. Additionally, financial institutions are motivated to take significant actions aimed at mitigating overall societal greenhouse gas emissions and increasing adaptive capacity because these actions would reduce overall uncertainty and decrease people and business' potential exposure to catastrophic risks in excess of their capacity.

Conclusive Remarks on a Future Climate Change Liability System

The law and economics analysis of the climate change remedies has been focused on the question of which would be the policy instrument most suited to provide incentives to industry and other sources to reduce greenhouse gas emissions. And the literature is still giving attention mainly to the comparison of carbon taxes and emission trading scheme (Nordhaus 2006).

Not so much attention has been addressed to another instrument to provide incentives to polluters to reduce emissions, largely used to internalize other environmental externalities, the liability system. With “liability” we intend the possibility of applying national tort law to the damage caused by climate change and the possibility for holding states liable under international law if emissions originating from a country were to cause damage to the citizens of other nations.

Even if it seems that the application of a liability system to climate change is merely a theoretical issue, in reality more and more public authorities or individuals have tried to sue large emitters of GHG, and, in some cases, claims were directed against governmental authorities for failure to take measures to reduce emissions of greenhouse gases.

As an example, in 2002, the small island state Tuvalu threatened to take the United States and Australia to the International Court of Justice as a result of their failure to stabilize GHG emissions, thus causing the melting of ice caps which consequently leads to a rise in sea levels which threatened its territory. Although for a change in government the application was never made, this example demonstrates the way in which international law could be used to impose liability for climate change-related harm.

Beyond this specific case, most of these claims would probably not qualify as liability suits in the strict sense, since it is usually not compensation for damage suffered that is asked by the plaintiffs, but rather injunctive relief in order to obtain a reduction of greenhouse gasses. Moreover, most of the claims brought so far, mainly in the United States, were either not successful, were withdrawn, or have not yet led to a specific result.

On a law and economics point of view liability is not only an instrument “to obtain compensation for damages resulting from climate change (the more traditional liability setting) but equally are looking at the question to what extent civil liability and the courts in general may be useful to force potential polluters (or governmental authorities) to take measures to reduce (the effects of) climate change” (Faure and Peeters 2011, p. 10).

A liability system could also play a role in mitigating climate change, and a question is open to what extent it is useful to use the civil liability system to strive for a mitigation of greenhouse gas emissions in addition to the existing framework which largely relies on carbon tax and emission trading systems.

References

- Faure M, Peeters M (2011) Climate change liability. Edward Elgar, Cheltenham
- Feess E, Hege U (2000) Environmental harm, and financial responsibility. *Geneva Pap Risk Insur Issue Pract* 25:220–234
- IPCC (2001) Climate change 2001: synthesis report. In: RT Watson and the Core Team (eds) A contribution of working groups I, II, III to the third assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge/New York
- Kaplow L, Shavell S (2002) On the superiority of corrective taxes to quantity. *American Law and Economics Review*, 4, pp. 1–17
- Kaul I, Conceicao P, Le Goulven K, Mendoza RU (2003) How to improve the provision of global public goods. In: UNDP (ed) *Providing global goods – managing globalization*. Oxford University Press, New York
- Mitchell T, Tanner TM (2006) *Adapting to climate change: challenges and opportunities for the development community*. Tearfund, Middlesex
- Nordhaus DW (2006) After Kyoto: alternative mechanisms to control global warming. FPIF discussion paper
- Nordhaus DW (2007) To tax or not to tax: alternative approach to slowing global warming. *Rev Environ Econ Policy* 1:26–40
- Pizer W (1999) Choosing price or quantity controls for greenhouse gases, resources for the future. *Climate Issue Brief* no 17
- Porrini D (2011) The (potential) role of insurance sector in climate change economic policies. *Environ Econ* 2(1):15–24
- Posner R (2004) *Catastrophe*. Oxford University Press, New York
- Weitzman ML (1974) Prices vs. quantities. *Rev Econ Stud* 41(4):477–491

Clinical Trials

- [Human Experimentation](#)

Coase, Ronald

Herbert Hovenkamp

The College of Law, The University of Iowa, Iowa City, IA, USA

Abstract

Ronald Coase (1910–2013) was a British born and trained economist who moved to the United States in 1951. He spent most of his career at the University of Chicago. Coase's principal contributions addressed the fact that moving resources through the economy by means of transactions is costly – an idea that he introduced in *The Nature of the Firm* (1937) and developed further in *The Problem of Social Cost* (1960). Over his career Coase argued in numerous papers that if transaction costs are modest, private bargaining is often better than legislation or taxation as devices for settling resource conflicts. His work was highly influential in the development of movements away from regulation and back to more market-centric devices for managing the private economy. Coase won the Nobel Prize in economics in 1991.

Biography, Impact and Legacy

Ronald Harry Coase (1910–2013) was born in suburban London. His father worked for the British post office as a telegraph operator, as did his mother until marriage. Coase attended the University of London and then the London School of Economics, receiving a Bachelor of Commerce degree in 1932. After lectureships at Dundee and Liverpool, he returned to LSE from 1935 to 1951. He then moved to the United States to the State University of New York at Buffalo. In 1958 he went to the University of Virginia, and in 1964 to the University of Chicago. For a time he was editor of the *Journal of Law and Economics*.

Coase received the Nobel Prize in Economics in 1991. By his own admission, he did not like mathematics – a fact that set him apart from most of the economists of his generation.

Coase was hardly the most prodigious writer among Nobel laureate economists, but what he wrote was highly influential. Indeed, in a real sense he may be called the father of the discipline of law and economics. His reputation rests heavily on two articles. “The Nature of the Firm” (Coase 1937) was written during the period 1932–1934 while Coase was an assistant lecturer at the School of Economics and Commerce, Dundee, Scotland (Coase 1994). He wrote “The Problem of Social Cost” (Coase 1960) while he was at the University of Virginia.

Coase has stated that “The Nature of the Firm” was conceived in 1931 and essentially finished in 1934. He was in his early twenties and just beginning his academic career. His article was intended to address a different issue than the ones that eventually made it prominent. Marginalists since the great industrial economist Alfred Marshall at Cambridge were troubled about why a single industry contains firms of different sizes and structures. For example, if fixed costs were at all substantial, one might expect the market to be taken over by a single firm, which would have lower per unit costs than any rival. In the highly influential eighth edition of his *Principles of Economics* (1920), Marshall developed the idea of the “representative firm,” a hypothetical mature firm with “normal” cost characteristics, although individual firms in various stages of development could vary. Marshall never specified precisely what made a firm “representative,” and identifying it was like identifying the “representative” tree in a forest (Marshall 1890; 1920).

This analytically unsatisfactory theory led to criticism and attempts at refinement. One influential critique was American economist John Maurice Clark's book on fixed costs (Clark 1923), which addressed the problem in terms of diverse technologies and pricing strategies: scale economies do not produce a single firm because firms are not identical. The subsequent rise of theories of product differentiation made the idea of a “representative” firm obsolete in microeconomics,

although it retained more traction in macroeconomics, particularly in Keynes. Firms in differentiated markets can have quite different sizes and structures. They compete by appealing to divergent consumer tastes.

In 1928 Lionel Robbins, head of the London School of Economics, strongly criticized Marshall's concept of the representative firm for failing to apply the very marginalist rigor that Marshall himself advocated (Robbins 1928). Arthur Cecil Pigou, Marshall's successor at Cambridge, developed the idea of the "equilibrium firm," arguing that a firm will expand when its marginal cost is lower than the market's supply price but contract when it is higher (Pigou 1928). The equilibrium firm is one whose marginal cost just equals the market supply price. In 1931 Cambridge economist E.A.G. Robinson added in *The Structure of Competitive Industry* that "management costs" must also be considered in any question about firm size and structure (Robinson 1931; Hovenkamp 2011). So Coase was not writing on a clean slate.

This history explains Coase's strange paean to Marshall in the opening paragraphs of "The Nature of the Firm." Coase stated his intent to use "two of the most powerful instruments of economic analysis developed by Marshall, the idea of the margin and that of substitution, together giving the idea of substitution at the margin" (Coase 1937). By 1937 the ideas of marginalism and substitution to equilibrium had become conventional in economics. They were not worth mentioning, except that Coase was pointing out a gap in Marshall's approach. Coase then observed that marginal cost includes all relevant incremental costs, including what he termed "marketing costs," by which he meant "the costs of using the price mechanism." The term "transaction costs," for which Coase is now popularly associated, did not appear in this article.

Coase's highly elegant model argued that for every production or distribution decision, a firm compares alternative approaches, including purchase on the market as an alternative to internal production, by various means. Internal production, internal management, and use of external markets are all costly. The firm's management

selects the alternative that maximizes firm value. The aggregate of these decisions accounts fully for the firm's size and "shape" – that is, the variety of markets in which it operates and the extent of its vertical integration. The elegance of Coase's argument lay not only in its simplicity but also its enormous range, extending far beyond vertical integration itself to such questions as whether to differentiate one's product or use more or less centralized governance, equity or debt financing, and the like. In the process "The Nature of the Firm" developed a powerful theme that came to dominate Coase's work – namely, a property right and a contract are simply alternative ways of getting something done. For example, an automobile maker's decision whether to build its own spark plugs or purchase them is simply a choice between property and contract.

Over his career Coase repeatedly criticized "blackboard" economists who abstracted from reality, repeatedly calling for more empirical research (e.g., Coase 1992). But the empirical research that went into "The Nature of the Firm" is minimal. Coase visited a few firms, conducted a very few interviews, and overheard some phone calls about procurement. His theory was purely analytic in the British tradition, assuming how a rational actor would select among alternative production decisions.

"The Nature of the Firm" lay ignored for 30 years after its publication, with leading texts on industrial organization not even mentioning it (e.g., Bain 1959). In 1942 the prominent economist and public intellectual Kenneth E. Boulding wrote an article discussing the leading literature on the theory of the firm over the preceding 10 years, but did not cite Coase's article (Boulding 1942). It was finally rediscovered after "The Problem of Social Cost" was published in 1960 (Cheung 1983).

In 1959 Coase published an article arguing that an auction-style market would be a better way to allocate radio spectrum than the largely political arrangements currently in use: "...it is not clear why we should have to rely on the Federal Communications Commission rather than the ordinary pricing mechanism to decide whether a particular frequency should be used by the police, or for

a radiotelephone, or for a taxi service. . .” (Coase 1959). That article contained this insight that came out of “The Nature of the Firm” but became the basis of “The Problem of Social Cost” a year later. Speaking of a cave, Coase noted that the law of property determines who owns it. However,

... the law merely determines the person with whom it is necessary to make a contract to obtain the use of the cave. Whether the cave is used for storing bank records, as a natural gas reservoir, or for growing mushrooms depends, not on the law of property, but on whether the bank, the natural gas corporation, or the mushroom concern will pay the most in order to be able to use the cave. (Coase 1959, at 25)

The principal purpose of governmental spectrum allocation, Coase observed, was to prevent interference that occurred when spectrum assignments conflicted with one another. Coase introduced the case of *Sturges v. Bridgeman* (1879), a nuisance dispute involving a physician who shared a building with a confectioner. The thumping of the confectioner’s mechanical mortar and pestle interfered with the physician’s use of his stethoscope. Coase pointed out that neither *Sturges* nor the spectrum case involved conflicts between a wrongdoer and a victim. They merely represented inconsistent property interests. In a well-functioning market, the interest would go to the person who was willing to pay the most for it.

Coase elaborated on this theme a year later in “The Problem of Social Cost” (Coase 1960). His foil was no longer the FCC but rather Pigou, who had died the previous year. Pigou was the first neoclassical economist to write extensively about how the costs of moving resources should be factored into economic analysis, although his concept of “costs of movement” was more inclusive than Coase’s “transaction costs.” Pigou argued that in cases involving multiple, unorganized users of rivalrous resources, individuals would tend toward excessive use. In such cases the state should intervene with taxes or regulations designed to encourage efficient use. One example that Pigou gave and Coase discussed was the factory that belched smoke, injuring downwind landowners. Clean air was the resource in question. To the extent the factory did not bear the full social cost of dirty air it would

overpollute. Pigou argued that the factory should be given legal liability so as to reduce or eliminate the smoke, or else assessed a tax that was “equivalent in money terms to the damages it would cost” (Pigou 1932, Chapter 9). Coase argued that it was incorrect to think of the factory as the “wrongdoer” and the property owners as victims. Both performed useful social activities that were simply inconsistent uses of land. His second point was that without transaction costs, private bargaining would address the problem, not necessarily by shutting the factory down, but rather by assigning the right to whoever valued it most highly.

Coase did not invent the term “Coase Theorem.” That credit belongs to George J. Stigler, Coase’s colleague at Chicago. Stigler also provided this definition: “Under Perfect Competition Private and Social Costs will be Equal” (Stigler 1966, p. 113). The definition was probably intended to capture Coase’s differences with Pigou.

Stigler’s initial definition caught Coase’s insights very poorly. Coase’s paper had virtually nothing to do with perfect competition. The markets in “The Problem of Social Cost” are largely bilateral monopolies, and Coase readily acknowledged that the price at which legal entitlements in such markets are transferred is indeterminate. Under perfect competition prices are at marginal cost. Finally Stigler’s definition trivializes the Coase Theorem by turning it into a minor and fairly obvious corollary of the First Welfare Theorem, which was already well known when “The Problem of Social Cost” was published (Blaug 2007). Coase corrected Stigler’s statement to say “with zero transaction costs private and social costs will be equal” (Coase 1988b, p. 158). Stigler later revised his definition to state “when there are no transaction costs the assignments of legal rights have no effect upon the allocation of resources among economic enterprises. . .” (Stigler 2003, at 77).

As formalized, the Coase Theorem is said to have two parts, or perhaps two different applications, which are not mutually exclusive. First, an “efficiency” thesis states that if transaction costs are zero, then the initial allocation of a right is irrelevant to efficiency because the right will be

traded to its highest value user. The final allocation maximizes private value among the bargainers. It also maximizes social value, *provided* that no outsider to the bargain is adversely affected – that is, there are no negative externalities. Second, an “invariance” thesis states that if transaction costs are zero, then where the right ends up is invariant to the underlying legal rule that creates it – “irrespective of the initial assignment of rights” (Coase 1992). One limitation is that the right must be “alienable,” meaning that the parties can contract around it through settlement. For example, in *Sturges* it does not matter whether Bridgman’s mortar and pestle is or is not declared a nuisance. Whether the machine is shut down depends entirely on whether Sturges values the right to be free of the noise more than Bridgeman values the right to use the machine. One problem with “inalienable” legislated rights is that the parties cannot bargain around them. For example, if a zoning law prohibited the use of Bridgeman’s machine, then the parties could not bargain to the efficient solution if use of the machine was more valuable than the interference it caused. At least for biological actors, the endowment effect can undermine the invariance thesis if an actor’s willingness to accept for a particular right is greater than his willingness to pay (Hovenkamp 1990; Kahneman et al. 1990).

Writing about the Coase Theorem has been voluminous, making “The Problem of Social Cost” the most cited law review article of all time. It drew an almost immediate response in tort and property law, two areas where the infant law and economics movement cut its teeth. In 1964 University of Chicago law professors Walter Blum and Harry Kalven acknowledged its importance in an article on tort liability. They observed, however, that the actors in Coase’s account were neighbors well aware of the accident possibilities before they occurred. The theorem would not work for automobile accidents, however, because prior to the accident the parties would not be in a position to negotiate over such issues as right of way (Blum and Kalven 1964; see Medema 2013).

By contrast, Guido Calabresi developed an alternative that relied on objective criteria for determining who would have won a bargain had

the parties been able to negotiate. In such cases liability should be assigned to the “least cost avoider” (Calabresi 1970). From there the debate spread into numerous areas, including questions such as when strict liability was a more efficient tort rule than negligence. In property law the literature considered whether common law rules such as nuisance or else private restrictive covenants were effective alternatives to zoning (e.g., Ellickson 1973).

Another issue was the choice between “alienable” rules that could be privately bargained and “inalienable” rules that could not be (Calabresi and Melamed 1972). Generally speaking, private injunction rules with alienable entitlements set up a mechanism like the one Coase contemplated. The rule creates a property interest in a plaintiff, if entitled to the injunction, but permits the parties to bargain around it. By contrast, a pure damages rule permits the conduct to continue but may require one person to pay the other an amount determined by the court. Once again, however, the parties are free to negotiate their own private arrangement. An “inalienability” rule, by contrast, assigns the right in one way and prohibits the parties from changing it by private agreement.

Generally speaking, inalienability rules make the most sense when transaction costs are high, meaning that the parties are unlikely to reach the efficient bargain, *provided* the state can by objective means determine which party would have won the right in a free bargain. For example, the common law rule requiring cars to yield to trains at grade crossings ordinarily creates an inalienability rule. When both are speeding toward the intersection, the car and the train are not in a good position to bargain over the right of way. Further, the costs of stopping and restarting are much higher for the train than for the car. So the state assigns the right of way to the train.

The choice between injunction rules and damages rules is more problematic. One view is that injunction rules are preferred when transaction costs are low and the parties are likely to bargain to an efficient result. By contrast, damages rules are superior when value determination is complex, perhaps because multiple parties are involved or there might be holdouts. One critique

of this view is that it implicitly assumes that the court is a better decision maker than the parties themselves (Polinsky 1980; Krier and Schwab 1995). The common law, it should be noted, tends to prefer injunctions more as damages are more difficult for an external observer to calculate. For example, breach of an agreement to sell land, thought to be unique, is usually remedied by specific performance. However, breach of an agreement to sell a commodity is generally remedied by expectancy damages. In 1972 then professor Richard A. Posner analyzed these and many other questions in his regularly updated book *Economic Analysis of Law* (Posner 1972), which was explicitly indebted to Coase and in a real sense institutionalized law and economics in legal analysis.

A “Coasean market” is one in which all affected parties must agree before a particular transaction can occur. In a traditional neoclassical market, by contrast, there might be thousands of buyers and sellers but only one of each is necessary to a deal. For example, when one buyer purchases bread from one seller, the rest of the market does not participate and is largely indifferent. The difference among markets is readily apparent in a case such as *Sturges vs. Bridgeman*. While Victorian London contained thousands of physicians, confectioners, and suitable houses, the “market” that Coase discusses involved a single seller, a single buyer, and a single duplex house. In the long run either *Sturges* or *Bridgeman* could avoid the conflict by moving way, thus indicating that Coase’s focus is not merely on a very tiny market but also on the short run. This small grouping is a market to the extent that the costs of exiting exceed the costs of reaching a bargain and staying. One important impact of the Coase Theorem was to increase economists’ focus on very small markets, such as the two parties to a tort dispute, a few homeowners in a subdivision, a husband and wife, or the relation between shareholders, creditors, and managers in a single firm.

The requirement that all parties in a Coasean market must agree poses difficulties as the number of parties increases or their interests are more diverse. For example, a smokestack factory might willingly compensate 100 downwind

landowners in order to keep running. But each one may be entitled to an injunction (abatement of the nuisance), so all must agree about how to share the award. More adjacent landowners, larger landowners, or those with more valuable homes will seek a larger share, and until these issues are resolved, there will be no agreement. The result could be endless cycles of coalitions and counter-coalitions. That this is a consequence of high transaction costs is by no means clear. A rational participant bargains as long as the cost of a further offer is less than the expected payoff. So if bargaining were indeed costless but there was any uncertainty about outcomes, bargaining would not stop. In these situations positive transaction costs force the agreement by making continuing bargaining more costly at the margin than any expected payoff. The transaction/bargain cost curve thus has a lopsided “U” shape, with endless bargaining when transaction costs are near zero, more successful bargaining when they are a little higher, and less successful as they rise to yet higher levels.

When multiparty Coasean bargains do occur, they can result in excessive stability. Exiting from them can be just as difficult as entering them in the first place. For example, zoning laws can usually be changed by the majority vote of a legislative body as economic conditions change. By contrast, contractual servitudes generally require the unanimous consent of all affected parties, producing significant holdup problems when the majority believes or market values indicate that a servitude has become counterproductive (Hovenkamp 2002).

Coase acknowledged many of these difficulties in his 1959 article on the Federal Communications Commission, although he paid little attention to them later and much of the transaction cost literature has ignored them. Coase noted that “when large numbers of people are involved, the argument for the institution of property rights is weakened and that for general regulations becomes stronger.” Speaking of smoke pollution, he acknowledged that “if many people are harmed and there are several sources of pollution, it is more difficult to reach a satisfactory solution through the market.” As a result, “in these circumstances it may be preferable to impose special regulations...” (Coase 1959, at 27, 29).

These admissions invite the question whether Coase really attacked Pigou fairly. The argument for Pigouvian taxes was not concerned about conflicting rights as between two bargainers where no one else was affected. Rather, it was with problems such as highway congestion or pollution, which affect many users, both spatially and often temporally. As a result Pigouvian taxes, such as a carbon tax, continue to have support among mainstream economists (e.g., Mankiw 2012, pp. 207–210; Medema 2011; Baumol 1972). The cost of fossil fuels includes not only production and distribution costs but also longer-run environmental costs. The affected interests include hundreds of millions of people and even future generations.

In the 1960s and 1970s, “The Nature of the Firm” was rediscovered. Together with “The Problem of Social Cost,” they became cornerstones in the development of “New Institutional Economics” (NIE) and its variations, sometimes called “organizational economics” or “transaction cost economics.” The earlier work of Oliver E. Williamson, particularly *Markets and Hierarchies*, probably did more than anything to bring “The Nature of the Firm” into the spotlight (Williamson 1975). NIE refocused economic study on “institutions,” an idea developed by the first generation of institutionalist economists a half century earlier, including Thorstein Veblen, Richard T. Ely, and John Commons. But NIE was dramatically different from the generally nontechnical, evolutionary, and often anti-marginalist conceptions of the original institutionalists (Hovenkamp 2013). The general thrust of NIE was to move economics away from large traditional markets to the study of very small ones, even viewing relationships inside the organization as a market. While the Coasean literature as applied to separate economic actors spoke of “transaction costs” as interfering with the efficient allocation of resources, the concept of “agency costs” came to describe costs internal to the firm that might obstruct efficient value maximization (e.g., Jensen and Meckling 1976). Another result was more refined studies of the risk and cost profiles that firms faced in deciding whether and how to integrate, including the significant costs of making costly, specialized

commitments to one’s trading partner (e.g., Klein et al. 1978; Coase 2000).

Coase made several other contributions to economics, including law and economics. One was his 1946 article “The Marginal Cost Controversy.” In 1938 Harold Hotelling had argued that, because marginal cost pricing is essential to competition, outcomes in industries with very high fixed costs, such as railroads and electric utilities, would be suboptimal (Hotelling 1938). Prices would be driven to marginal cost, with insufficient surplus to cover fixed costs. The correction was government subsidies permitting such firms to recoup their fixed cost investment. Coase’s rejoinder was to develop the concept of two-part pricing, with an entry or access fee to cover the fixed cost component and a per use variable fee to cover the marginal component (Coase 1946). Most of the subsequent literature on Coase’s argument concludes two things. *First*, two-part pricing will rarely yield optimal outcomes when competitive providers set their own prices, although they are often more efficient than purely linear pricing. *Second*, however, two-part tariffs can be (and are) an effective way to encourage closer-to-optimal output in price-regulated markets by bringing the per use price closer to the marginal cost (Tirole 1988, pp. 142–146; Brown et al. 1992).

Two of Coase’s important contributions in the early 1970s concerned the diverse topics of the durable goods monopolist and the scope of public goods. An article on durability and monopoly developed the “Coase conjecture” that in the very act of selling, the monopolist of a durable good dissipates its monopoly power (Coase 1972). It ends up competing with its own previous output, resold on secondary markets. Durability varies considerably, from nearly perfect in the case of land (Coase’s opening example) to highly imperfect in the case of clothing (a used suit at Goodwill is a poor competitor for a new suit at Macys). The value of such a monopolist’s output, Coase argued, depends on its ability to make a credible commitment to limit future output. For example, while van Gogh’s painting *The Starry Night* (1889) is priceless, people can obtain a copy for \$5 because the painting is in the public domain and no one can make a credible commitment that future output will be limited. In

addition, the durable goods monopolist may be able to profit by leasing rather than selling.

In “The Lighthouse in Economics” (Coase 1974), Coase wondered about the extent to which traditionally defined public goods really constitute a market failure. The lighthouse had appeared frequently in the economics literature as a public good that needed to be supplied by the government. However, Coase observed, privately owned lighthouses existed and were typically supported by a harbor tax or equivalent assessment against the vessels that benefitted from them. The real problem lays in developing an appropriate pricing mechanism and accounting for free riders – in particular, ships that might pass without actually using the harbor, thus benefitting from the lighthouse without paying the tax. Later critics observed, however, that private lighthouses either did not exist at all or else were short-lived relatively unsuccessful ventures (Bertrand 2006; Barnett and Block 2007). The harbor tax, if assessed by a public authority, was Pigouvian in any event.

Coase revisited many of the themes that defined his career in his Nobel Prize Lecture in 1991, entitled “The Institutional Structure of Production” (Coase 1992). He reiterated the theme that transaction costs are what give the legal system its importance and called for further empirical study of the role of transaction costs in real-world economies. He also lamented that his theories had been much less influential among economists than among lawyers – a view that was largely undermined by Oliver Williamson’s receipt of the Noble Prize in 2009. Coase’s work remains as alive and controversial as ever and has cast a long shadow on the disciplines of both economics and law.

References

- Bain JS (1959) *Industrial organization*. Wiley, New York
- Barnett W, Block W (2007) Coase and Van Zandt on lighthouses. *Public Financ Rev* 35:710–733
- Baumol WJ (1972) On taxation and the control of externalities. *Am Econ Rev* 62:307–322
- Bertrand E (2006) The Coasean analysis of lighthouse financing: myths and realities. *Camb J Econ* 30:389–402
- Blaug M (2007) The fundamental theorems of modern welfare economics, historically contemplated. *Hist Polit Econ* 39(2):185–207
- Blum WJ, Kalven H (1964) Public law perspectives on a private law problem—auto compensation plans. *U Chi L Rev* 31:641–723
- Boulding KE (1942) The theory of the firm in the last ten years. *Am Econ Rev* 32:791–802
- Brown DJ, Heller WP, Starr RM (1992) Two-part marginal cost pricing equilibria: existence and efficiency. *J Econ Theory* 57:52–72
- Calabresi G (1970) *The cost of accidents: a legal and economic analysis*. Yale University Press, New Haven
- Calabresi G, Melamed A (1972) Property rules, liability rules, and inalienability: one view of the cathedral. *Harvard Law Rev* 85:1089–1128
- Cheung SNS (1983) The contractual nature of the firm. *J Law Econ* 26:1–21
- Clark JM (1923) *Studies in the economics of overhead costs*. University of Chicago Press, Chicago
- Coase RH (1937) The nature of the firm. *Economica* 4(n.s.):386–405
- Coase RH (1946) The marginal cost controversy. *Economica* 13:169–182
- Coase RH (1959) The federal communications commission. *J Law Econ* 2:1–40
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Coase RH (1972) Durability and monopoly. *J Law Econ* 15:143–149
- Coase RH (1974) The lighthouse in economics. *J Law Econ* 17:357–376
- Coase RH (1988a) The nature of the firm: origin. *J Law Econ Organ* 4:3–17
- Coase RH (1988b) *The firm, the market and the law*. University of Chicago Press, Chicago
- Coase RH (1992) The institutional structure of production. *Am Econ Rev* 82:713–719
- Coase RH (1994) Duncan Black. In: Ronald H (ed) *Coase, essays on economics and economists*. University of Chicago Press, Chicago, pp 187–189
- Coase RH (2000) The acquisition of fisher body by General Motors. *J Law Econ* 43:15–32
- Ellickson RC (1973) Alternatives to zoning: covenants, nuisance rules, and fines as land use controls. *U Chi L Rev* 40:681–781
- Hotelling H (1938) The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6:242–269
- Hovenkamp H (1990) Marginal utility and the Coase Theorem. *Cornell L Rev* 75:783–801
- Hovenkamp H (2002) Bargaining in Coasean markets: servitudes and alternative land use controls. *J Corp L* 27:519–530
- Hovenkamp H (2011) Coase, institutionalism, and the origins of law and economics. *Ind L J* 86:499–542
- Hovenkamp H (2013) *The opening of American law: neo-classical legal thought, 1870–1970*. Oxford University Press, New York
- Jensen MC, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. *J Fin Econ* 3:305–360

- Kahneman D et al (1990) Experimental tests of the endowment effect and the Coase Theorem. *J Polit Econ* 98:1325–1346
- Klein B et al (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J Law Econ* 21:297–326
- Krier JE, Schwab SJ (1995) Property rules and liability rules: the cathedral in another light. *New York U Law Rev* 70:440–483
- Mankiw NG (2012) *Principles of economics*, 6th edn. Cengage Learning, Independence, Ky
- Marshall A (1890; 8th ed. 1920), *Principles of economics*. Macmillan, London
- Medema SG (2011) Of Coase and carbon: The Coase theorem in environmental economics, 1960–1979 (SSRN working paper, Dec. 20, 2011), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1929086
- Medema SG (2013) Rethinking market failure: ‘The problem of social cost’ before the ‘Coase Theorem’ (SSRN working paper, Jan. 25, 2013), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2188728
- Pigou AC (1928) An analysis of supply. *Econ J* 38:238–257
- Pigou AC (1932) *The economics of welfare*, 4th edn. Macmillan, London
- Polinsky MA (1980) Resolving nuisance disputes: the simple economics of injunctive and damage remedies. *Stanford Law Rev* 32:1075–1112
- Posner RA (1972, 8th ed. 2010) *Economic analysis of law* little. Little Brown/Aspen, Boston/New York
- Robbins L (1928) The representative firm. *Econ J* 38:387–404
- Robinson EAG (1931) *The structure of competitive industry*. Nisbet, London
- Stigler GJ (1966) *The theory of price*, 3rd edn. Macmillan, New York
- Stigler GJ (2003) *Memoirs of an unregulated economist*. University of Chicago Press, Chicago
- Sturges v. Bridgeman (1879) 11 Ch. D 852
- Tirole J (1988) *The theory of industrial organization*. MIT Press, Cambridge
- Williamson OE (1975) *Markets and hierarchies: analysis and antitrust implications*. Free Press, New York

property rights and that the operation of the price system requires these rights to be defined. Coase was more interested in how property rights are (or should be) allocated and exchanged than in their content or definition. He insisted that factors of production must be considered as property rights, but conversely, property rights, even when they relate to nuisances, are nothing more than extra production costs.

Introduction

Ronald H. Coase was concerned with law since he studied at the London School of Economics. He was there greatly influenced by his professor then colleague Arnold Plant, who wrote extensively on the importance of the delimitation of property rights and on the influence of their structure on the economic system (see Coase 1977). “The problem of social cost” (Coase 1960) was addressing economists, and one of its main insights was to explain “that what are traded on the market are not, as is often supposed by economists, physical entities, but the rights to perform certain actions, and the rights which individuals possess are established by the legal system” (Coase 1992, p. 717). In fact, Coase developed this idea while writing on the Federal Communications Commission (Coase 1959).

The Role of Property Rights

In 1959, Coase famously argued that the price system could be used to allocate radio frequencies. His strategy was to demonstrate that frequencies are unspecific compared to other goods and services. First, just like a land must be possessed to avoid multiple uses of it, property rights on frequencies must be defined to avoid simultaneous uses of the same frequencies: “A private-enterprise system cannot function properly unless property rights are created in resources, and, when this is done, someone wishing to use a resource has to pay the owner to obtain it. Chaos disappears; and so does the government except that a

Coase and Property Rights

Elodie Bertrand

CNRS and University Paris 1 Panthéon-Sorbonne (ISJPS, UMR 8103), Paris, France

Abstract

In “The problem of social cost” (1960), Ronald H. Coase argued that what are exchanged are

legal system to define property rights and to arbitrate disputes is, of course, necessary” (Coase 1959, p. 14). Coase was thus stressing that what are exchanged are property rights (like John Roger Commons before him) and that the operation of the price system requires these rights to be defined.

Second, Coase added that interferences between adjacent frequencies did not make property rights on frequencies specific either. Based on the *Sturges v. Bridgman* case (1879), he argues that either the confectioner has the right to make noise and imposes costs on his doctor-neighbor, who cannot longer practice, or the doctor has the right to practice in silence and he imposes costs on his neighbor. This is exactly the same with the owner of a land who impedes others to use it: “What this example shows is that there is no analytical difference between the right to use a resource without direct harm to others and the right to conduct operations in such a way as to produce direct harm to others. In each case something is denied to others: in one case, use of a resource; in the other, use of a mode of operation” (ibid.: 26). Likewise, this is the right to use frequency in a certain way that is exchanged, not the frequency itself: “What does not seem to have been understood is that what is being allocated by the Federal Communications Commission, or, if there were a market, what would be sold, is the right to use a piece of equipment to transmit signals in a particular way. Once the question is looked at in this way, it is unnecessary to think in terms of ownership of frequencies or the ether” (ibid.: 33).

This “change of approach” (Coase 1960, p. 42) is detailed in “The problem of social cost”. Factors of production (including those that create external effects) must be thought of as property rights (and conversely): “A final reason for the failure to develop a theory adequate to handle the problem of harmful effects stems from a faulty concept of a factor of production. This is usually thought of as a physical entity which the businessman acquires and uses (an acre of land, a ton of fertiliser) instead of as a right to perform certain (physical) actions. . . . If factors of production are thought of as rights, it becomes easier to understand that the right to do

something which has a harmful effect (such as the creation of smoke, noise, smells, etc.) is also a factor of production. . . . The cost of exercising a right (of using a factor of production) is always the loss which is suffered elsewhere in consequence of the exercise of that right” (ibid.: 43–44). Nuisances are defined as reciprocal conflicts over the use of a property right. Like for other factors of production, the cost of this right is an opportunity cost, and the price system makes the exchanges efficient, if it operates without cost.

The role of the judge would then just be to define property rights, no matter how, but in a definite and predictable way (ibid.: 19). Exchanges on these property rights (including those whose use implies effects on others) could then take place and yield an optimal result, independent from their initial allocation: this is the idea Stigler named “the Coase theorem”.

However, transaction costs may prevent some exchanges of rights, and, when this is the case, the initial allocation of rights is not modified or, at least, not until the optimal allocation is reached. In these conditions, the initial distribution of property rights influences the economic result: “with positive transaction costs, the law plays a crucial role in determining how resources are used” (Coase 1988, p. 178). In this case, what should be done?

Either the initial delimitation of rights is given but inefficient, and the economist or the policy-maker must compare the values of production yielded by different institutional arrangements (market, firm, regulation, status quo) and choose the one in which it is the highest, taking into account the costs of operation of these arrangements and the costs of changing from one to another (Coase 1960, pp. 16–18).

Either property rights are not yet allocated and common law judges should take into account this economic influence of their decisions when allocating property rights: “It would therefore seem desirable that the courts should understand the economic consequences of their decisions and should, insofar as this is possible without creating too much uncertainty about the legal position itself, take these consequences into account when making their decisions. Even when it is

possible to change the legal delimitation of rights through market transactions, it is obviously desirable to reduce the need for such transactions and thus reduce the employment of resources in carrying them out" (ibid.: 19). This entails distributing, right from the start, the property right to the person who values it the most, that is to say, imitating the result of the market. If exchanges cannot take place, this could improve efficiency. Even when transaction costs do not prevent exchanges of the right, limiting the need for exchanges economizes on these costs. And Coase couples this normative role of judges to the empirical claim that they actually are, at least partly, aware of the reciprocity of the problem and of the economic consequences of their decisions: they introduce economic efficiency considerations in their deliberations, as several cases analyzed in "The problem of social cost" would suggest (see Bertrand 2015). This was the very beginning of the debate over the efficiency of the common law, more famously brought to the fore by Posner (1972).

Coase, however, also mentions that the allocation of property rights by statutory law (by contrast to common law) is generally inefficient, because it protects harmful producers – gives them the right to pollute – beyond what would be economically desirable (Coase 1960, pp. 26–27).

Regarding the analysis of property rights in law and economics, Coase introduced three important ideas. First, property rights may be analyzed with the theory of markets and contracts; this kind of work was developed by Armen Alchian, Harold Demsetz, and Douglass North. Second, externalities are the symptom of ill-defined property rights, as was substantiated by Demsetz (1967). Third, and this intuition was expanded by Douglas Allen (1991) and Yoram Barzel (1989), transaction costs are the costs of establishing and maintaining property rights.

What Are Coasean Property Rights?

Coase was more interested in how property rights are (or should be) allocated and exchanged than in the content or definition of these property rights.

"The problem of social cost" uses the term "property right", but in the examples on which the analysis is based, agents are "liable" or not, and the reciprocal situations of liability are not exactly symmetrical.

Calabresi and Melamed's (1972) typology helps to reinterpret Coase's argument. When the rancher is "liable" (Coase's words), this means that he has to pay the farmer for the damage caused to his corn. The "entitlement" is here protected by a "liability rule" (Calabresi and Melamed's terms): the agreement of the farmer is not necessary for the exchange to take place, and the price is externally determined. In the reciprocal case, when the rancher is not "liable", the farmer has to negotiate with him so that he diminishes his herd; this means that this time the entitlement is protected by a "property rule": the rancher's agreement is necessary, and the price is determined during the process of negotiation.

Merrill and Smith (2001, 2011) explain that Coase thinks property rights as bundles of use rights and that he is the one who transmitted this legal realist view of property to (law and) economics. This view is very different from the traditional conception of property, attached to "things" and excluding the world from this "thing" (in rem property), that we find in William Blackstone and Adam Smith. Coase is rather referring to a bundle of in personam rights, attached to persons and obtained against certain other persons. This is most visible here: "We may speak of a person owning land and using it as a factor of production but what the land-owner in fact possesses is the right to carry out a circumscribed list of actions" (Coase 1960, p. 44). Coase chose this conception because he was confronted to radio frequencies which are not "things" (Coase 1988, p. 11): he was dealing with harmful effects and was concerned with exchanges of rights; but this choice obscured the in rem character of property rights.

Conclusions

Coase's view of property rights as bundle of rights comes from his will to conceive harmful effects as

unspecific by thinking of them as just another factor of production. As soon as 1959, Coase asserts that, in order to practice, the doctor must own not only his examination room but also the right to use it in silence; likewise the confectioner must own not only his machinery but also the right to use it noisily. The right to harm or to be protected from harms complements the classical factors of production: it is a right to use a certain resource in a certain manner, and it is a cost, to be taken into account among the other costs of production. This idea permitted considering the right to harm or to be protected from harms as a factor of production. But this is also how, while Coase wanted to introduce property rights in economics, he assimilated them to unspecific factors of production, allowing them to enter into the analysis just as other costs of production. In Coase's examples, property rights are just costs. He insisted that factors of production must be considered as property rights, but conversely, property rights, even when they relate to nuisances, are nothing else than extra production costs. We therefore understand how this vision of property rights disregards any notion of causality, responsibility, enmity, or moral inalienability and more largely any historical, social, or moral dimension.

Cross-References

- [Coase, Ronald](#)
- [Coase Theorem](#)
- [Externalities](#)
- [Transaction Costs](#)

References

- Allen DW (1991) What are transaction costs. *Res Law Econ* 14:1–18
- Barzel Y (1989) *Economic analysis of property rights*. Cambridge University Press, Cambridge
- Bertrand E (2015) 'The fugitive': the figure of the judge in Coase's economics. *J Inst Econ*. 11(2):413–435
- Calabresi G, Melamed AD (1972) Property rules, liability rules, and inalienability: one view of the cathedral. *Harv Law Rev* 85:1089–1128
- Coase RH (1959) The federal communications commission. *J Law Econ* 2:1–40

- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Coase RH (1977) Review of: selected economic essays and addresses by Arnold Plant. *J Econ Lit* 15:86–88
- Coase RH (1988) *The firm, the market and the law*. The University of Chicago Press, Chicago
- Coase RH (1992) The institutional structure of production. *Am Econ Rev* 82:713–719
- Demsetz H (1967) Toward a theory of property rights. *Am Econ Rev* 57:347–359
- Merrill TW, Smith HE (2001) What happened to property in law and economics? *Yale Law J* 111:357–398
- Merrill TW, Smith HE (2011) Making Coasean property more Coasean. *J Law Econ* 54:S77–S104
- Posner RA (1972) *Economic analysis of law*. Little, Brown, Boston

Coase Theorem

Steven G. Medema
Department of Economics, University of
Colorado Denver, Denver, CO, USA

Definition

Assuming the property rights are well defined and that the costs of transacting are zero, parties to an externality will resolve the dispute efficiently, and the outcome will be unaffected by to which party rights are initially assigned.

Introduction

The Coase theorem was derived from the negotiation result laid out by Ronald Coase in his 1960 article, "The Problem of Social Cost" (1960), after having first been articulated in his discussion of the allocation of broadcast frequencies a year earlier (Coase 1959). The theorem, so named by George Stigler (1966, p. 113), has been stated in a variety of ways by the thousands of authors who have invoked it over the last five decades, but the essentials are as follows: Assuming the property rights are well defined and that the costs of transacting are zero, parties to an externality will resolve the dispute efficiently, and the outcome will be unaffected by to which party rights are

initially assigned. In short, rights matter, but to whom they are assigned initially does not. The theorem thus suggests that, under assumed conditions, the assignment of rights related to pollution externalities or of liability for accidents resulting from the use of (perhaps defective) consumer products, the remedy mandated for breach of contract, the rules governing trespass, etc. are largely irrelevant. It matters crucially that there are legal rules to govern these phenomena, but to whom the relevant rights are assigned does not. An efficient and invariant allocation is guaranteed regardless.

The Coase theorem is a cornerstone of the economic theory of externalities and of the economic analysis of law; yet, it remains the subject of controversy. It has been the subject of challenges and defenses both intuitive and mathematical, experimented with and subjected to empirical assessment. (Medema and Zerbe (2000) provide a survey that includes a plethora of references to the literature discussed in the present essay.) There are many who believe that the theorem is not valid as a proposition in economic logic, but the theorem is nonetheless referenced regularly as an accepted truth in the scholarly and textbook literatures and has been applied in various ways across virtually every subfield of economics and of law, to the political realm, and beyond. Where their *is* widespread agreement is on the theorem's domain of *direct* relevance – that it is highly circumscribed and likely zero. At issue here is the theorem's assumption of zero transaction costs, a frictionless world that, one could argue, is more highly restrictive than the world of perfect competition – the economist's ideal type.

Underpinnings

The zero transaction costs assumption remains imprecisely defined even to this day owing to the ambiguity surrounding the notion of transaction costs. Meanings attributed to the zero transaction costs assumption range from the minimalist idea that there are no costs associated with engaging in negotiations per se (in essence, that talk is free) to the vastly more expansive idea that all relevant information can be acquired costlessly and thus that mutual gains can be realized instantaneously (Allen 1990). One gets a sense at times that, in the

hands of those supporting the theorem, the working definition of transaction costs is “all impediments to the Coase theorem's validity.” While this can give the Coase theorem a tautological air, the fact is that the frictionless world of the theorem has its counterparts in physics and elsewhere, and its unrealistic nature does not leave it without significance as a framework for thought experiments. More troubling on this score is that the Coase theorem assumes a free lunch that, at a minimum, there are no opportunity costs of time; the only question is how vast this free lunch domain is assumed to be.

Many of the critiques that are said to invalidate the Coase theorem involve violations of the zero transaction costs assumption, or at least can be objected to on those grounds. This is particularly true of the critiques grounded in game theory, the strategic behavior accompanying which is only possible because of some form of imperfect/incomplete information. Thus, while there is no question – as has been demonstrated time and again – that game theoretic analysis reveals very clearly that there are any number of settings in which the Coase theorem's efficiency and invariance predictions do not hold up, there is substantial question as to whether the conclusions drawn actually go to the validity of the Coase theorem itself.

The second fundamental assumption underlying the theorem is that property rights – that is, who has the right to do what – are fully specified over the relevant resources. These rights make clear the baseline against which any negotiations will take place and precisely what it is that is being exchanged between parties through any negotiation process. Absent such well-defined rights, policing and enforcement costs may be sufficiently high to preclude negotiations entirely or, at the very least, inhibit the exploitation of all relevant gains from exchange (Demsetz 1964). Some have argued that the property rights assumption is subsumed under or simply an extension of the assumption of zero transaction costs (or at least the most broad version of it), as, within such a world, there will be no uncertainties about the status of rights and no costs of policing and enforcement. Of course, these rights must be

alienable, for otherwise the associated costs of transaction are effectively infinite.

The Consumer Problem

When Coase laid out his negotiation analysis in “The Problem of Social Cost,” both his analysis and the illustrations that he employed dealt with externality relationships among business firms. While, as any number of subsequent commentators pointed out, this presents its own set of potential problems – e.g., entry-related long-run asymmetries, non-separable cost functions, non-convexities in production sets, and the like – these objections can be overcome with relative ease (though validity here, like beauty, is sometimes in the eye of the beholder). Where the Coase theorem has run into difficulty is when consumers are party to the externality, whether as victims of, say, pollution emitted by a factory or when the externality is among to individual agents, such as when a neighbor plays his/her stereo too loud.

The challenges for the theorem pointed to here take two forms, one of which has been present in the literature since the mid-1960s and the other of which has come to the fore more recently via scholarship in the area of behavioral economics. The first of these is the problem of income effects, that the initial assignment of rights raises the income/wealth of one party relative to the other (Mishan 1971). The different patterns of resulting demands can both give rise to varying negotiated solutions to the externality issue and, in a large numbers situation, varying levels of prices and outputs in the marketplace. Thus, while the negotiated solution may be efficient (in the sense that all gains from trade have been exhausted), the outcome is not invariant – or even, strictly speaking, comparable – across alternative specifications of rights. One result of this challenge has been a propensity to add an “income effects aside” qualification to statements of the theorem or statements of the theorem that include the efficiency result but not the invariance claim – though these are far from universal in the literature.

The second issue on the consumer side arises when the initial distribution of rights impacts the valuation that agents place on the rights in question (Kahneman et al. 1990). The problem arises

because of the potential that a given right may be valued differently by *A* when he/she owns that right than when that same right is owned by *B*. The potential for a divergence between an individual’s willingness to pay for a right (WTP) and the amount which he/she is willing to accept in payment (WTA) to give up that right holds out the prospect that outcomes of the exchange process will vary if $WTA \neq WTP$ and, in fact, that such divergences may preclude exchange altogether if rights are assigned to one party rather than the other. The fact that a number of experimental treatments of the Coase theorem and of other economic contexts have provided evidence for the reality of such divergences has further called into question the validity of the theorem’s invariance result when consumers are party to the dispute in question.

Why Does This Matter?

If the Coase theorem’s validity depends on the existence of a fictional world many steps removed from reality and even then might not hold water, why has it been the subject of so much controversy and, even with that, come to occupy such a prominent place in legal-economic analysis?

First, the received theory of externalities at the time when Coase formulated his negotiation result implicitly assumed a similarly fictional world. In fact, even today the textbook analysis of externalities and of Pigovian remedies for them carries on that framework. The idea that Pigovian instruments – taxes, subsidies, and regulations – can be used to achieve efficient resolutions of externalities assumes that there are no costs, information, or otherwise associated with coordination. But the Coase theorem shows that, if coordination is costless, exchange, too, can generate efficient solutions to externality issues. That is, the claim that Pigovian instruments are *necessary* for the efficient resolution of externalities is shown to be invalid. Moreover, if one is going to invoke the reality of transaction costs against the Coase theorem, consistency requires taking into account the costs associated with governmental coordination. The implication of this is that there is no globally optimal means of dealing with externalities; the appropriate response (which may be

allowing the status quo to persist) can only be determined on a case-by-case basis.

Second, the Coase theorem shows that people will, of their own volition, efficiently and invariantly resolve externality issues if they are free to do so – that is, if transaction costs do not get in the way. This result, then, is the one that would be preferred by both parties against all other feasible alternatives. Given this, the argument can be (and has been) made that judges should assign rights so as to facilitate this outcome – that is, should “mimic the market” – so that the results which agents would arrive at in a frictionless world can be realized in the world in which we live. Differently put, the Coase theorem provides the philosophical foundation for the idea of efficiency as justice.

Third, the Coase theorem suggests that private agreements can be utilized to resolve externality issues within an appropriate legal framework – not that the theorem’s predictions of an efficient and invariant result carry through in reality, but that significant efficiency gains can be realized if agents are placed within a context where rights are well defined and transaction costs are reasonably low. Such arrangements, it is argued, may in many instances (especially with small numbers of agents) prove to be superior, in an efficiency sense, to those arrived at via more traditional externality remedies. Thus, by demonstrating the veracity of markets/exchange in a frictionless world, the theorem suggests that its underlying processes may offer the best means of dealing with externalities in contexts not too far removed from its basic assumptions. Phenomena as disparate as out-of-court legal settlements, property rights solutions to common pool problems, and marketable emissions permits are seen as evidence of this implication of the Coase theorem’s insights.

The Reciprocity Issue

The Coase theorem literature has had associated with it, almost from the start and in both economics and law, a ideological cast – a cast that has to do in part with divergent views regarding the relative efficacy of market and governmental coordination but also has at least as much to do with the question of the appropriateness of

assigning rights to parties said, according to the conventional wisdom, to be the cause of the externality rather than the ostensible victims of it. But here, too, the Coase theorem set conventional wisdom on its head.

Given that the Coase theorem was born and came of age during a period of great concern over industrial pollution, its suggestion that efficiency will obtain regardless of to which party rights are initially assigned raised some disquiet, as the specter of victims being required to bribe polluters to achieve a reduction in pollution emissions raised its head. Similar concerns arose on the legal front, as scholars raised the possibility that victims would be required to bribe criminals to avoid being robbed or potential victims of tortious harm having to bribe potential injurers into taking precaution against, as in the case of products liability, acting recklessly or producing dangerous products.

The conclusion that the assignment of rights does not impact the allocation of resources reflects perhaps the most important insight provided by the Coase theorem: the reciprocal nature of externalities. This reciprocity can be conceptualized in two ways. The first goes to the notion of “causation,” where the reciprocity issue informs us that it is improper to label one party as *the cause* of the harm. The agent typically labeled the “victim” of the factory’s pollution is as much the cause of the harm as the factory itself: Take away either factory or neighbor, and the externality disappears. The second arena of reciprocity is on the costs front, as each party can be conceived of as a source of uncompensated costs imposed on the other. If the factory has the right to pollute, it visits uncompensated costs upon its neighbor by virtue of the pollution generated. If, on the other hand, the neighbor has the right to be free from pollution, costs are imposed upon the factory which must install pollution abatement equipment, move locations, cease production, or compensate the neighbor for harm caused. In sum, the imposition of costs/harm runs in both directions.

The theorem informs us that the externality will be resolved in efficient and invariant fashion regardless of to which party the relevant rights are initially assigned. But as we move away from the

theorem's idealized world of zero transaction costs, it also suggests that the assignment of rights and the form that these rights take has (perhaps significant) allocative import. Assignments of rights in one direction or another may do more to facilitate bargaining, owing to asymmetries in transaction costs, meaning that judicial decisions have the potential to either encourage negotiation or foreclose it. And when transaction costs are prohibitive, the utilization of liability rules rather than property rules may be expected to generate more efficient outcomes in the end. More generally, the theorem's emphasis on the reciprocal nature of externalities points us to the conclusion that the most efficient resolution of the problem may not be to restrain the actions of the party traditionally identified as the "cause" of the harm.

Conclusion

Questions of the theorem's validity notwithstanding, there can be no question that its influence in economics and in law has been considerable, even if the prolonged debate over it has generated more heat than light. The theorem suggested to economists the possibility of utilizing exchange or market processes to deal with real-world instances of externality where these possibilities had not previously been contemplated. In the legal realm, as well as the economic, the theorem emphasized that traditional notions of causality can be impediments to efficiency and that judicial decisions are not always the ultimate arbiter of rights. Perhaps most importantly, though, the frictionless world contemplated by the theorem brought to the fore the role played by transaction costs in market and exchange processes and the need to come to grips with the influence that legal and other institutions have on the magnitude of these costs.

Cross-References

- [Coase and Property Rights](#)
- [Coase, Ronald](#)
- [Development and Property Rights](#)
- [Efficiency](#)

References

- Allen DW (1990) What are transaction costs? *Res Law Econ* 14:1–18
- Coase RH (1959) The federal communications commission. *J Law Econ* 2(1):1–40
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Demsetz H (1964) The exchange and enforcement of property rights. *J Law Econ* 7:11–26
- Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. *J Pol Econ* 98(6):1325–1348
- Medema SG, Zerbe RO Jr (2000) The Coase theorem. In: Boudewijn BBA, Geest GD (eds) *The encyclopedia of law and economics*. Edward Elgar, Aldershot, pp 836–892
- Mishan EJ (1971) The postwar literature on externalities: an interpretive essay. *J Econ Lit* 9(1):1–28
- Stigler GJ (1966) *The theory of price*. Macmillan, New York

Further Reading

- Ayres I, Talley E (1995) Solomonic bargaining: dividing a legal entitlement to facilitate coasean trade. *Yale Law J* 104(5):1027–1117
- Buchanan JM (1973) The Coase theorem and the theory of the state. *Nat Resour J* 13:579–594
- Calabresi G (1991) The pointlessness of pareto: carrying Coase further. *Yale Law J* 100:1211–1237
- Calabresi G, Melamed AD (1972) Property rules, liability rules and inalienability: one view of the cathedral. *Harv Law Rev* 85(6):1089–1128
- Coase RH (1988) Notes on the problem of social cost. In: *The firm, the market, and the law*. University of Chicago Press, Chicago, pp 157–185
- Cooter R (1982) The cost of Coase. *J Leg Stud* 11(1):1–33
- Demsetz H (1972) When does the rule of liability matter? *J Leg Stud* 1(1):13–28
- Ellickson RC (1986) Of Coase and cattle: dispute resolution among neighbors in Shasta County. *Stanford Law Rev* 38(3):623–687
- Farber DA (1997) Parody lost/pragmatism regained: the ironic history of the Coase theorem. *Va Law Rev* 83:397
- Halteman J (2005) Externalities and the Coase theorem: a diagrammatic presentation. *J Econ Educ* 36(4):385–390
- Hoffman E, Spitzer ML (1982) The Coase theorem: some experimental tests. *J Law Econ* 25(1):73–98
- Medema SG (1994) Ronald H Coase. Macmillan, London
- Medema SG (1999) Legal fiction: the place of the Coase theorem in law and economics. *Econ Philos* 15(2): 209–233
- Medema SG (2009) *The hesitant hand: taming self-interest in the history of economic ideas*. Princeton University Press, Princeton
- Parisi F (2003) Political Coase theorem. *Public Choice* 115(1/2):1–36

- Posner RA, Parisi F (2013) *The Coase theorem*, vol 2. Edward Elgar, Cheltenham
- Samuels WJ (1974) The Coase theorem and the study of law and economics. *Nat Resour J* 14:1–33
- Usher D (1998) The Coase theorem is tautological, incoherent or wrong. *Econ Lett* 61(1):3–11

Coase Theorem and the Theory of the Core, The

Varouj A. Aivazian¹ and Jeffrey L. Callen²

¹Department of Economics, University of Toronto, Toronto, ON, Canada

²Rotman School of Management, University of Toronto, Toronto, ON, Canada

Abstract

Aivazian and Callen (1981) and a number of their subsequent papers use cooperative game theory and core theory to show that the Coasean efficiency result is not robust when there are more than two players. Drawing primarily on their results, this chapter systematically explains the main argument and its extensions as follows. First, the Coase theorem could break down when there are more than two participants because the core of the negotiations may be empty under one set of property rights and nonempty under another. Second, transaction costs will tend to aggravate the empty core problem and make it more likely that the Coasean efficiency result will fail. Third, Pareto optimality can be achieved when the core is empty by the imposition of constraints on the bargaining process and the use of penalty clauses and binding contracts. Overall, the results indicate that it is important to distinguish between transaction costs (when the core exists) and costs due to the empty core because each has different implications for rationalizing institutions. This chapter also summarizes experimental results indicating that the existence of the core is an important determinant of negotiations generally and the Coase theorem in particular. It also points out that some of the problems raised for Coasean

efficiency by the empty core also arise under alternative (non-core) notions of coalitional stability.

Introduction

The Coase theorem states that with well-defined property rights and in the absence of transaction costs, Pareto-efficient allocations will emerge through negotiations among the players to internalize any externality among them, regardless of the initial assignment of property rights (Coase 1960). This result obtains because participants will costlessly recontract around property rights assignments that fail to be Pareto efficient. Coase (1960) also emphasizes the central importance of transaction costs for resource allocation, focusing on efficient property right structures when transaction costs are significant.

As argued by a number of scholars, the bargaining mechanism over property rights in the Coase theorem can be fruitfully framed in terms of cooperative game theory (Arrow 1979; Davis and Whinston 1965). Focusing on core theory, a branch of cooperative game theory, the Coase theorem can be interpreted as: with zero transaction costs, the grand coalition will always emerge regardless of the initial allocation of property rights among the players, and irrespective of whether or not the core of a superadditive characteristic function exists. (Telser (1994) provides a compelling discussion of the power of core theory). Aivazian and Callen (1981) and a number of their subsequent papers employ cooperative game theory and core theory to show that while the Coasean efficiency result is robust for the case of a two-person game, it may fail when there are more than two players. We review these results drawing on Aivazian and Callen (1981, 2003), Aivazian et al. (1987), and Aivazian et al. (2009) as well as on some of the papers that commented on the original 1981 Aivazian-Callen study.

Aivazian and Callen (1981) show that the Coasean efficiency result may fail in a zero transaction cost environment with at least three players and two externalities, in which there are gains from cooperating and forming coalitions to

internalize the externalities. Specifically, in their example, the core is empty under one set of property rights, but nonempty under the other. With an empty core, cycling among coalitions could occur, preventing attainment of the grand coalition and Pareto efficiency. In response to the Aivazian and Callen counterexample to his theorem, Coase (1981) asserts that the zero transaction cost environment underlying his theorem is uninteresting in and of itself. He argues that if transaction costs are imposed on the negotiations in the Aivazian-Callen example, an empty core is less likely to obtain implying that the counterexample is essentially uninformative. However, extending their counterexample to allow for a reasonable transaction cost technology that is convex in the number of coalition partners, Aivazian and Callen (2003) demonstrate that transaction costs tend to aggravate the empty core problem making the breakdown of Coasean efficiency even more likely.

The Empty Core Argument without and with Transactions Costs

The original Aivazian and Callen (1981) analysis involves two polluting firms (A and B) and a laundry (C) and shows that when the polluting firms are liable (when the laundry has the property rights), the Pareto efficient outcome emerges; but when they are not liable, the core is empty and negotiations cycle without necessarily converging to the grand coalition or any other specific outcome. This can be demonstrated by representing the Aivazian and Callen (1981) example in the form of the following normalized characteristic function where V denotes joint coalitional profits:

$$V(i) = 0 \text{ all } i = A, B, C \quad (1a)$$

$$V(A, B) = a, V(A, C) = b, V(B, C) = c \quad (1b)$$

$$V(A, B, C) = d \quad (1c)$$

where a, b, c, d are positive constants, and $d > a, b, c$, for superadditivity. The Pareto optimal outcome corresponds to the grand coalition outcome

$V(A, B, C)$. Note that the characteristic function will be different under different property rights since what each coalition can guarantee itself depends on the prevailing property rights arrangements (Shubik 1984, Chap. 19). Necessary and sufficient condition for the core to be empty (when A and B are not liable) is

$$d < 1/2(a + b + c). \quad (2)$$

If the latter inequality obtains, the grand coalition outcome is not guaranteed so that specific property rights matter for efficiency.

Aivazian and Callen (2003) extend their 1981 paper to allow for transaction costs in the negotiation process by making the reasonable assumption that the costs of forming a coalition are convex in the number of players in the coalition, that is, coalition formation costs increase at an increasing rate with the number of coalition partners. This is reasonable since the number of communication channels required to obtain agreement among coalition members is also convex in the number of members. Two important conclusions emerge. First, if the core is empty in the absence of transaction (coalition formation) costs, then it is necessarily empty with such costs. Second, even if the core is not empty in the absence of transaction costs, such costs could generate an empty core.

What Have We Learned from the Empty Core Argument?

Several lessons emerge from the original Aivazian and Callen (1981) paper and the literature it has spawned (see, e.g., Bernholz (1997), De Bormier (1986), Hurwicz (1995), and Mueller (2003)). First, the Coase theorem may break down when there are more than two participants because the core of the negotiations may be empty under one set of property rights and nonempty under another. As a consequence, even in the absence of other transactions costs, the empty core is likely to impose particular costs of its own. Specifically, cycling induced by the empty core will tend to increase bargaining costs, diminish the value of the exchange opportunity as the negotiation

process is prolonged, and the exchange is postponed, potentially keep at least one coalition unsatisfied and yield a non-Pareto allocation of resources if eventually the grand coalition does not form (Aivazian and Callen (1981, 2003); Shubik (1983), p. 150–151). In fact, Bernholz (1997) argues that the empty core in the Aivazian and Callen example is equivalent to cyclical social preferences. Second, the empty core problem arises as the number of participants increases only when additional participants bring in additional externalities (see Mueller (2003), De Bornier (1986), and the discussion in Aivazian and Callen (2003) on this point). Third, transaction costs may well aggravate the empty core problem, especially if they are incurred prior to the bargaining process (Aivazian and Callen 2003; Anderlini and Felli 2006). Fourth, Pareto optimality may be achieved when the core is empty if there are reputational (transaction) costs with the breaking of agreements (Guzzini and Palestini (2009), or, from a normative perspective, by the imposition of constraints on the bargaining process (e.g., limiting negotiations to only certain sub-coalitions) and the use of penalty clauses and binding contracts (Bernholz 1997; Telser 1994; Aivazian and Callen 2003). As a consequence, it is important to distinguish between transaction costs (when the core exists) and costs due to the empty core because each has a different implication for rationalizing institutions. As Aivazian and Callen (2003, p. 291–292) emphasize:

It is wrong to conclude, therefore, that once transaction costs are introduced, then the problem of the empty core disappears and a Pareto optimal solution obtains. Rather, in such circumstances negotiations may break down more quickly and which specific coalition structure (the grand coalition or a proper sub-coalition) obtains cannot be specified a priori. Even if transaction costs were to force an equilibrium, nothing insures that the equilibrium is Pareto optimal ... It may seem difficult to distinguish empirically between institutional arrangements that arise because of the nonexistence of the core from those that arise from the transaction costs of bargaining when there is a core. After all, the nonexistence of the core will also manifest itself in transaction costs, through the opportunity cost of (negotiation) time. However, the fact that an empty core can arise in the absence of bargaining costs, although these costs exacerbate the empty core

problem, means that the costs generated by an empty core are fundamentally different from the transactions costs of bargaining. Indeed, what is unique about the empty core is that, in addition to direct bargaining costs, it gives rise to costs such as the erosion of the value of the exchange opportunity as it is postponed or the possibility of settling down to a non-Pareto optimal coalition (a proper sub-coalition).

Non-core Coalitional Stability

Many of the issues raised by the empty core also arise under alternative (non-core) notions of coalitional stability. As Aivazian et al. (1987) argue, for the Coase theorem to obtain, the grand coalition must be stable and, moreover, no other coalition can be similarly stable because otherwise a Pareto optimal allocation cannot be guaranteed. Aivazian et al. (1987) extend the Aivazian-Callen (1981) example to Aumann and Maschler (1964) bargaining set notions of coalitional stability by showing that while a specific Pareto optimal allocation of resources obtains for one set of property rights, a non-Pareto optimal allocation may well obtain for another set of property rights that involves bargaining. Indeed, under one type of bargaining set stability, they find that every coalition but the grand coalition is stable, completely vitiating the Coase theorem.

Testing the Implications of the Empty Core

It is unlikely that archival data are available that would allow one to test the implications of the empty core problem for the Coase theorem. Instead, Aivazian et al. (2009) investigate the Coase theorem experimentally in a bargaining game in which the final allocation of payoffs differ in terms of whether the core exists and in the initial allocation of property rights among the players. The experimental results indicate that the existence of the core is an important determinant of negotiations generally and the Coase theorem in particular. They find that when the core is empty and property rights are ill defined, Coasean efficiency breaks down. In particular, the number

of non-Pareto optimal agreements and negotiation rounds with cycling are significantly larger when the core is empty than when it exists, particularly when property rights are ill defined.

Conclusion

The upshot of the empty core issue for the Coase Theorem can be summarized as follows (Aivazian and Callen 2003, p. 296):

“In the real world opportunities for exchange are sometimes manifold and the bargaining strategies potentially complex. The Coase Theorem masks this reality by presupposing that exchange occurs between two parties . . . with more than two parties, and at least two externalities, coalitional behavior may predominate. In which case, under some property rights arrangements the core may not exist; as a result, the Coase Theorem may fail to hold. Transactions costs may well exacerbate the empty core problem. In such circumstances, specific property rights arrangements, and contractual schemes such as penalty clauses, binding contracts, and restrictions on the sequence of bargaining, may emerge to attenuate the problems engendered by the nonexistence of the core”

Cross-References

- [Coase and Property Rights](#)
- [Coase, Ronald](#)
- [Coase Theorem](#)
- [Coase Theorem: Empirical Tests](#)

References

- Aivazian VA, Callen JL (1981) The Coase theorem and the empty core. *J Law Econ* 24(1):175–181. Reprinted in R.A. Posner and F. Parisi (eds.) *The Coase Theorem. Economics Approaches to Law Series*. 2013. Edward Elger Publishing.
- Aivazian VA, Callen JL (2003) The core, transaction costs, and the Coase theorem. *Constit Polit Econ* 14(4): 287–299
- Aivazian VA, Callen JL, Lipnowski I (1987) The Coase theorem and coalitional stability. *Economica* 54(216):517–520. Reprinted in R.A. Posner and F. Parisi (eds.) *The Coase Theorem. Economics Approaches to Law Series*. 2013. Edward Elger Publishing.
- Aivazian VA, Callen JL, McCracken S (2009) Experimental tests of core theory and the Coase theorem: inefficiency and cycling. *J Law Eco* 52(4):745–759. Reprinted in R.A. Posner and F. Parisi (eds.) *The Coase Theorem. Economics Approaches to Law Series*. 2013. Edward Elger Publishing.
- Anderlini L, Felli L (2006) Transactions costs and the robustness of the Coase theorem. *Econ J* 116(508): 223–245
- Arrow KJ (1979) The property rights doctrine and demand revelation under incomplete information. In: Boskin MJ (ed) *Economics and human welfare: essays in honor of Tibor Scitovsky*. Academic Press, New York, pp 23–29
- Aumann RJ, Maschler M (1964) The bargaining set for cooperative games. In: Dresher M, Shapley LS, Tucker AW (eds) *Advances in game theory*. Princeton University Press, Princeton, pp 443–476
- Bernholz P (1997) Property rights, contracts, cyclical social preferences and the Coase theorem: a synthesis. *Eur J Polit Econ* 13:419–442
- Coase R (1960) The problem of social cost. *J Law Econ* 3(1):1–44
- Coase R (1981) The Coase theorem and the empty core: a comment. *J Law Econ* 24(1):183–187
- Davis OA, Whinston AW (1965) Some notes on equating private and social cost. *South Eco J* 32(2):113–126
- De Bornier JM (1986) The Coase theorem and the empty core: a reexamination. *Int Rev Law Econ* 6(2):265–271
- Guzzini E, Palestini A (2009) The empty core in the Coase theorem: a critical assessment. *Econ Bull* 29(4): 3095–3103
- Hurwicz L (1995) What is the Coase theorem? *Jpn World Econ* 7:49–74
- Mueller DC (2003) *Public choice III*. Cambridge University Press, New York
- Shubik M (1983) *Game theory in the social sciences: concepts and solutions*, vol 1. M.I.T. Press, Cambridge, MA
- Shubik M (1984) *Game theory in the social sciences: a game theoretic approach to political economy*, vol 2. M.I.T. Press, Cambridge, MA
- Telser L (1994) The usefulness of core theory in economics. *J Econ Perspect* 8(2):151–164

Coase Theorem: Empirical Tests

Elodie Bertrand

CNRS and University Paris 1 Panthéon-Sorbonne (ISJPS, UMR 8103), Paris, France

Abstract

This entry examines three classic tests of the Coase theorem: two empirical studies on the

most famous examples of externalities – Coase’s ranchers and farmers (Ellickson, *Stanford Law Rev.* 38(3):623–687, 1986) and Meade’s bees (Cheung, *J Law Econ* 16(1):11–33, 1973) – as well as the seminal laboratory experiments of Hoffman and Spitzer (*J Law Econ* 25(1):73–98, 1982). I will insist on the difficulty of testing the theorem without measuring transaction costs, and on another obstacle to negotiation over and above these costs: a moral or social prohibition on exchange.

Introduction

In *The Problem of Social Cost*, Coase (1960) used examples to suggest that, in the presence of externalities, if transaction costs are nil and if property rights are clearly defined and allocated, agents bargain over rights and achieve an optimal output that is independent of the initial allocation of rights. This proposition was to be called the “Coase theorem” by Stigler (1966, p. 113).

The theorem has long been tested against the facts, whether already existing (empirical studies) or purposefully constructed (laboratory experiments). Strictly speaking, however, these studies do not in fact test the Coase theorem. The reasons for this are twofold: first, the conclusion of almost six decades of debate over the validity of the theorem is generally taken to be that all criticisms can be subsumed under the category of transaction costs and that this renders the “theorem” tautological (Medema and Zerbe 2000). Second, transaction costs never being nil, the theorem does not apply to the real world (most of Coase’s seminal article actually examined the consequences of these costs). Regarding experiments, they reduce transaction costs to a minimum. As for empirical studies, they can only test a “generalized Coase theorem” (Bertrand 2011): if the gain from a transaction concerning a right (the use of which provokes side effects) is greater than its cost, then the transaction takes place. I will insist on the difficulty of testing such a proposition without measuring transaction costs, and on another obstacle to negotiation over and above these costs: a moral or social prohibition on exchange. For a

detailed review of the tests of the Coase theorem, see Medema and Zerbe (2000).

Empirical Studies of the Coase Theorem with Externalities

Some empirical tests of the theorem have concerned pretrial settlements (Galanter 1983) and transactions on nuisances after trials (Farnsworth 1999), but the majority test the prediction of the absence of any effects engendered by a change in the law: notably with regard to share tenancy arrangements (Cheung 1969), rules governing divorce (Peters 1986), the effect of a payment of bonuses to unemployed workers when they obtain a job or to employers when they hire unemployed people (Donohue 1989), and generally with a very specific attention to sports (Hylan et al. 1996; Cymrot et al. 2001; Szymanski 2007). I will focus on two tests which concern well-known examples of externalities: Coase’s cattle that destroy crops on neighboring land (Ellickson 1986), and Meade’s bees that pollinate orchards while foraging for nectar (Cheung 1973). The cattle case was also examined by Vogel (1987), who analyzed the effects of the changes in trespass law in the different counties of California during the second half of the nineteenth century: the assignment of rights to farmers tended to increase farm output, contrary to the prediction of the theorem (this movement was studied over a longer period than that of Ellickson, which explains their different results). Hanley and Sumner (1995) also studied the resolution of externalities due to red deer in Scotland, and their conclusions are close to Ellickson’s.

Ellickson (1986) intended to test the realism of the “Parable of the Farmer and the Rancher” (p. 624), the numerical example developed by Coase (1960) in which, on the basis of their formal entitlements, a cattle raiser and a crops farmer negotiate the size of the herd in exchange for a monetary payment, and arrive at the same size whatever the initial allocation of rights is. In 1982, therefore, Ellickson turned to Shasta County, in the north of California, where two legal regimes coexisted: most of the county was under the historical “open-range” regime, in

which a cattleman is not liable for trespass damages even when negligent, but one district had switched to a “closed-range” regime in 1973, where the rancher is always liable for damage even in the absence of negligence.

Although the allocation of resources (e.g., spaces respectively devoted to crops and cattle) is not affected by changes in liability law, the explanation does not turn on the monetary negotiations invoked by Coase: neighbors rather refer to social norms to resolve their disputes; and since these norms are independent from formal law, the resolution of incidents is as well.

In particular, the parties to a dispute refer to an informal rule according to which the cattle owner is liable for the actions of his animals. This goes against the formal rule of the main part of the county, which implies that “norms, not legal rules, are the basic sources of entitlements” (Ellickson 1986, p. 672). Other norms detail the manner of resolving an incident: most of the time, the victim downplays the incident, which will be quickly solved by an exchange of civilities, and neighbors expect reciprocity. In any case, the resolution of the conflict must be informal, without recourse to law or courts, and monetary compensation is forbidden: inhabitants “regard a monetary settlement as an arms-length transaction that symbolizes an unneighborly relationship” (ibid.: 682).

This monograph therefore shows that neighbors do not solve their disputes with monetary transactions, but through social norms, which Ellickson explains by high transaction costs. But, first, he does not measure them: he simply infers them from the facts that exchanges do not take place and that the law would be costly to learn (a cost itself inferred from the fact that inhabitants do not know it). Second, it could be argued that transaction costs are low: small number of parties, easily identifiable, easy monetary assessment of the damages, sharing of the same social norms, etc. Third, other obstacles than transaction costs may explain the absence of exchange: social norms can impede a market from developing (Bertrand 2011). Take the norm that condemns monetary settlements: Ellickson explains this by reference to transaction costs, but we could just as well assert that it is this

norm that prevents transactions in the first place, and hence that the norm lies at the origin of the impossibility of negotiations. Transaction costs would be irrelevant since agents are not willing to negotiate.

The absence of monetary payments and of reference to formal rules thus raises questions about the legitimization of the legal structure of rights and of their alienability, which may deter agents from negotiating these rights, quite apart from transaction costs. In the Shasta County case, there is no norm indicating that you can transact over the right to destroy your neighbor’s crops. In fact, we here encounter the basic problem of externality: a market does not exist, and it may be difficult to create one out of nothing because of social norms. By contrast, the next example will stress that a market does seem to exist for what was long considered as externalities: pollination and nectar services.

Cheung’s “Fable of the Bees” (1973) tests the realism of Meade’s (1952) example of the externality between beekeepers and orchard owners. It was written at Coase’s request, who was not satisfied with Johnson’s study (1973) – the latter being more focused on the institutional setting of the pollination service market. Cheung’s investigation was conducted in the state of Washington in spring 1972; it covered a sample of 9 beekeepers and a total of approximately 10,000 spring colonies.

Cheung first observes that a marketplace exists for transactions on pollination and nectar services. On the one hand, an orchard owner can rent the pollination services of an apiarist who places her hives in the orchard; he pays her a monetary pollination fee that depends on the number, density, and strength of hives. On the other hand, an apiarist who wants to place her hives among crops for honey production pays an apiary rent to the orchard owner, mostly in the form of honey, and depending on the honey crop. Note, however, a difference with Coase’s parable of the rancher and the farmer: an orchard owner can choose a beekeeper from among others and vice versa; they are not compelled to negotiate or to find a solution with their neighbor.

Nevertheless, we here have evidence that contracts with monetary exchanges can deal with so-called externalities. It seems that the possibility of exchanges in pollination services (or social

permission) developed in the period after WWI, alongside with the specialization of farms which required specific pollinators. The right of having his plants pollinated was said to be exchangeable and a price was suggested: the set of social norms necessary for the operation of a market was available. Still, regarding the habit of giving some honey to the orchard owner in exchange for the right to place hives in his orchard, the question remains whether this should be considered as a market transaction or a social norm.

In any case, the markets for pollination and nectar services are unusual in that the enforcement of contracts, whether oral or written, relies heavily on social norms. In addition, externalities remain at the margins of the market for pollination services, and they are dealt with by social norms. For example, my neighbor, who also cultivates apples, benefits from the hives I rent for my apple trees. These positive externalities are solved by a social norm of neighborliness called “the rule of the orchards,” by which we both rent the same number of hives per acre (Cheung 1973, p. 30). Cheung explains the absence of monetary exchange to solve this externality by the cost of such a negotiation (not measured, but itself inferred from the absence of such an exchange). As with trespass incidents, we could alternatively explain the absence of negotiation by the existence of an already satisfying norm or, prior to it, by other norms that would deter monetary negotiations between neighbors and make transaction costs irrelevant.

It would be excessive to infer from Ellickson’s and Cheung’s studies that the generalized Coase theorem is empirically confirmed, since both authors have a tendency to explain their observations precisely by this assumption. This kind of empirical study faces the problem of measuring the gains and costs of exchange. However, the control possible in the laboratory allows for a more precise determination of them.

Experiments of the Coase Theorem with Monetary Negotiations

Laboratory experiments concerning the theorem began in 1982 with the publication of studies by

Prudencio (1982) and Hoffman and Spitzer (1982). It is the last protocol, closest to the Coasean parable of the rancher and the farmer, that has been used the most. Experiments of the theorem seek to test the robustness of results given variation of the implicit or explicit assumptions: high number of agents (Hoffman and Spitzer 1986); incomplete information (Prudencio 1982; McKelvey and Page 2000); introduction of transaction costs (time limit in Prudencio 1982 or cost of an offer in Rhoads and Shogren 1999); uncertainty about the payments (Shogren 1992); property rights earned and/or legitimized (Hoffman and Spitzer 1985a), not allocated (Harrison and McKee 1985) or uncertain (Cherry and Shogren 2005); non-convexity (Shogren et al. 2002); empty core (Aivazian et al. 2009); and physical discomfort (Coursey et al. 1987). Schwab’s (1988) experiment does not concern externalities but rather contract presumptions. Another series of experiments consists in a comparison of the efficiency of different solutions to externalities (e.g., Plott 1983; Harrison et al. 1987). The results of the experiments on the endowment effect were applied to refute some of the theorem’s predictions (see Kahneman et al. 1990) but are not specific to the theorem and do not exhibit theorem-like mechanisms.

It is Hoffman and Spitzer’s (1982) first set of experiments, with two persons and complete information, whose design is the closest to the Coasean parable. The respective monetary payoffs (net profit) of the two subjects, who were randomly assigned the letters A (rancher) and B (farmer), depend on the value of a discrete number (the size of the herd): a couple of payoffs are associated to each number (0, 1, etc.) and only one number maximizes the sum of these payoffs. A’s payoff increases when the number increases and conversely for B, which renders the “externality” negative. The allocation of the property right is translated by the designation of one of the subjects as the “controller”: she has the right to unilaterally choose the number (and hence the payoffs). The possibility of negotiation comes from the fact that the other subject may influence her choice by proposing to transfer a part of his own payoff. The controller is randomly designated through a heads or tails game.

As for the results, exchanges take place and are efficient (23 out of 24 decisions choose the number that maximizes the sum of payoffs). The authors infer that this result, confirmed by their other sets (1982) and other experiments (1985a; 1986), “creates a strong presumption in favor of the Coase Theorem” (1985b, p. 1011). Nevertheless, most of the exchanges are not mutually advantageous. The controller sacrifices a part of her gain for fairness. Typically, whereas the controller, B, could obtain \$12 (and A 0) by unilaterally choosing the number 0, A and B sign an agreement by which the controller chooses the number 1 (B earns \$10, and A 4) – which is the maximum total payoff – and share this total gain equally (\$7 for each), which means the controller sacrifices \$5 (Hoffman and Spitzer 1982, p. 86 and 92). As Harrison and McKee (1985, p. 655) conclude, “the Coase Theorem is [therefore] behaviorally ‘right for the wrong reasons.’” In these experiments, subjects agree on the optimal issue, but only because one of them sacrifices a part of her gains. Why does this subject accept the exchange?

Admittedly, fairness is a classic result of bargaining experiments, the robustness of which has been confirmed in large measure by Hoffman and Spitzer’s (1982) own experiments. Fairness is here favored by public face-to-face negotiation; complete information (McKelvey and Page 2000 obtain less good results with incomplete information); and a random allocation of the right, without legitimacy and without insistence on its meaning (the controller obtains her individual maximum more often when the initial allocation of the right follows a preliminary game and when her authority is legitimated by the monitor (Hoffman and Spitzer 1985a), or when she can learn the meaning of her unilateral right (Harrison and McKee 1985)).

But fairness is here obtained by sacrifice. A likely explanation of this non-mutually advantageous exchange is that of a moral or social incentive raised by the experimental design and the instructions (Bertrand 2014). First, contracts are given to the subjects, from which they infer that the right is exchangeable and how they can exchange it. This amounts to morally authorizing the exchange regarding this externality, and even to encouraging it by implying that it is expected

from participants that they use this possibility. Then, the list of payoffs in function of the number gives the subjects the monetary value of the externality, which impedes the perception of the moral problem of giving value to something that should not have one (Kelman 1985, pp. 1038–1039). Finally, instructions are not indifferent regarding the question of exchange, and take care, under the shelter of neutrality, to avoid any vocabulary concerning externalities, constraint, or causality.

Confronted with some of these criticisms, Coursey et al. (1987) built an experiment precisely to test the existence of a moral ban on bargaining in the presence of externalities. They replicate the 1982 experiment, but with the possibility of physical discomfort for the subject who suffers from the externality (keeping in his mouth a very bitter-tasting liquid for 20s). This allegedly allows testing the moral ban against negotiating over something that insults dignity, what they call the “dignity hypothesis.” Pairs of subjects reached the optimal result, and authors hence reject this hypothesis. But here again, the sources of bargaining breakdown are in the most part avoided (Bertrand 2014, pp. 454–456). In particular, and paradoxically in comparison to what the authors wanted to test, the moral obstacles to the exchange are at least partially removed: the contract is provided, and the instructions and consent insist on the possibility of exchange and the possibility of A taking the liquid. Individuals placed in such a situation know what they have to do to please the monitor, or may simply internalize her authority (Milgram 1974).

Conclusion

Since the Coase theorem is circular, what is tested is precisely whether mutually advantageous bargains are indeed realized. Such realization may encounter three obstacles: (1) transaction costs, (2) problems of agreement over the distribution of the surplus, and (3) moral or social prohibition of exchange. This entry has shown that the authors of these tests commonly appeal to transaction costs to legitimize every result as efficient. They thereby rationalize their observations by appeal to

the assumption of efficiency rather than test its correspondence to the real world. And they underestimate the other obstacles.

What are the lessons for the theorem? As stressed by Medema, these studies enlighten us about “situational behavioral norms” (1997, p. 129) and the limits to the assumption of individual maximization, hence calling into question the behavioral assumptions that ground the theorem, and indeed the law and economics movement more generally. Cheung’s bees and Hoffman and Spitzer’s experiment bring to light that markets for what was seen as an “externality” can exist: the right over an “externality” is sold and bought. Two elements have nevertheless been overlooked by Coase (1960): first, the entitlements to which agents refer are not only determined by the law; second, impediments to bargaining other than transaction costs exist, these being moral or social obstacles to such an exchange.

Cross-References

- [Coase Theorem](#)
- [Coase Theorem and the Theory of the Core, The](#)
- [Endowment Effect](#)
- [Nuisance](#)

References

- Aivazian VA, Callen JL, McCracken S (2009) Experimental tests of core theory and the Coase theorem: inefficiency and cycling. *J Law Econ* 52(4):745–759
- Bertrand E (2011) What do cattle and bees tell us about the Coase theorem? *Eur J Law Econ* 31(1):39–62
- Bertrand E (2014) Autorisation à l’échange sur des externalités. De l’interdiction à l’obligation. *Revue Economique* 65(2):439–459
- Cherry TL, Shogren JF (2005) Costly Coasean bargaining and property right security. *Environ Resour Econ* 31(3):349–367
- Cheung SNS (1969) The theory of share tenancy. University of Chicago Press, Chicago
- Cheung SNS (1973) The fable of the bees: an economic investigation. *J Law Econ* 16(1):11–33
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Coursey DL, Hoffman E, Spitzer ML (1987) Fear and loathing in the Coase theorem: experimental tests involving physical discomfort. *J Leg Stud* 16(1):217–248
- Cymrot DJ, Dunlevy JA, Even WE (2001) “Who’s on first”: an empirical test of the Coase theorem in baseball. *Appl Econ* 33(5):593–603
- Donohue JJ III (1989) Diverting the Coasean river: incentive schemes to reduce unemployment spells. *Yale Law J* 99(3):549–609
- Ellickson RC (1986) Of Coase and cattle: dispute resolution among neighbors in Shasta county. *Stanford Law Rev* 38(3):623–687
- Farnsworth W (1999) Do parties to nuisance cases bargain after judgment? A glimpse inside the cathedral. *Univ Chicago Law Rev* 66(2):373–436
- Galanter M (1983) Reading the landscape of disputes: what we know and don’t know (and think we don’t know) about our allegedly contentious and litigious society. *UCLA Law Rev* 31:4–71
- Hanley N, Sumner C (1995) Bargaining over common property resources: applying the Coase theorem to red deer in the Scottish Highlands. *J Environ Manag* 43(1):87–95
- Harrison GW, McKee M (1985) Experimental evaluation of the Coase theorem. *J Law Econ* 28(3):653–670
- Harrison GW, Hoffman E, Rutström EE, Spitzer ML (1987) Coasian solutions to the externality problem in experimental markets. *Econ J* 97(386):388–402
- Hoffman E, Spitzer ML (1982) The Coase theorem: some experimental tests. *J Law Econ* 25(1):73–98
- Hoffman E, Spitzer ML (1985a) Entitlements, rights, and fairness: an experimental examination of subjects’ concepts of distributive justice. *J Leg Stud* 14(2):259–297
- Hoffman E, Spitzer ML (1985b) Experimental law and economics: an introduction. *Columbia Law Rev* 85(5):991–1036
- Hoffman E, Spitzer ML (1986) Experimental tests of the Coase theorem with large bargaining groups. *J Leg Stud* 15(1):149–171
- Hylan TR, Lage MJ, Treglia M (1996) The Coase theorem, free agency, and major league baseball: a panel study of pitcher mobility from 1961 to 1992. *South Econ J* 62(4):1029–1042
- Johnson DB (1973) Meade, bees, and externalities. *J Law Econ* 16(1):35–52
- Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. *J Polit Econ* 98(6):1325–1348
- Kelman M (1985) Comment on Hoffman and Spitzer’s “Experimental law and economics”. *Columbia Law Rev* 85(5):1037–1047
- McKelvey RD, Page T (2000) An experimental study of the effect of private information in the Coase theorem. *Exp Econ* 3:187–213
- Meade JE (1952) External economies and diseconomies in a competitive situation. *Econ J* 62:54–67
- Medema SG (1997) The trial of homo economicus: what law and economics tells us about the development of economic imperialism? In: Davis JB (ed) *New economics and its history*, suppl. 29. Duke university Press, Durham, pp 122–142

- Medema SG, Zerbe RO Jr (2000) The Coase theorem. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics*, vol 1. Edward Elgar, Cheltenham, pp 836–892
- Milgram S (1974) Obedience to authority. Harper & Row, New York
- Peters HE (1986) Marriage and divorce: informational constraints and private contracting. *Am Econ Rev* 76(3):437–454
- Plott CR (1983) Externalities and corrective policies in experimental markets. *Econ J* 93(369):106–127
- Prudencio YC (1982) The voluntary approach to externality problems: an experimental test. *J Environ Econ Manag* 9(3):213–228
- Rhoads TA, Shogren JF (1999) On Coasean bargaining with transaction costs. *Appl Econ Lett* 6(12):779–783
- Schwab S (1988) A Coasean experiment on contract pre-sumptions. *J Leg Stud* 17(2):237–268
- Shogren JF (1992) An experiment on Coasian bargaining over ex ante lotteries and ex post rewards. *J Econ Behav Organ* 17(1):153–169
- Shogren JF, Moffett R, Margolis M (2002) Coasean bargaining with nonconvexities. *Appl Econ Lett* 9(15):971–977
- Stigler GJ (1966) *The theory of price*, 3rd edn. Macmillan, New York
- Szymanski S (2007) The champions league and the Coase theorem. *Scott J Polit Econ* 54(3):355–373
- Vogel KR (1987) The Coase theorem and California animal trespass law. *J Leg Stud* 16(1):149–187

Codes of Conduct

George Tsourvakas
School of Journalism and Mass Communication,
Aristotle University of Thessaloniki,
Thessaloniki, Greece

Abstract

Codes of conduct are self regulated rules adopted by people or organizations. All types of associations and organizations today have professional codes of conduct to avoid negative behaviors and to improve quality practices.

Keywords

Accountability; Best practices; Business ethics; Corporate social responsibility; Ethical code; Professional guide; Sustainable development

Synonyms

Accountability; Corporate social responsibility; Professional guides; Quality entrepreneurship; Related issues are business ethics; Sustainable development, and code of best practices (Schwartz 2004)

Definition

A code of conduct is both a sum of the ideal values or principals and a guide for good practice for individuals and organizations; it targets the reduction of negative externalities, the internalization of some social costs, and, at the same time, increasing benefits to society.

Theoretical Framework

Why do more and more organizations today try to follow, provide, and measure their social contribution based on a code of conduct? First, a code of conduct is a precaution mechanism. Following appropriate social behavior is less costly and risky than the costs of penalties, law suits, punishments, harm done to a brand name, customer boycotts, negative word of mouth, (Schwartz 2004).

Second, the strict enforcement of codes of conduct is a voluntary and self-regulated mechanism for companies that is not imposed by legislation and regulations. They are a signal and means to inform people that a company wishes to stay in the market for a long time and create a respected brand name; therefore, it is very important for companies to build a good image (Carroll and Buchholtz 2006).

Third, codes of conduct are an ex-ante strategy and not ex-post, which means that elements of positive ecological and social behavior could form part of a business strategy, giving companies the opportunity to think about how to operate in a way that offers something to society or without doing harm. This strategy could make companies very innovative and creative (Porter and Kramer 2006).

Fourth, companies that enforce codes of conduct in their policies and strategy significantly reduce their operational and transactional costs. This is because green companies save energy not only for society but themselves. Moreover, all types of stakeholders are happier to cooperate or work with socially responsible organizations, e.g., suppliers, distributors, shareholders, and bankers are more positive towards dealing with social responsible companies, reducing the cost of searching for, bargaining for, and enforcing contracts. Hence, employees who work in socially responsible organizations are more motivated and committed to the organizational dream and are professional in their work. Furthermore, codes of conduct can help organizations to achieve a balance between different stakeholder's interests (Thomsen 2001).

Fifth, companies that follow codes of conduct gain a positive reputation, and this is a guarantee to customers and citizens that they will not undertake opportunistic behavior such as higher prices, low quality, misleading practices, or intrusive behavior, but that they will respect them and show good will and trust. Thus, codes of conduct are a trust mechanism. Consequently, they make a positive contribution to the community, society, culture, and ecology development through payments, donations, and philanthropy for a better future. Therefore, socially responsible firms pursue a win-win strategy (Kotler and Lee 2005).

Empirical Evidence

Relevant cases that provide empirical evidence of this benefit are most national and international codes of conduct for hospitals and communication organizations or for individual doctors and journalists. According to their associations, there are behaviors that should be avoided, such as harassment and discrimination, advertising, disclosure of private information, vulgar language, being influenced by power, bribery, and corruption. At the same time, there are good practices that will benefit citizens and society, such as respecting people's dignity, explaining medical

procedures, asking for permission for research, offering services voluntarily in emergency situations, and improving national culture and language.

Conclusion

All types of organizations could adopt altruistic corporate codes of conduct which complement legal requirements with a positive internal and external micro and macro environment.

Companies could care for all stakeholders, for instance employees, by providing training and healthy and safe conditions, and by discussing problems and respecting privacy, and at the same time offering sports, cultural, social, and ecological activities for their members. The same could be done for the micro external environment for customers, suppliers, distributors, and competitors. For instance, for customers, companies could offer safe products without their being produced by child labor or using animal testing, with logical prices, and with the products being eco-friendly and recyclable. Finally, with regards to the community, social, cultural, and physical macro environment, companies could incorporate into their professional codes proposals to solve social problems. For instance, communication organizations could support educational improvements and knowledge regarding educational problems that will change citizen's living conditions and lead to a better society.

References

- Carroll AB, Buchholtz AK (2006) *Business and society: ethics and stakeholder management*, 6th edn. Thomson, Toronto
- Kotler P, Lee N (2005) *Corporate social responsibility: doing the most good for your company and your cause*. Wiley, Hoboken
- Porter M, Kramer M (2006) Strategy and society: the link between competitive advantage and social responsibility. *Harv Bus Rev* 84(12):78–92
- Schwartz MS (2004) Effective corporate codes of ethics: perceptions of code users. *J Bus Ethics* 55(4): 323–343
- Thomsen S (2001) Business ethics as corporate governance. *Eur J Law Econ* 11(2):153–164

Codes of Ethics

Luc Van Liedekerke¹ and Peter-Jan Engelen^{2,3}

¹Faculty of Applied Economics, University of Antwerp, Antwerpen, Belgium

²Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

³Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

Definition

A written set of guidelines issued by an organization to its stakeholders (primarily its employees) to help them conduct their actions in accordance with the primary values of the organization.

Introduction

The term “code of ethics” is often used interchangeably with other terms like code of conduct, deontic code, compliance code, integrity code, business code, etc. It is possible to range all these codes along a continuum with at the one end compliance-directed codes and at the other end ethics-directed codes. Compliance-directed codes are mainly focused on guaranteeing that its users stick to certain legal rules, while ethics codes start out from the values and norms of the organization. While this distinction is useful in order to characterize a code, almost all business codes today are a mixture of compliance and ethics. As will become clear in the historical section, there is also a clear legislative reason why most codes today are a combination of ethics and compliance.

Codes come in a huge diversity of form and content. An example that invited a lot of mockery is the 52 pages dress code launched by the Swiss bank UBS in 2010. It contained advice like “You can extend the life of your knee socks and stockings by keeping your toenails trimmed and filed” and discussed issues like eating garlic, the color of lingerie for women, and the length of thermic underwear (UBS 2010). On the other hand, we find examples of ethics codes that were crucial in

the formation of an organization and continue to dominate its identity. A prime example is the Johnson and Johnson Credo. Launched in 1943, this one pager is still cherished by the company as the reference document that guides their decision-making (Johnson & Johnson 1943).

One could argue that the first ethics code was “do not eat the apple,” and it immediately signals two fundamental issues with ethics codes: they are hard to implement and if broken imply considerable risk. Enron is a more recent example of the same. Its elaborate ethics code was launched in 2000 just one year before its spectacular collapse made abundantly clear that nobody had ever taken this code serious (Sims and Brinkmann 2003). Codes can be found in many places. Medical professions (Hippocratic Oath) and liberal professions are among the earliest to use codes of ethics as central element in self-regulatory efforts and still use them today (Higgs-Kleyn and Kapelianis 1999; Gaumnitz and Lere 2002). But not only barristers also pirates had their written codes of ethics (Kukla 2014). In this contribution we will not go into the many deontic codes for liberal and medical professions but instead concentrate on the history and use of ethics codes in business (Brooks 1989). We further narrow down the analysis to the Anglo-Saxon use of code of ethics and compliance as it can be argued that the Anglo-Saxon way of encoding ethics in business organizations has become the dominant form, certainly among large, stock-quoted companies (Rodriguez-Dominguez et al. 2009).

Evolution in Codes of Ethics

The history of ethics/compliance codes in (USA) business is driven by a cycle of scandal and regulatory response (Farrell et al. 2002). The first wave of scandals we would like to mention is best known through the Lockheed scandal but was not limited to Lockheed alone (Shaplen 1978). The Lockheed scandal encompassed a series of bribes and contributions made by Lockheed personnel in the process of negotiating the sale of aircraft. In fact in the mid-1970s, the US Securities and Exchange Commission

investigated over 400 US companies who admitted making questionable or illegal payments in excess of \$300 million to foreign government officials, politicians, and political parties. In response President Carter enacted the Foreign Corrupt Practices Act (FCPA). This led to a first wave of compliance and ethics codes mainly among large, exporting companies with extensive contacts in the public sector. Defense industry scandals in the 1980s (defense contractors charged, for instance, 400\$ for a simple hammer) lead to the Defense Industry Initiative (DII) that aimed and provided guidelines for building strong internal compliance programs in the defense industry (Kurland 1993). It is probably fair to say that this is one of the reasons why the defense industry has until today among the best-developed compliance and ethics programs.

A regulatory milestone for compliance programs was the reform of the Federal Sentencing Guidelines in 1991. These guidelines describe the elements of an organization's compliance and ethics program that are required to be considered for eligibility for a reduced sentence. It was the first time the government provided clear guidance on how a good compliance program needed to look like. After its introduction, it became clearly beneficial for companies to have a compliance program as judicial risk was directly linked to the presence of a decent compliance program (Ferrell et al. 1998). It resulted in a second, much larger wave of compliance programs in the USA, with expansions mainly in other Anglo-Saxon countries. Sentencing guidelines with reference to compliance were introduced in several countries among others the UK, France, Australia, Canada, Israel, and Singapore.

Around 2000 a new wave of scandals hit the USA with Enron undoubtedly the most familiar household name. During this time, Sarbanes-Oxley (SOX) regulation was the response by the regulator, and it contained, among other things, another review of the sentencing guidelines (published in 2004). The remarkable part about this revision is that every reference to the word "compliance" in the 1991 version was now replaced by "ethics and compliance." This fundamentally changed the target of compliance

programs (Canary and Jennings 2008). It was no longer just "following the law;" instead creating an ethical culture inside the organization was now the goal which is vastly more complicated and uncertain. The reason for this change was that by 2000 many companies, most notably Enron itself, had established compliance programs, but the wave of scandals made painfully clear that these compliance programs failed to have any influence whatsoever on the organizational attitude toward the law (Sims and Brinkmann 2003). Clearly more and better compliance was needed. The change would have to reach deeper: the ethics of the organizations and its representatives needed to change.

The last wave of scandals was the financial crisis of 2008. Again governments reacted by regulatory change (Dodd-Frank in the USA, Mifid2 and Solvency2 in Europe, the Financial Instruments Exchange Act in Japan, etc.), with strong influence on compliance and ethics programs mainly in the financial sector.

In general one can say that over the past decades, ethics and compliance programs have increased dramatically. There are several reasons for this. First, there is a trend among legislators to push risks downward toward the individual firm. A good example from the financial sector is anti-money laundering (AML) regulation. While it used to be the regulator that had to look out for money laundering, now it is the financial service provider himself that carries the main responsibility to control for AML. Second, the cost of non-compliance has clearly increased over the past decades. Spectacular failures like Siemens, HSBC, or BNP each time involving billion dollar settlements have pushed companies to invest more in compliance and ethics programs. Third, the risk approach in business tends to take ethics and compliance much more serious and integrates it into an overall risk strategy.

As the regulatory environment becomes ever more complex, it is very likely that ethics and compliance departments will continue to grow. Nevertheless, ethics and compliance departments today remain effectively small. Most companies rely on one or at most a couple of full-time equivalents (FTE) to run their ethics departments.

Budgets are limited to at most a couple of million dollars, except for heavily regulated industries like finance, defense, and pharmaceuticals where it can easily reach hundreds of millions.

Usefulness of Codes of Ethics

Despite the undeniable rise of ethics and compliance programs in business, it is hard to pinpoint any real progress with respect to integrity in business. For instance, survey after survey indicates that occupational fraud, corruption, and other forms of maleficence are not going down (Carberry et al. 2018). Ethics codes and compliance programs have therefore consistently been criticized from different sides. Managers resent the million dollar investment without clear return. Business ethicists have criticized the programs as at best shallow and at worst a hindrance to ethical conduct (Benson 1989; Frankel 1989; Stevens 1994; Painter-Morland 2010). A lot has to do with the crucial issue of how exactly a code of ethics is enacted inside the organization (Schwartz 2004; Munter 2013). Today implementation of ethics codes is primarily done through signing of the code (often part of the labor contract) and online training. Mandatory exercises on topics such as privacy, insider trading, and bribery are concluded by a ten-question quiz at the end. Many employees resent these mindless training sessions that are experienced as a series of box-checking routines, and according to business ethicists, instead of increasing ethics awareness, such an approach can actually invoke a cynical attitude toward ethics and refusal to take the code of ethics serious.

While this criticism is warranted, one should be realistic about the target of an ethics code (Lere and Gaumnitz 2003). If the target is a culture of integrity inside the organization (as is implied by the sentencing guidelines), it is impossible to reach such a goal only through compliance (Cressy and Moore 1983). It demands leadership at the top and efforts by audit, legal, HR, and several other departments that are connected to company culture (Sims and Brinkmann 2002). Compliance programs are just one element in the building of organizational integrity. Far more

crucial, for instance, are incentive structures inside the organization. If commercial pressure is consistently high and reward is never linked to integrity indicators, one should not be surprised that compliance seems ineffective (Schwartz 2001). The danger today is that we end up with schizophrenic organizations with on the one hand increased control through ethics and compliance programs and on the other hand increased commercial pressure that drives ethical breaches. Such a schizophrenic organization simply breeds ethical cynicism toward the values and norms that the organization tries to push.

From this it follows that ethics and compliance programs should not be built nor judged in isolation (Collins 2012; Hoffman et al. 2001). They should be part of a general strategy that aims at corporate integrity (Chen and Soltes 2018). Today the success or failure of ethics and compliance programs is often measured in a very limited way. A survey by Deloitte in 2016 pointed out that the most common way is to measure completion rates of training programs and to deem training effective if enough employees – perhaps 90% or 95% – finish it (Deloitte 2017). Such an indicator mistakes legal accountability for compliance effectiveness. When the US Department of Justice published in 2017 an evaluation of corporate compliance in which this type of mistakes was pointed out, the document was immediately recuperated by business as a reference point that the justice department would from now on use to determine whether an ethics programs were effective (US DOJ 2017). It is another proof of the close interaction between regulation and the specific form of ethics and compliance programs.

Following a wave of accidents in the 1970s, the chemical sector launched the responsible care code and developed extensive health and safety programs. The impact of these programs has been clearly visible and measurable, for instance, in the number of workplace accidents. But this demanded a clear investment and consistent implementation with clear, measurable targets linked to financial incentives for the employee as well as for the company itself (e.g., the price of insurance was directly linked to accident rates). If codes of ethics want to have a clear impact on

corporate integrity, a similar road needs to be followed. Better measurement of ethics inside the organization, consistency with payment and promotion structures, pushing a speak-up culture in which employees are incentivized to speak about ethical issues, a demand to report ethical breaches under a no fault policy, follow up not just of things gone wrong but also of “near misses,” and effective hotlines are all measures that companies can take in order to improve the impact of ethics codes. The benefits would not only reach the company but also the society in general.

Cross-References

- [Audit Committees](#)
- [Codes of Conduct](#)
- [Corporate Criminal Liability](#)
- [Corruption](#)
- [Cost of Crime](#)
- [Organizational Liability](#)
- [Whistle-Blower Policy](#)

References

- Benson GC (1989) Codes of ethics. *J Bus Ethics* 8(5):305–319
- Brooks LJ (1989) Corporate codes of ethics. *J Bus Ethics* 8(2–3):117–129
- Canary HE, Jennings MM (2008) Principles and influence in codes of ethics: a centering resonance analysis comparing pre-and post-Sarbanes-Oxley codes of ethics. *J Bus Ethics* 80(2):263–278
- Carberry EJ, Engelen PJ, Van Essen M (2018) Which firms get punished for unethical behavior? Explaining variation in stock market reactions to corporate misconduct. *Bus Ethics Q* 28(2):119–151
- Chen H, Soltes E (2018) Why compliance programs fail – and how to fix them. *Harv Bus Rev* 2018:116–125
- Collins D (2012) *Business ethics: how to design and manage ethical organizations*. Wiley, Hoboken
- Cressey DR, Moore CA (1983) Managerial values and corporate codes of ethics. *Calif Manag Rev* 25(4):53–77
- Deloitte (2017) In focus: 2016 Compliance Trends Survey. Online available at <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/governance-risk-compliance/us-advisory-compliance-week-survey.pdf>. accessed on 14 June 2018.
- Farrell BJ, Cobbin DM, Farrell HM (2002) Codes of ethics: their evolution, development and other controversies. *J Manag Dev* 21(2):152–163
- Ferrell OC, LeClair DT, Ferrell L (1998) The federal sentencing guidelines for organizations: a framework for ethical compliance. *J Bus Ethics* 17(4):353–363
- Frankel MS (1989) Professional codes: why, who and with what impact? *J Bus Ethics* 8(2–3):109–115
- Gaumnitz BR, Lere JC (2002) Contents of codes of ethics of professional business organizations in the United States. *J Bus Ethics* 35(1):35–49
- Higgs-Kleyn N, Kapelianis D (1999) The role of professional codes in regarding ethical conduct. *J Bus Ethics* 19(4):363–374
- Hoffmann WM, Driscoll D, Painter-Morland M (2001) Integrating ethics. In: Moon C, Bonny C (eds) *Business ethics: facing up to the interests*. The Economist Book, London, pp 38–54
- Johnson & Johnson (1943) The credo. Online available at <https://www.jnj.com/credo>. accessed on 14 June 2018.
- Kukla R (2014) Living with pirates: common morality and embodied practice. *Camb Q Healthc Ethics* 23(1):75–85
- Kurland NB (1993) The defense industry initiative: ethics, self-regulation, and accountability. *J Bus Ethics* 12(2):137–145
- Lere JC, Gaumnitz BR (2003) The impact of codes of ethics on decision making: some insights from information economics. *J Bus Ethics* 48(4):365–379
- Munter D (2013) Codes of ethics in the light of fairness and harm. *Bus Ethics A Eur Rev* 22(2):174–188
- Painter-Morland M (2010) Questioning corporate codes of ethics. *Bus Ethics A Eur Rev* 19(3):265–279
- Rodriguez-Dominguez L, Gallego-Alvarez I, Garcia-Sanchez IM (2009) Corporate governance and codes of ethics. *J Bus Ethics* 90(2):187
- Schwartz M (2001) The nature of the relationship between corporate codes of ethics and behaviour. *J Bus Ethics* 32(3):247–262
- Schwartz MS (2004) Effective corporate codes of ethics: perceptions of code users. *J Bus Ethics* 55(4):321–341
- Shaplen R (1978) Annals of crime: the Lockheed incident. *New Yorker* 53:48–91
- Sims RR, Brinkmann J (2002) Leaders as moral role models: the case of John Gutfreund at Salomon brothers. *J Bus Ethics* 35(4):327–339
- Sims RR, Brinkmann J (2003) Enron ethics (or: culture matters more than codes). *J Bus Ethics* 45(3):243–256
- Stevens B (1994) An analysis of corporate ethical code studies: “where do we go from here?”. *J Bus Ethics* 13(1):63–69
- UBS (2010) UBS Dress Guide for Client Advisors, Switzerland. Online available at <https://static1.squarespace.com/static/55e0c62fe4b096b6619ff3d9/t/571bde1907eaa0d2558be4f6/1461444174809/Ubs+Dress+Code+English+2010.pdf>. accessed on 14 June 2018.
- US DOJ (2017) Evaluation of corporate compliance programs, US Department of Justice, Washington, DC. Online available at <https://www.justice.gov/criminal-fraud/page/file/937501/download>. accessed on 14 June 2018.

Cognitive Law and Economics

Angela Ambrosino¹ and Marco Novarese²

¹Department of Economics and Statistics Cognetti de Martiis, University of Turin, Torino, Italy

²Department of Law and Economics, University of Eastern Piedmont, Centre for Cognitive Economics, Alessandria, Italy

Definition

The tendency to consider the Behavioral Law and Economics and Cognitive Law and Economics as different sides of the same coin has been widespread inside the discipline. That was the consequence of a miscomprehension of what behavioral economics and cognitive economics are. These two research areas arise from a shared critique to standard neoclassical economics assumption of agents' perfect rationality and a common idea that economic agents, in the real world, are heterogeneous and more cognitive complex than what the theory assumed, but soon they diverge pursuing different goals and partially applying different research tools. Particularly BL&E is more concerned with what agents do, while CL&E is more about how agents think.

Hence we need a proper discussion of what Cognitive Law and Economics is as well as we need a proper definition of Behavioral Law and Economics.

Introduction

Do we really need an autonomous definition for Cognitive Law and Economics or it is the same of Behavioral Law and Economics? The tendency to consider the two approaches as different sides of the same coin has been widespread inside the discipline. That was the consequence of a miscomprehension of what behavioral economics and cognitive economics are. These two research areas arise from a shared critique to standard neoclassical economics assumption of agents' perfect rationality and a common idea that economic agents, in

the real world, are heterogeneous and more cognitive complex than what the theory assumed, but soon they diverge pursuing different goals and partially applying different research tools. Hence we need a proper discussion of what Cognitive Law and Economics is as well as we need a proper definition of Behavioral Law and Economics.

Other entries in this encyclopedia show how and when law meets economics (see Law and Economics or Behavioral Law and Economics or Nudge or Financial Education). When law scholars started applying the insights offered by neo-classical economics to their inquiry, the aim of this new approach to law was to develop both a positive and a normative theory of law on which to build efficient legal norms. Law and economics (L&E) uses economic models and econometric tools to develop its research in two ways:

1. Pursuing efficiency: efficiency is considered from two different points of view; on the one hand, it means that common law (judge-made law) is efficient, and on the other, from a normative point of view, it also means that law must be efficient.
2. Its emphasis on incentives and people's responses to those incentives.

L&E has been widely criticized (Ellickson 1989) in that applying economic tools is not sufficient to investigate the logic underlying the law and that the reductionist approach of economics cannot enable L&E to develop a proper positive theory of law and it excludes any consideration about justice.

L&E has been strongly influenced by the changes and debates that have characterized the development of economics since the middle of the last century (Rachlinski 2000). In recent years, the results obtained by the behavioral economics have given new emphasis to the first criticisms brought against law and economics. Behavioral economics shows that human behavior deviates from the perfect rationality assumption, and these deviations are not completely random, so it is possible to model and predict human behavior when it is affected by biases. During the 1990s, Jolls et al. (1998) investigate the opportunities offered by

behavioral economics to develop a new approach to law based on a more exhaustive theory of human behavior whereby better understanding of the foundations of individual behavior should strengthen both the descriptive power of models and their normative power. Their pioneering work gives rise to Behavioral Law and Economics (BL&E). During these same years, inside economics is developing another important research approach called cognitive economics (CI) (Bourguine and Nadal 2004). Cognitive economics shares with the behavioral approach the idea that human behavior is complex and that economic theory must ground its theories on a better understanding of cognitive decision-making processes. Cognitive economics retrieve the tradition of what Sent (2004) define “Old Behavioral Economics” that is the approach by Herbert Simon, instead that Kahneman’s.

Nevertheless, the two approaches follow (almost partially) different paths of inquiry. Cognitive economics puts itself in opposition to neoclassical economics investigating economic problems as complex phenomena. Its inquiry focuses on the analysis of the micro-foundations of human behavior and applies an interdisciplinary approach. Cognitive economics strongly criticizes the assumptions of standard economics and focus on the complexity of decision-making processes of heterogeneous agents. It questions the predictions of standard economics models and the rigidity of the formal tools applied. It is aimed at understanding decision-making processes, but it differs from behavioral economics, whose methodology is based on the analysis of the effectively exhibited behaviors. Cognitive economics’ central idea is that each phenomenon can be investigated with different tools and from different points of view. For example, cognitive economics investigates interdependent decisions using game theory not as a formal tool to predict specific outcomes but as a framework of analysis that allows investigating the complexity of agents’ decision-making processes (Schelling 1960); the outcomes of the game do not simply depend on strategies, but they are strongly linked to social context, path dependence dynamics, and focal points. Cognitive economics focus on norms and institutions (Rizzello and

Turvani 2000, 2002), but while law and economics has been much influenced by behavioral economics, the cognitive analysis of institutions has not been considered until recently.

Ambrosino (2016) shows two main explanations for this lack of interest in the cognitive theory of institutions:

1. The different concept of norms underlying the two research fields.
2. The cognitive theory of institutions is still far from developing a normative theory, and it focuses its inquiry on the positive level.

Nevertheless in the last few years, part of the literature points out the relevance of the analysis of the role of institutional forces and social norms in constraining and coordinating heterogeneous individuals, and cognitive economics and law and economics start to be connected and a new path of inquiry is arising.

The next sections are organized as follow: section “[Why Behavioral Law and Economics is not Cognitive Law and Economics](#)” explains why Cognitive Law and Economics (CL&E) is not the same as BL&E, particularly, “[Toward a Cognitive Approach to Law and Economics](#)” describes the main feature of CL&E, and “[Main Critiques to Behavioral Law and Economics](#)” focuses on the main critiques that such approach moves to behavioral law and economics. Section “[Toward a Cognitive Law and Economics Inquiry](#)” provides an example of how CL&E contributes to the inquiry into law.

Why Behavioral Law and Economics is not Cognitive Law and Economics

Toward a Cognitive Approach to Law and Economics

The cognitive theory of institutions is grounded on the idea that it is not possible to investigate the rise and evolution of institutions without investigating individual decision-making processes (North 2005). The institutional and the individual levels of analysis are interconnected, so that an institutional change may be the starting point for

modification of agents' behavior, and new cognitive classifications or new routines of behavior can engender a slow process of institutional change (Hayek 1982; Hodgson 2004; Ambrosino 2014). Cognitive theory of institutions assumes that agents are heterogeneous. Heterogeneity means that agents can exhibit different behaviors even if they belong to the same social and cultural context. That heterogeneity doesn't prevent coordination because agents are different, but they are made up of the same ingredients (Hayek 1982). Hence, they are able to understand each other, to build correct expectations about each other's behavior, and to share common social norms.

Recently such research filed shows points of contact with that part of the legal theory that firmly critiques BL&E. Such connection opens the door to a proper cognitive approach to L&E.

Particularly, Gregory Mitchell's main works seems to represent the main contribution to developing inquiry into the "individual-institution" framework already described by the cognitive theory of institutions (Hodgson 2004; Ambrosino 2014). Mitchell's critique of BL&E "provides reasons why legal theory should refrain from broad statements about the manner in which all legal actors process information, make judgments and reach decisions and why others should be skeptical of such broad claims by the legal decision theorists" (2002b, p. 33); "legal decision theorists should recognize the need for greater caution and precision in drawing of descriptive and prescriptive conclusions from empirical research on judgment and decision making" (2002b p. 32). Mitchell's contribution is based on a strong belief in the utility of psychological and other empirical research for legal analysis.

It emerges a new approach to law that shares with cognitive institutional economics the idea that agents are heterogeneous and that simply introducing the existence of "standard" biases in modeling human behavior does not enable the development of efficient predictive models; the perfect rationality assumption is not an appropriate instrument with which to investigate agents' behavior, and a proper theory of human behavior is needed. This approach suggests that the existence of cognitive biases in legal contexts must be

investigated in the field and with respect to specific contexts through "social facts studies" (Mitchell et al. 2011): a social facts study applies different research methods to explain case-specific descriptive or causal claims, and it is focused on the context-specific features of the case at hand. The analysis of how agents should behave cannot be separated from the investigation of the specific social context and cultural and social relations. A multidisciplinary approach is necessary to develop better inquiry into the complexity of decision-making processes in legal contexts. Legal theory, hence, moves toward a new approach, in which the cognitive determinants of agents' behavior are investigated; it highlights the importance of (i) agents' cognitive predispositions, (ii) learning processes and the influence of past experience, and (iii) the role of context. Moreover a cognitive inquiry into the diffusion of normative behavior and institutional change can furnish key into the opportunities offered by the development of prescriptive rules in shaping individual behavior. It emerges a new metacognitive approach to legal theory in which norms are concrete instruments with which to induce agents to develop different ways of processing information.

CL&E, following a social facts analysis, shows how to build appropriate decision tools based on objective casual claims. Scientific research results can be applied to normative purposes. They should constitute a sort of "social authority": an organizing principle for courts' or legislator' use of social science to create or modify a rule of law (Monahan et al. 2009). In the perspective of CL&E, social research and legal theory partially lose the need to furnish normative models. Producing case-specific evidence through reliable social science principles and methods, they become the research instruments that give judges and courts, and more generally the legislator, the information and the tools with which to evaluate and create new rules of law.

Main Critiques to Behavioral Law and Economics

Part of the literature inside legal theory criticizes BL&E both under a theoretical and a methodological point of view and points out relevant

elements of contact with cognitive economics that has opened the door to a new path of inquiry.

BL&E arise to pursue two main aims: first, explain why people do not act as they should in context of interest for legal theory (the benchmark being that agents should behave as the perfect rationality assumption expects), and second, bring people to act as they should proposing “a form of paternalism, libertarian in spirit, that should be acceptable to those who are firmly committed to freedom of choice on grounds of either autonomy or welfare” (Sunstein and Thaler 2003, p. 1160).

To pursue such aims, BL&E applies the tools and the insights furnished by behavioral economics. It is not surprising that BL&E today is exposed to quite the same critiques as behavioral economics (Ambrosino 2016).

The first critique to BL&E is strictly related to one of the cornerstone ideas inside B&E. It is a common opinion in B&E that it is possible to incorporate the complexity of the cognitive determinants of human behavior into the standard formal models of the neoclassical approach. The idea is that the assumption of perfect rationality can be replaced with a new concept of rationality – in which the existence of deviations from the perfect rationality assumption is explained by introducing new variables corresponding to particular biases assumed as commonly shared among agents – that better explains the complexity of real decision-making processes. Behavioral economics returns to being a research approach completely compatible with mainstream economics (Davis 2013). This tendency to build formal models has also taken place in the behavioral approach to L&E (Korobkin and Ulen 2000). The replacement of the perfect rationality assumption guarantees that BL&E models, compatible with the mainstream, produce strong normative outcomes. The first criticism to BL&E concerns the way in which scholars introduce into their inquiries insights drawn from the cognitive and psychosocial research of the past 30 years (Mitchell 2002a, 2002b, 2003a). BL&E grounds its research on the evidence of the existence of cognitive biases in human behavior and argues that such biases

are widespread in the population and are responsible for predictable and systematic errors (Korobkin and Ulen 2000). Nevertheless BL&E scholars fail in their attempt to criticize the perfect rationality assumption because they do not develop a new concept of rationality including the complexity of human decision-making processes. BL&E substitutes the neoclassical assumption of perfect rationality with an assumption of “equal incompetence” (Mitchell 2002a). This assumption is based on empirical research that shows homogeneous behavioral tendencies among agents. BL&E uses these behavioral tendencies to compile a list of common deviations from rationality that characterizes the entire population, and it develops normative models prescribing how agents have to behave and how decision-makers should intervene to shape agents’ behavior and avoid their errors. B&LE overlooks the substantial empirical evidence that people are not equally irrational and that human behavior is strongly influenced by situational variables: “The only way the lessons of behavioral decision research on bounded rationality can be manageably incorporated into behavioral models for use in the law is if these lessons apply widely and uniformly. If the rationality of behavior depends on particular characteristics of the legal actor or on even just a few characteristics of the situation at hand, then the development of behavioral models that are both realistic and predictive becomes enormously complex” (Mitchell 2002a p. 83). CL&E argues that BL&E do not understand that heuristic processing is only one mode of thought and that agents often do not act as expected, and it suggests the need for a legal theory focused on finding solutions to specific problems rather than on developing a general model of legal behavior. Heuristics can lead to favorable solutions but in many cases they can also give rise to errors. BL&E relies on the results obtained by behavioral research developed in other branches of economic theory and generalizes their significance. One of the main contributions is the pioneering work of Kahneman and Tversky (1974). These authors argue that their “studies on inductive reasoning have focused on systematic errors because they are diagnostic of

the heuristics that generally govern judgment and inference” (1974, p. 313). But this does not mean that the so-called K-T man can be reduced simply to the use of rules of thumb and heuristics in judgment. It seems an excessively simple explanation of human decision-making. “The likelihood that a particular decision or judgment will deviate from the ideal behavior derived from norms of rationality depends on a range of personal and situational factor. Even inside the relatively controlled environment of the laboratory, we see considerable variation in cognitive performance among individuals depending on their cognitive abilities, educational background, and affective state” (Mitchell 2002a, p. 109). CL&E suggests legal theory should not seek a general model of judgment and decision-making, but it should develop a contextualist approach that seeks to identify the conditions under which irrational behavior occurs. BL&E has important normative, methodological, and empirical limitations that prevent it from achieving descriptive and predictive accuracy. The libertarian paternalism suggesting that planners can improve social welfare by setting default rules that create benefits for those who commit errors but cause little or no harm to those who are fully rational (Sunstein and Thaler 2003) assumes the pervasiveness of irrational tendencies but ignores less invasive forms of intervention that may help agents overcome their errors without altering the substantive rights of the parties (Mitchell 2005). BL&E describing behavior as rational or irrational requires a normative standard against which the behavior may be judged (Mitchell 2003b). The behavioral approach assumes that rationality requires logical consistency and coherence in the formation and ordering of beliefs and preferences (Kahneman 1994; Simon 1997). Rationality as coherence operates as a closed system. Individual defines goals and beliefs and behavior must be logically consistent and coherent with respect to those goals and beliefs. In the case of legal judgment, when evidence of an irrational judgment is found, many different explanations are possible, some of which make the irrationality of the decision questionable (Mitchell 2003b). A behavior in a particular context may be at the same time rational and irrational

depending on the goals, the interpretation of the situation, and the tools used by any agent involved in the decision-making process.

The second main criticism concerns the methods employed to test for cognitive biases and errors (Mitchell 2002b, 2003b). BL&E research underestimates situational and individual variations in behavior and employs relatively weak tests of the hard-core assumptions of agents’ cognitive feature. The point is that the core of the research in heuristics and biases is based on statistical significance tests on experimentally generated and aggregate data. This body of research formulates in general terms the conditions under which events of various sorts occur and provides an interesting set of findings in general terms but with unspecified practical implications. Judgments are summarized by averaging across all the experimental subjects. That means that in BL&E analysis, if individual differences among judges emerge, these differences are treated as “errors,” and an “average judge” is considered the most meaningful summary of judges. This approach has the advantage of ensuring generalizability. Therefore, rather than examining individual variation in judgment and choice, behavioral decision theorists typically assume that “to a first approximation, the thought processes of most uninstitutionalized adults are quite similar, and any variation in subjects’ responses is attributed to measurement error or random variance” (Mitchell 2002b, p. 46). The rigor of experimental research is purchased at the price of generalizability of results, and this trade-off operates most directly in those fields that use laboratory experiments to study how humans navigate complex social environments like BL&E. Such critique is strongly related to the debate emerged in psychology about the danger of relying on “statistical significance” as a measure of behavioral tendencies. Scientists (and journals) publish studies and analyses that “work” and place those that do not in the file drawer (Rosenthal 1979). One answer to this problem of publication bias is that we can trust a result if it is supported by many different studies. But this argument breaks down if scientists exploit ambiguity in order to obtain statistically significant results (Simmons et al. 2011).

Toward a Cognitive Law and Economics Inquiry

Hence Cognitive Law and Economics is aimed at developing a legal theory in which the peculiarity of decision-making in legal contexts can be really explained. The critique of the equal incompetence assumption suggests the need for a new analysis in which heterogeneous agents are considered (Mitchell 2002a, b, 2003a, b).

Evidence on cognitive biases must be investigated in legal contexts so as to build an original and consistent map of evidence. CL&E aspires to develop a contextualist approach. A contextualized approach acknowledges that features of the person, the situation, and the task have an impact on the nature and quality of judgment.

Experiments are only one of the tools that should be applied to examine variations in individual behavior. The need for an interdisciplinary approach arises from the recognition that multiple forces combine to produce particular behaviors. The cognitive theory of institutions has yet developed interesting inquiries into coordination processes (Schelling 1960) and into the relevance of learning in the process through which people conform to social or formal rules.

More recently, an example of the kind of inquiry CL&E can develop is given by Mitchell (2009) idea of a metacognitive approach to regulation. Such approach is based on his discussion about the role of second-level thought in shaping human behavior. BL&E describes judgment as the product of a non-deliberative thought process based on cognitive heuristics and rules of thumb. Psychological models of actors developed inside BL&E show that biases in judgment and errors often arise at the level of first-order thoughts; thoughts occur at the direct level of cognition and are not intentional and not deliberative. These models assume that agents are incapable of going beyond these first-order thoughts and that this is the cause of irrational and discriminatory behavior. This emphasizes the role of automatic and intuitive thoughts while neglecting the role played by controlled and deliberative thoughts. It leaves no room for self-correction, arguing that individuals lack self-awareness of

their biases, and it ignores the substantial evidence that agents learn through experience. Second thoughts may be the products of conscious effort, but they may also be automatic corrections working at the unconscious level. The propensity to engage in self-correction varies among persons and situations, but all cognitively normal people are able to engage in some amount of “metacognition” about their own thoughts (Loires 1998). People may differ in their propensity for such reflection depending on their education, upbringing, values, or genetic endowment, but everyone possesses some level of ability in rethinking their own thoughts.

Regulation should take it into consideration. If second thoughts apply, law will not simply change the prices of different behaviors for the purposes of a rational analysis of the costs and benefits of different courses of action. Rather, law will focus on altering the ways in which agent processes information. Under this point of view, law is a system of second thoughts that functions both consciously and unconsciously. Hence, law can contribute to influencing thoughts and behaviors in legal contexts. Mitchell provides concrete applications of his theory of law. The author (Monahan et al. 2009; Mitchell 2010; Mitchell et al. 2011) enters the debate on the proper scope of expert witness testimony that purports to summarize general social science evidence to provide context for the fact-finder to decide case-specific questions. Mitchell’s analysis focuses on the *Dukes v. Wal-Mart* case on gender discrimination toward female employees. Dukes’ plaintiffs submitted expert statistical evidence showing that female employees were faring worse in the aggregate than male employees, and a report by a social science expert identified a common source of this discrimination across all Wal-Mart facilities (Mitchell 2010, p. 136). The social science expert based his report on the “social framework analysis” method (Fiske and Borgida 1999). This method consists in using social science research as a framework for analyzing the facts of a particular case. The reliability of such analysis is based on the reliability of the research on which the general conclusions applied to the case at hand are based. In *Dukes v. Wal-Mart*, the expert

summarized research on gender bias, organizational culture, and anti-discrimination measures and applied it to interpret the facts in the discovery material supporting the claims of the Dukes plaintiffs. Mitchell argues that testimony based on that social framework analysis should be restrained from making any linkage between general social science research findings and specific case questions. In the specific case of *Dukes v. Wal-Mart*, he based his critique on two main points: (1) in social framework analysis, experts use their personal judgment rather than scientific method to link social science to specific cases; in some sense, social framework analysis make the same mistake that BL&E does in extending the experimental economics results to its research purposes without dealing with context-specific research. (2) The expert corroborated his report with statistical evidence. But the statistical evidence was itself subject to dispute with regard to the proper unit of analysis. The plaintiffs argued for an aggregate-data approach. This choice did not allow consideration of context-specific differences due to store-by-store variation in male-female outcomes and to local control over personnel matters. This use of statistical evidence is an example of how statistical results can vary depending on the many decisions that researchers have to make while collecting and analyzing data (which outliers to exclude, which measures to analyze, and so on). Mitchell argues that there are social science techniques and methods that allow development of opinions about the parties or behaviors involved in a particular case; such evidence has been referred to as “social facts” (Mitchell et al. 2011). Social facts are special types of adjudicative facts produced by applying social science techniques to case-specific data in order to help prove some issue in the case. A wide variety of social science methods can be used to produce social facts. The design of a social fact study depends on what a party hopes to learn. Mitchell divides the search for social facts according to three main goals:

1. Obtaining descriptive information: getting the facts right is important, but doing so can be difficult when the relevant facts are in the possession of a large number of nonparties.

2. Obtaining explanatory information: gain a better understanding of the issue in a case. Many research methods can be applied, such as interview, survey, observational study, and experimental simulation.
3. Testing specific hypotheses: the ideal way to test causal hypotheses is through the use of experiments in which participants’ behaviors are recorded to assess how changes in the experimental conditions affect the behavior in question (Mitchell et al. 2011).

Social facts constructed by a proper scientific method possess scientific reliability and fit the facts of a particular case. Such reliability depends on the reliability of the scientific method applied. Mitchell shows that when addressing such a complex task as deciding a legal dispute, it is necessary to rely on rigorous interdisciplinary research tools that help prove some issue in the case.

CLE remains a very recent research project; its finding can be still considered a preliminary attempt to develop a proper interdisciplinary inquiry to law and economics. Moreover this approach is still mainly focused on a positive ground. As shown in this section, CL&E is a very relevant and promising research field.

Cross-References

- [Behavioral law and economics](#)
- [Law and Economics](#)
- [Nudge](#)

References

- Ambrosino A (2014) A cognitive approach to law and economics: Hayek’s legacy. *J Econ Issues* 48(1):19–49
- Ambrosino A (2016) Heterogeneity and law: toward a cognitive legal theory. *J Inst Econ* 12(2):417–442
- Bourgine P, Nadal JP (eds) (2004) *Cognitive economics: an interdisciplinary approach*. Springer, London
- Davis JB (2013) Economics imperialism under the impact of psychology: the case of behavioral development economics. *Oeconomia* 1:119–138
- Ellickson RC (1989) Bringing culture and human frailty to rational actors: a critique of classical law and economics. *Chicago Kent Law Rev* 65:23–55

- Fiske ST, Borgida E (1999) Social framework analysis as expert testimony in sexual harassment suits. In Estreicher S (ed.) *Sexual harassment in the workplace: proceedings of New York University 51st annual conference on labor*. New York, pp 575–577
- Hayek FA (1982) *Law, legislation and liberty*. Routledge, London
- Hodgson GM (2004) Reclaiming habit for institutional economics. *J Econ Psychol* 25:651–660
- Jolls C, Sunstein CR, Thaler R (1998) A behavioral approach to law and economics. *Stanford Law Rev* 50:1471–1552
- Kahneman D (1994) New challenges to the rationality assumption. *J Inst Theor Econ* 150:18–36
- Kahneman D, Tversky A (1974) Judgment under uncertainty: heuristics and bias. *Science* 185:1124–1131
- Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. *Calif Law Rev* 88(4):1051–1144
- Loires G (1998) From social cognition to metacognition. In: Yzerbyt VY, Loires G, Dardenne B (eds) *Metacognition: cognitive and social dimensions*. Sage, London, pp 1–15
- Mitchell G (2002a) Why law and economics' perfect rationality should not be traded for behavioral law and economics' equal incompetence. *Georgetown Law J* 91:67–167
- Mitchell G (2002b) Thinking behavioralism too seriously? The unwarranted pessimism of the new behavioral analysis of law. *William Mery Law Rev* 43:1907–2021
- Mitchell G (2003a) Tendencies versus boundaries: levels of generality in behavioral law and economics. *Vanderbilt Law Rev* 56:1781–1812
- Mitchell G (2003b) Mapping evidence law. *Michigan State Law Review*, 1065–1148
- Mitchell G (2005) Libertarian paternalism is an Oxymoron. *Northwest Univ Law Rev* 99(3):1245–1277
- Mitchell G (2009) Second thoughts. *McGeorge Law Rev* 40:687–722
- Mitchell G (2010) Good causes and bad science. *Vanderbilt Law Rev Banc Roundtable* 63:133–147
- Mitchell G, Monahan L, Walker L (2011) Case-specific sociological inference: meta-norms for expert opinions. *Sociol Methods Res* 40:668–680
- Monahan J, Walker L, Mitchell G (2009) The limits of social framework evidence. *Law Probab Risk* 8(4): 307–321
- North D (2005) *Understanding the process of economic change*. Princeton University Press, Princeton
- Rachlinski JJ (2000) The “New” law and psychology: a reply to critics, skeptics, and cautious supporters. *Cornell Rev* 85:739–766
- Rizzello S, Turvani M (2000) Institution meet mind: the way out of an impasse. *Constit Polit Econ* 11:165–180
- Rizzello S, Turvani M (2002) Subjective diversity and social learning: a cognitive perspective for understanding institutional behavior. *Constit Polit Econ* 13:201–214
- Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychol Bull* 86:638–641
- Schelling TC (1960) *The strategy of conflict*. Harvard University Press, Cambridge, MA
- Sent EM (2004) Behavioral economics: how psychology made its (limited) way back into economics. *Hist Polit Econ* 36:735–760er
- Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366
- Simon H (1997) *Models of bounded rationality*. MIT Press, Cambridge, MA
- Sunstein C, Thaler R (2003) Libertarian paternalism is not an Oxymoron. *Univ Chicago Law Rev* 70:1159–1202

Coherence

► Rationality

Collective Redress

► Class Action and Aggregate Litigations

Commercial Benefits

► Preferential Tariffs

Commercial Concessions

► Preferential Tariffs

Commercial Sex and Health

Giovanni Immordino and Francesco F. Russo
Department of Economics and Statistics,
University of Naples Federico II and CSEF,
Naples, Italy

Abstract

The spread of HIV/AIDS and of other sexually transmittable infections is affected by the legal regime governing prostitution. In this entry we briefly discuss a theory of commercial sex, public intervention, and the health consequences. We present some evidence and the

main motivations for unprotected commercial sex in addition to examining the theoretical and empirical effects of different policies for the health consequences of commercial sex.

Definition

Link between the legal regime governing commercial sex and health.

Introduction

The spread of HIV/AIDS and of other sexually transmittable infections (STI henceforth) is an important negative externality of commercial sex. The feasibility and effectiveness of public intervention differ substantially depending upon the legal regime in place. The reason for this is that different policies affect the endogenous evolution of the risks in the market for prostitution services differently. Any policy intervention will change the identity of the clients and of the sex workers who decide to join, or not join, the market, and it will also have an impact on the decision regarding which market to join (legal or illegal). This will endogenously change the risks associated with the markets, thus further influencing demand and supply, and consequently the identities, and so on. This is the reason why even well-meant interventions can lead to unintended consequences which not only reduce the effectiveness of the intervention but sometimes even make things worse.

Finally, we should be aware that legal regime governing prostitution might have other consequences besides affecting public health. For instance, Immordino and Russo (2015b) show that if prostitution is legal or regulated, individuals tend to justify it significantly more than if it is prohibited.

A General Theory of Commercial Sex and Health Consequences

Immordino and Russo (2015a) look at a specific negative externality associated with prostitution,

the spread of HIV/AIDS and of other STIs, from a general and theoretical point of view. In particular, they set up an equilibrium model of prostitution and calibrate it to real-world data to perform a quantitative policy analysis exercise. They model demand and supply of prostitution, given the policies implemented by the government, as a function of the risk of infection, of the health status, and of the outside earning opportunities for both clients and sex workers. They consider the policies within three different regimes. In a *laissez-faire* regime, prostitution is legal and unlicensed and no policy is implemented. In a *prohibition* regime, prostitution is illegal and the enforcement consists of random audits and fines. Under *regulation*, prostitution is legal, conditional on compliance with a series of requirements such as tax payments, entry fees, or participation restrictions. Two markets coexist in this regime: a legal market, where all the agents comply with the regulations, and an illegal market where they do not. The enforcement against the illegal sector consists of random audits and fines. Key features of this model are the interaction between the demand and supply of legal and illegal prostitution and the endogenous evolution of the risks: any policy intervention influences the characteristics of clients and sex workers that join the two markets and, therefore, the risks associated with buying and supplying prostitution, further influencing demand and supply and so on. For instance, introducing mandatory health checks for sex workers decreases the probability of infection for customers, thereby attracting more demand for prostitution by healthy customers. This in turn decreases the average risk of infection for sex workers, thereby inducing more healthy individuals to join the supply side of the prostitution market and so on.

Motivations for Unprotected Commercial Sex

Unprotected sex increases the risk of HIV/AIDS and the transmission of other STIs. It is therefore extremely important to understand why sex

workers and clients might engage in unprotected sex. Concerning sex workers, the main reasons are lack of awareness about HIV/AIDS and about safe sex practices (Bhave et al. 1995), the high cost and poor availability of condoms, violent practices by clients who force them to have unprotected sex, and extra compensation that some customers are willing to pay for unprotected sex. With respect to clients, the main reasons are intrinsic preference not to use condoms and a lack of information about the risks.

Bhave et al. (1995), in a study of Bombay sex workers, show that monetary incentives are the most important factor deterring sex workers from using condoms, even for those who are aware of the risks of unprotected sex and of the preventive role of condoms. The importance of monetary incentives is also documented in two studies that use micro data to quantify the risk compensation required by sex workers for not using condoms. The conclusion of both studies is that sex workers accept the extra risk because clients are willing to pay them very well for it (Rao et al. 2003; Gertler et al. 2005). Specifically, Rao et al. (2003) studied a random sample of commercial sex workers from the red-light area of Sonagachi in Calcutta. Using price data for transactions with and without condoms, they were able to estimate the compensating differential for safe sex. They found a loss in average prices between 66% and 79% in the case of condom use. Similarly Gertler et al. (2005), using data from Mexico, estimate a 23% premium for unprotected sex (46% in case of a “very attractive” sex worker). Overall, the strong monetary disincentive for safe sex practices can have a substantial adverse impact on preventing the spread of HIV/AIDS and of other STIs.

Moving to the demand side of the market, Cameron and Collins (2003) use UK micro data to identify the male clients’ characteristics that influence their demand for sex services. They find, among other things, that a high risk of HIV/AIDS infection has a negative impact on sex service usage. Their results show support for the view of “the man who pays for sex as a rational economic actor,” suggesting that interventions on the side of the client might be effective.

Effects of Public Interventions: Theory and Evidence

The conventional policy recommendations to reduce unprotected sex are to intervene on the supply side of the market by implementing educational campaigns, improving access to inexpensive condoms, enforcing laws against human trafficking and rape, etc. Moreover, since an important motivation for unprotected commercial sex is the clients’ willingness to pay, complementary interventions on the demand side are also necessary in order to increase condom use.

However, the legal regime in place matters a great deal with respect to the effectiveness of different policies on the health consequences of commercial sex. Immordino and Russo (2015a) show theoretically that in a *prohibition* regime, targeting enforcement at customers only has a worse outcome on reducing the negative health externality than targeting enforcement at sex workers only. They also show that, in a *regulation* regime, taxation can increase harm. The reason is that higher taxes, by raising the cost of operating legally, decrease the number of healthy individuals that join the legal market, which is, therefore, smaller but riskier. Conversely, increased enforcement against the illegal market increases the equilibrium quantity in the legal market, also through a decrease in risk. Establishing a licensing system that prevents risky individuals from joining the legal market, on the other hand, reduces the risk and, therefore, increases the demand for legal prostitution. The impact on harm is theoretically ambiguous, but, for many model parameterizations, the decreased risk prevails so that harm also decreases. Finally, in comparing regulation and prohibition, they find that regulation is better for harm minimization, while the *laissez-faire* regime is always dominated (see also Immordino and Russo 2016).

Gertler and Shah (2011) studied empirically the effects of different policies in Ecuador, where a *regulation* regime prevails. Prostitution is legal, but sex workers must obtain and periodically renew a license at their own expense. This license, among other requirements, is conditional on health status and, in particular, on being free

from STIs. Licensed prostitutes can solicit in nightclubs and bars or in the streets, and typically, licensed and unlicensed prostitutes share the workplace in both nightclubs and in the streets. The authors exploit a nationwide dataset that combines sociodemographic, economic, and health information for the sex workers and regional, within-country variation of police enforcement against unlicensed prostitutes in the streets and in bars. They find that increased enforcement against unlicensed street prostitution is associated with a smaller diffusion of STIs among sex workers, which is in line with the theoretical results in Immordino and Russo (2015a). Gertler and Shah (2011) also show that public intervention can lead to unintended consequences. Indeed, enforcement against *unlicensed street prostitution* pushes some sex workers to move to the bars where there is a lower demand for unprotected sex and lower STI rates among clients, with the result that the overall risk of infection decreases. However, increased enforcement against *unlicensed prostitution in bars and nightclubs* is associated with higher infection rates. The reason is that an increase in enforcement raises the cost of being a sex worker without a license in a bar with respect to both complying with the license and with being unlicensed on the streets. Therefore, the question is if, in response to a tighter enforcement, the unlicensed sex workers will choose to obtain and renew the license or to move to the streets where unprotected sex is more common and STIs among clients more frequent. Gertler and Shah (2011) show that migration toward the street sector prevails, with an overall increase in the infection rate.

Cross-References

- [Crime and Punishment \(Becker 1968\)](#)
- [Liberty](#)
- [Pornography](#)
- [Prostitution, Demand and Supply of](#)
- [Sex](#)
- [Sex Offenses](#)
- [Shadow Economy](#)
- [Underground Economy](#)

References

- Bhave G, Lindan CP, Hudes ES, Desai S, Wagle U, Tripathi SP, Mandel JS (1995) Impact of and intervention on HIV, sexually transmitted diseases, and condom use among sex workers in Bombay, India. *AIDS* 9(Suppl 1):S21–S30
- Cameron S, Collins A (2003) Estimates of a model of male participation in the market for female heterosexual prostitution services. *Eur J Law Econ* 16: 271–288
- Gertler P, Shah M (2011) Sex work and infection: what's law enforcement got to do with it? *J Law Econ* 54(4): 811–840
- Gertler P, Shah M, Bertozzi SM (2005) Risky business: the market for unprotected prostitution. *J Polit Econ* 113: 518–550
- Immordino G, Russo FF (2015a) Regulating prostitution: a health risk approach. *J Public Econ* 121:14–13
- Immordino G, Russo FF (2015b) Laws and stigma: the case of prostitution. *Eur J Law Econ* 40: 209–223
- Immordino G, Russo FF (2016) Optimal prostitution policy. In: Cunningham S, Shah M (eds) *Handbook of the economics of prostitution*. Oxford University Press, Oxford, pp 332–347
- Rao V, Gupta I, Lokshin M, Jana S (2003) Sex workers and the cost of safe sex: the compensating differential for condom use among Calcutta prostitutes. *J Dev Econ* 71:585–603

Common Law System

Cristina Costantini

Department of Law, University of Perugia,
Perugia, Italy

Abstract

Moving from a clear definition of the diverse and concurring meanings attributed to the syntagma “Common Law,” the essay rediscovers the genealogical construction of the English legal tradition. A particular emphasis is given to the relationship between auxiliary or competing jurisdictions and, with the proper methodological approach of comparative law, to the main traits, which shape legal mentality and legal style in a different way from the canonical morphology of the Civil Law tradition.

Defining the Expression “Common Law”

The expression “Common Law” has been used with various extensions and with different, even if interrelated, meanings.

First of all, it denotes the constructive result of the act of mapping or splitting the worldwide *nomos* (the conceptual space of the normative globe) into definite families and systems, each of them identified by its proper genealogy and its governing principles. In this perspective, “Common Law” is counterposed to “Civil Law,” the first denoting the legal model which flourished in the United Kingdom (except Scotland), later transplanted into the United States of America and into the Commonwealth countries with diverse extent of local adaptation; the second naming the legal order which embraced Continental Europe. The spatial dichotomy traces its origins back to the dissimilar relation entertained by the aforesaid legal experiences with a common denominator, the Roman Law. While the Civil Law systems developed from the conscious recovery and the strategic interpretation of the stratified precepts condensed in Justinian’s *Corpus Juris Civilis*, the Common Law sapiently exploited the inner potentialities offered by feudal law to resist against the hegemonic expansion of Roman texts and culture. Therefore, the Civil Law family founded its prestige on the venerable authority of Roman maxims; on the contrary, the Common Law family exalted its own uniqueness and insular authenticity by the means of a fierce immunity from Roman fascination. The relics of the ancient bipartition are kept by the other nomenclature conventionally employed to juxtapose these legal entities: on the one side, it evoked the Anglo-American legal family, which retains in the name the pride of its autochthon sources; on the other side, it made manifest the Romano-Germanic legal family, which exhibits in the epithet the satisfaction for such an illustrious descent.

Historically, the opposite recall to Roman Law molded the conceptual structure of the systems in dissimilar ways; shaped otherwise their methodological outlooks, respectively based on the assumption (in the Civil Law family) and on the rejection (in the Common law family) of the

codification of laws; and justified heterogeneous techniques of legal education and legal reasoning. On this ground Common Law and Civil Law came to different choices both at an ontological and epistemological level.

In a second and more specific meaning, the expression “Common Law” is closely linked to the notion of *Equity*, in order to denote the complex structure of the English Legal system, the double soul, which gave substance and form to English Law wholly considered. In particular, “Common Law” and “Equity” designate two different bodies of rules, principles, and remedies, originally settled and administered by two separate jurisdictions, respectively by the *Curia Regis*, with its internal partitions, and by the Chancellor and the Court of Chancery.

In a third sense, “Common Law” is contrasted or opposed to “statutory law,” with the aim of stressing the diverse sources of legal rules. While the Common Law is the case law, that is, the law declared or created by the Courts, statutory law is the ensemble of written laws passed by the legislature or other government agency and expressed with the requisite formalities.

The Genealogy of the English Common Law: The Jurisdictional Project

Two main visions compete for the genealogical reconstruction of the Common Law identity: one corresponds to a mythological thought supporting a legitimating project; the other represents the constitutive framework conventionally used to emplot the narratological history of English system, where a proper emphasis is given to the elements which built the specificity of Common Law and measured the distance from Continental Law.

In the first perspective, the Common Law presented itself as an organic bulk of memories and customs inherited from a “time out of mind” and handed down without solution of continuity. The immemorial law has the nature of an unwritten tradition, nourished by a tacit and original knowledge, that is close to nature and to divine law (Fortescue and Chrimes 1942; Goodrich 1990).

Therefore, the superiority of the English laws was validated even by theological and philosophical arguments.

In a different manner, the orthodox account depicts the Norman Conquest of 1066 as a catastrophe irrupted into the history of England, a revolutionary event which impressed the proper direction to the future course of English Law. The pristine foundations of the Common Law are underestimated, if not concealed, and William the Conqueror, Duke of Normandy, is considered as the illuminate king responsible for the creation of the political and social context in which the principal institutions of a distinguished system could be developed (Baker 2000, 2002; Milsom 1969; Plucknett 1956).

First of all, William I imposed a particular form of feudal structure, based on a sapiently articulated association of lordship and tenures and influenced by the very conception of liegeancy. The King was proclaimed as the supreme Lord of the landed property, and as a consequence, this assertion prevented the introduction of allodial estates, over which one could exercise full and unrestricted ownership. At that time the European *nomos* was split into two opposite models, into two ideological visions of government and society competing with each other: on the one side, there was the Roman model with its clear demarcation between the concepts of *imperium* (public power) and *dominium* (private ownership); on the other, there was the feudal model with its confusing mixture of the public and the private (Samuel 2013).

Another relevant aspect of William's policy was administrative centralization, also extended to law. Enhancing the medieval conception of sovereignty, the King was the pinnacle atop the feudal hierarchy and retained both legislative power and jurisdictional capacity: he was Lord Tenant in Chief and, metaphorically, the fountain of justice, supervising the declaration of new rules and the dispensation of new remedies. The English Common Law could not originate and exist, as we actually know it, if a corps of advisers and courtiers, named *Curia Regis*, wasn't established in all its functions. This council embodied the center of royal administration:

initially it was undivided and peripatetic, following the King during his itinerant circuits across the realm; later it was transformed into a fragmented and mainly stable organ, which began to sit regularly at Westminster. The internal evolution of the *Curia Regis* led to the formation of the three royal courts which molded the forms and the contents of a law common throughout the kingdom.

The Exchequer was the first department to be deposited, dealing with finance and taxation. Another division was settled to examine the petitions that affected the King's interests, because of their nature heard and discussed *Coram Rege*, at the presence of the Monarch. It was the King's Bench with jurisdiction over issues recognized as "public" in their orientation, such as those nowadays included within criminal and administrative law. In order to grant a secure and constant justice to all the litigants, a historical compromise was reached in 1215, inserting in the text of Magna Carta a specific clause, according to which the "common pleas" – that is to say all the suits in which the King had no interest – would be heard by a permanent body of judges set in Westminster. As a corollary, a new court originated, which decided not *coram rege* but *in banco*, known as the *Common Bench* or as the *Court of Common Pleas* (Brand 1992).

By the fourteenth up to the seventeenth century, the Common Law was a matter of three royal courts competing for litigation (Samuel 2013). The English Common Law developed and flourished as a jurisdictional project.

The consciousness of unicity and originality that marked the Common Law Tradition was supported even by the specificity of procedural technicalities required for the effective functioning of the Royal Courts.

The first characteristic was the systems of writs (Maitland et al. 1936). The plaintiff who wished to start a lawsuit before one of the Royal Courts had to purchase a writ from the Chancery section of *Curia Regis*, a sort of granted permission authorizing the commencement of the proceedings. This was due to the fact that the Courts in Westminster, before becoming the ordinary and regular courts of law, had an exceptional jurisdiction, allowed by royal favor; as a consequence, litigants

formally had not the right to go to the royal courts, but they needed a sort of pass to benefit from that kind of justice for which they had paid. In its materiality the writ was a strip of parchment usually written in Latin and sealed with the Great Seal; in its juridical consistency it was an order to do justice based on a definite and compelling formula, composed by selected words, apt both to introduce a particular type of action and to settle a likewise specific procedure. From the beginning of the thirteenth century they were collected and reported in a proper book entitled the *Register Brevium*. By the middle of the same century, as a result of their exponential growth, a political and juridical contention arose among the King, the Feudal Lords, and the common law judges. The King claimed to preserve the reserve of justice he bore by virtue of his paramount sovereignty; the Feudal Lords craved to avoid the infringement of their signorial jurisdiction; the Common Law judges aspired to strengthen their authority even though they were overwhelmed by litigation. A temporary composition of these tensions was reached in 1258 with the Provisions of Oxford, which fixed the orthodoxy of the common law system, insofar as no unprecedented writs could be issued by Chancery without the consent of the King's Council. This form of institutional conciliation generated order and certainty but imposed an asphyxial fixity, a stagnant and bogged immobility. The subsequent Statute of Westminster II allowed a certain, albeit limited, openness through the recognition of "*writs in consimili casu*," so to make justiciable all those instances presenting a great similarity with others for which the writ was already dispensed. It should be emphasized that the choice of the correct writ was not a mere formality. First of all, an improper selection among the suitable writs undermined the whole procedure, while the absence of an appropriate writ was equivalent to the absence of a legal remedy; secondly, the internal classification of writs was at the basis of the legal taxonomy of actions and, in course of time, represented the structural framework of the substantive outlines of the common law.

Moreover, from the origins onward, the system of writs fashioned the common law mentality in a

very different manner from the civil law way of reasoning and arguing. The common lawyers privileged analogy and factual appreciation, since the concrete facts of a dispute had to be compared with the models of factual situations already sanctioned in the *Register Brevium*. The civil lawyers privileged logic deduction, since the proper solution of a juridical issue had to be derived from a coherent complex of superior and outstanding principles.

The second and most important procedural feature was the presence of the jury in the course of the ordinary trial. Firstly used in criminal procedure as a means of evidence, the jury was subsequently imported into private lawsuits as a sworn body of lay people convened to render an impartial verdict and to judge on the facts of the controversy.

Common Law and Equity: A Mutable Relationship

The growth of the Common Law marked the victory of a centralized authority over a cluster of diffuse and competing powers. Nevertheless the balance achieved among the King, the Feudal Lords, and the Common Lawyers ultimately sacrificed the creative progress of the system of law. The English Common law froze and arrested; historically it was not too far from a shocking paralysis. The apparent involution was also made worse by the emergence of grave defects which undermined the efficiency of the Royal Courts and seriously compromised their popularity. Moreover, the set of remedies created and administered by the central Courts became progressively inadequate and irrespective of the new or increased exigencies. In particular, the common law courts were not technically equipped to grant personal remedies, that is, to order a party to do or not to do something, the only remedies dispensed being monetary remedies (debt or damages). On the other hand, if the facts of the claim could not be adapted to the formula of an established writ, the search for justice was utterly denied and the trial could not commence. The lack of a proper system of appeal courts and a mounting set of

judicial vices (among which corruption and delay) completed a not comforting framework.

However, in spite of everything, the English law was able to replenish itself from the inside. The most urgent need was to overcome the strictness of the formulary system, which could preclude the effective access to the common law courts. The historical escape resided in the overriding and residuary power of the king to dispense justice outside the regular system. By the end of the thirteenth century and during the first half of the fourteenth, the unsatisfied claimant came to address a specific petition (bill) to the King in piteous terms, asking for his grace and mercy to make manifest in respect of some complaint. Initially the King examined these formal requests directly, by himself, but due to the considerable increase of their number in times, he began to pass the petitions to the Chancellor. As it has brilliantly pointed out, the Chancellor was in direct connection with all the parts of the constitution (Holdsworth 1966). He was the secretary of state of all departments; to him was entrusted the Great Seal by which all the acts of state and royal commands are authenticated; as the head of the Chancery, formerly conceived as an administrative department, he had also the power to draw and seal the same royal writs necessary to start legal proceedings before the Courts of Common Law. Moreover, most early Chancellors were ecclesiastics, keepers of the King's conscience. When the petitions were passed from the King to the Chancellor, he formerly decided on behalf of the monarch, admitting "merciful exceptions" to general law with the aim to ensure that the King's conscience was right before God (Watt 2008). By the end of the fourteenth century, the Chancellor decided in his own name and on his own authority. Moving from his ecclesiastic affiliation, the Chancellor exercised a transcendent form of justice, beyond the common law jurisdiction and based on an innovative mixture of canon precepts, Christian discretion, and Roman Law. The procedure was different if compared with that enacted by the Royal Courts of Westminster and integrally devoted to amend the presumably guilty conduct of the defendant. In particular, all the actions were commenced by an informal complaint, in order to

make unnecessary the selection of a correct writ; the pleading was conducted in English (not in Latin, nor in French); there was no jury; the final judgment was expressed in the form of a decree, more precisely in a decree of injunction or in a decree of specific performance; the Chancellor decided questions of fact as well as issues of law. One of the most important features of the new procedure was the proper form of the act whereby it starts, called *writ of subpoena*, which, in spite of its name, was something other than the old writs enacted in Common Law Courts. While the latter explicitly mentioned the cause of action against the defendant, the writ of subpoena was limited to command the physical presence of the litigant before the Chancellor – upon pain of forfeiting a sum of money – in order to answer to the complaints made against him by the plaintiff, without revealing the specific reasons led at the basis of the claim.

This was the process which led to the formation of a separate corpus of principles, remedies, and rules, autonomous from the common law, strictly considered. A line of demarcation crossed the former and undivided space of jurisdiction: the proper domain of the Courts of Westminster was disjointed from the sphere of the Chancellor, newly rediscovered (Maitland et al. 2011).

From a constitutional point of view, the same process made complex the originally unitary vision both of the Chancellor and of the so-called Curia Cancellariae, the staff of clerks over which he presided. In course of time, it was possible to recognize two sides of each of the aforesaid bodies. Looking at the figure of the Chancellor, on the one hand he had the power to seal the writs needed to bring in court a common law action, but in this guise he didn't act as a judge, he didn't hear the polemical arguments of the parties, insofar as he simply granted a writ on the basis of the plaintiff's claim, leaving to the three Royal Courts the decision on the conformity of the writ to the law of the land. On the other hand, the Chancellor gradually exercised an extraordinary form of justice, not to supersede or to contradict the principles expressed within the boundaries of the common law jurisdiction, but to mitigate the excessive rigor of the Courts and to adequate remedies to the new substantial needs.

Conversely, looking at the “Curia Cancellariae,” it was possible to detach the growing of a Common Law side and an Equity side (as it was called at the end of the formative period), or a Latin side and an English side, giving relevance to the official language used in the course of the different procedures. The Latin side was requested to examine that petition which concerned the person of the King, whenever justice was demanded against the sovereign. The proceedings were enrolled in Latin and developed in a very similar manner to that followed in the three courts of law. The English side gave rise to the new form of justice, we have discussed above, originally perceived as ancillary to the common law system.

The history of this other source of English legal system was marked by progressive metamorphoses, which affected both its nature, function, and appraisal and the subtle relationship with the common law, properly considered. As a result, the whole morphology of English law took different appearances across epochs and times.

In the first direction, it's possible to detach a transformation with respect to the same structure of the organ that administered the new jurisdiction: gradually the Lord Chancellor ceased to be an individual taking decisions and became the Chief of a Court rendering justice in the name of the King. At first this Court was known as the “Court of conscience” and lately was designed as the main Court of Equity jurisdiction, using a concept – that of equity – with a polymorphous cultural heritage, as it was located at the intersection of Greek, Roman, and Christian traditions. In this perspective the English legal system institutionalized, in the proper form of a permanent jurisdiction, what in other legal experiences remained a theoretical or philosophical concept with particular and limited transpositions in the juridical domain.

In the second direction, Common Law and Equity grew up in a changeable relationship. As it has been noted, in the first genealogical period Equity was structured and recognized as a kind of supplementary justice and law: it was not a self-sufficient system, but at every point it presupposed the existence of common law; it was auxiliary, accessory, supplemental; it came to

rectify the severe asperity of common law, with a discretionary appreciation of the particularities embedded in single cases. Profound evolutions occurred in the course of the sixteenth century, when Equity jurisdiction appeared to be settled and consolidated. In this period, in fact, three main factors concurred to transmute the primordial aspects of Equity. First of all, Cardinal Wolsey was the last ecclesiastical Chancellor; from this time onward laymen, and eventually even great lawyers, were designated as Lord Chancellors of England. Consequently, Equity had to manage the secularization of its proper sources, the collapse of its theological foundation, and the possible collision with common law rules, which could be indirectly transposed by the new figure of Chancellor. The device used for the preservation of its autonomy marked the second renewed feature of the jurisdiction: to outlast, Equity came to adopt the same rhetoric of Common Law, putting the observance of precedents and codified rules before the consideration of the specificities of individual cases. Finally, both Tudors' and Stuarts' dynasties manipulated and adapted the same vision of Equity in order to achieve political goals (Costantini 2008). In this perspective the original nature of Equity as an extraordinary form of justice was converted into an *instrumentum regni*, into a means used to justify and support royal prerogatives against the restrictions imposed by the Law administered by the central Courts. To that end, new Courts – other than the historical Chancery Court – apt to dispense equitable remedies, were instituted: the Court of Requests, by late 1530 competent in civil matters for poor petitioners seeking relief for minor legal issues, and the Star Chamber, created by Henry VII in 1487 out of one of the traditional functions of King's Council and composed by the same members of the Privy Council when they were dealing with criminal prosecutions. The sphere of competence of the Star Chamber was progressively and strategically extended: from an agency of control and social discipline, it was transformed into an instrument of absolutism, as well as into a dreadful thread to those liberties guaranteed by common law. The Court inflicted sanctions of various nature and

intensity, from the monarch's displeasure, passing through fines and imprisonment, to conclude with corporal punishments. The procedure was a further adaptation of canon law and differed in many regards from the common law one. The effectiveness of these newborn Courts was evidence and measure of the respect with which the authority of the Sovereign was regarded.

The institutional framework, refashioned according to royal will and aspirations, laid the foundations for a parallel alteration of the peaceful relationship between Common Law and Equity jurisdictions. The harmony gave way to discord, tension, and frictions. The conflict patently arose in 1616, in James I's days, as a fierce battle among strong personalities (Simpson 1984). The leading role in the dispute was acted, on the one side, by Lord Ellesmere, in his quality of Chancellor and able common lawyer, and on the other, by the Lord Chief Justice, sir Edward Coke, the best active proponent of the reasons of the common law. The main object of contention was a recurrent practice of the Chancery Court, namely, hearing a case in Chancery after judgment had been given at common law. Obviously, the Common Law Courts were jealous of their powers and believed that this separate judgment – within the boundaries of a separate jurisdiction – was a kind of irregular appeal, contrary to statute. Therefore they made use of legal devices, and especially of writs of habeas corpus and prohibition, to fight against the illegal invasions and violations of the Prerogative Courts. Finally, given the institutional relevance, the quarrel was referred to the King, James I Stuart, in 1616. At this moment Lord Ellesmere, Francis Bacon, and the Duke of Buckingham worked together as the historical engineers of Coke's downfall (Baker 2002). Coke was dismissed from office and James I stated by decree that in case of conflict between Common Law and Equity, the rules of Equity would prevail. After Ellesmere's death, the relation between the two jurisdictions was reassessed in a cordial mood.

Another form of reversal should impress the course of English legal history. It was no longer related to the correlative dynamics of rival jurisdictions, but it concerned the proper nature and

structure of Equity. What, at first, was a device apt to soften the asperities of Common Law now hardened into law. It was an historical irony: the system of remedies and principles, which, in the past, concretely corrected the deficiency and the inadequacy of Common Law, finally was pervaded by even worse defects. During the seventeenth and eighteenth century the Court of Chancery was transformed from the elected place of relief and comfort to the proper locus of anguish and despair. All the possible aberrations were collected in Chancery: the procedure had become complex and cumbersome; the mass of documents and procedural acts were elephantine; the length of judicial processes had definitely sacrificed the instance of justice.

This phase of stagnation and involution continued up to the nineteenth century, the age of the great reforms, which led the foundation of the renewed system of jurisdiction introduced by the Judicature Acts in 1873–1875. This statute abolished the old Courts and established three levels of central justice: the first was the High Court, the second consisted of the Court of Appeal, and at the third was posed the Judicial Section of the House of Lords. The High Court historically derived from the final embodiment of the old courts of law and is internally structured into three divisions, the Queen's Bench Division (QBD), the Chancery Division (ChD), and the Family Division. It's important to underline that the Victorian legislation caused the procedural fusion of the old rules (originally divided into the two bodies of Common Law and Equity); the specialization of the three divisions is only a matter of convenience, insofar as all the judges are empowered to administer both Law and Equity. According to the latest reform, the appeal body of the House of Lords has been transformed into a new Supreme Court, with an independent seat from that of the House of Lords.

Mentality and Style Within the Common Law Tradition

From a comparatistic point of view, legal traditions can be defined, distinguished, and identified

assuming legal mentality and legal style as proper markers, or demarcation devices. The concept of mentality relates to the complex of elements pertaining thinking, discourse, narrative, symbolic, and social practices shared or recognized by a given community and characterized by persistence in time. The idea of style describes the forms and the perceptive qualities of a legal system; it is a replication of patterning that derives from a series of choices made within some set of constraints (Lang 1987).

The Common Law tradition is structured on the basis of a specific legal mentality and had elaborated an original legal style, which concurs to justify the differences between Common Law and Civil Law with regard to four main aspects: legal education, form of judgment, legal taxonomy, and epistemological attitude to law.

First of all, as a matter of fact, there is a close relationship between any system of law and the professionals, the experts who operate it (Dawson 1968). Historically, the peculiarity of English Common Law Tradition and its proper resistance to continental influences rest on the spaces and methods of legal education, on the subtle link between Bench and Bar, on the internal organization of legal profession. Traditionally, common lawyers were formed by practitioners, not by law professors, into the elitist space of the Inns of Court, not into the broader dimension of the Universities, where only Roman Law and Canon Law could be taught and transmitted. Change came only in the second half of the nineteenth century, when the academic study of law was established.

Moreover, by the end of the thirteenth century it had become a general custom that judges of the central courts could only be appointed from the professional bar. This trait came to distinguish the English legal tradition, assuring the peculiar strength of its professional elite.

Nowadays the legal profession in England and Wales is divided into two branches: those of barristers and solicitors. At the beginnings of their institutions each of them retains a proper monopoly (court representation for the barristers and conveyancing for the solicitors), but the reforms passed during the 1990s abraded this pristine privileges.

Binding force of judicial decisions represents the second marker of Common Law tradition (Lasser 2004). While on the Continent the internal coherence of the legal systems was granted by the act of normative codification, in England the notorious spirit of anticodification urged to find out another juridical device apt to achieve the same goal. The doctrine of precedent (or the doctrine of *stare decisis*), by which the Courts are obliged to respect their prior decisions, was affirmed during the nineteenth century, after the reorganization of the hierarchical structure of central jurisdiction and the introduction of an official system of law reporting. The effects of these principles occur both vertically (the decisions of a superior court bind all the inferior courts) and horizontally (a court is bound to follow its own precedents unless there is a strong reason to not do so). Arguing that a too rigid adherence to precedent may lead to injustice in a particular case and also unduly restrict the proper development of the law, with the Practice Direction of 1966, the Law Lords of the English House of Lords decided to depart from a previous decision when it appears right to do so. It is important to stress that what is binding is the only *ratio decidendi*, the rule of law heated by the judge as a necessary step in reaching the conclusions (Cross and Harris 1991), to be ascertained by an analysis of the material facts of the case, and to be distinguished from the *obiter dicta*, that is, things said by the way. The doctrine of precedent comes to shape legal reasoning in a proper manner, giving a specific emphasis to legal argumentation and to the precise appreciation of facts.

The third main difference between Common Law and Civil Law is the internal taxonomy. While in the Civil Law systems the principal dichotomy is posed between substantive and procedural law, and then, within substantive law, between private and public law, in the Common Law tradition the formative history of jurisdictions has prevented from the creation of rigid distinctions. The rules are agglutinated around three blocks – persons, things, and remedies – and, even after the procedural fusion caused by the Judicature Acts, maintain the mark impressed at their origin, so to be recognized as the

expressions respectively of the Common Law or of the Equity system of justice.

Cross-References

► [Civil Law System](#)

References

- Baker JH (2000) The common law tradition: lawyers, books, and the law. Hambledon Press, London
- Baker JH (2002) An introduction to English legal history, 4th edn. Butterworths, London
- Brand P (1992) The making of the common law. Hambledon Press, London
- Costantini C (2008) Equity breaking out: politics as justice. *Pólemos* 1:9–20
- Cross R, Harris J (1991) *Precedents in English law*. Oxford University Press, Oxford
- Dawson JP (1968) The oracles of the law. University of Michigan Law School, Ann Arbor
- Fortescue J, Chrimes SB (1942) *De laudibus legum anglie*. University Press, Cambridge
- Goodrich P (1990) Languages of law: from logics of memory to nomadic masks. Weidenfeld and Nicolson, London
- Holdsworth WS (1966) A history of English law. Methuen, London
- Lasser M (2004) Judicial deliberations: a comparative analysis of judicial transparency and legitimacy. Oxford University Press, Oxford
- Maitland FW, Chaytor AH, Whittaker WJ (1936) The forms of action at common law: a course of lectures. Cambridge University Press, Cambridge
- Maitland FW, Chaytor AH, Whittaker WJ (2011) *Equity: a course of lectures*. Cambridge University Press, Cambridge
- Milsom SFC (1969) Historical foundations of the common law. Butterworths, London
- Lang B (Ed) (1987) The concept of style. Coneee University Press, Ithaca
- Plucknett TFT (1956) A concise history of the common law. Little Brown, Boston
- Samuel G (2013) A short introduction to the common law. Edward Elgar, Cheltenham UK
- Simpson AWB (1984) Biographical dictionary of the common law. Butterworths, London
- Watt G (2008) *Equity & trusts law*. Oxford University Press, Oxford

Commons, Anticommons, and Semicommons

Enrico Bertacchini

Department of Economics and Statistics

“Cognetti de Martiis”, University of Torino,
Torino, Italy

Abstract

The notions of commons, anticommons, and semicommons are presented here to highlight their connections concerning how forms of ownership beyond the classical boundaries of private property affect the management of resources. While the three concepts have been presented as expressing specific dilemmas for the management of the resources or distinct property regimes, they may be seen as components of a unified interpretative framework which recognizes resources as collection of multiple attributes and addresses the complexity of mixed property regimes by studying the interaction of common and private uses.

Definition

A semicommons exists when the management of a resource is characterized by the coexistence of both common and private uses. Because a resource may be comprised of a bundle of attributes, which can be put to various productive uses, the introduction of semicommons highlights how different property regimes may coexist to govern the simultaneous uses of the resource. At the same time, a semicommons expresses a tragedy, because it induces problems of strategic behavior by agents in governing the externalities which emerge in the interaction between common and private uses. The semicommons therefore relates to and extends the notions of commons and anticommons in understanding how the allocation of property rights affects the management of resources.

Common Sense

► [Rationality](#)

Introduction

Commons, anticommons, and semicommons refer to three notions property law scholars have introduced to explain and analyze how forms of ownership beyond the classical boundaries of private property affect the management of resources and cope with creating incentives and dilemmas for multiple owners.

As the three concepts have been elaborated in different and subsequent periods, they reflect an evolution in the appreciation by property theory scholars of the complexity of governing resource systems through property regimes. On one hand, commons and anticommons pose two symmetric dilemmas and reflect the tension between too many use privileges and overfragmentation of private interests. On the other hand, the semicommons, with its dynamic interaction between private and common property over the same resource, offers new insights as to conditions in which mixed property regimes emerge and fragmentation solutions are favored to avoid strategic behavior in multiple uses.

This entry provides an overview of the main connections between the three concepts with a particular focus on explaining the most recent notion of the semicommons. The insights drawn from the scholarly contributions show that commons, anticommons, and semicommons may be seen as components of a comprehensive interpretative framework toward a better understanding of how resource systems are regarded as collections of multiple attributes and accommodate multiple uses that are most efficiently pursued at different scales, whether simultaneously or over time.

Commons and Anticommons Revisited

In studying common property regimes, scholars have long identified the so-called tragedy of the commons, which occurs when open-access or common ownership resources become overexploited due to the uncoordinated actions of the common right holders (Gordon 1954; Hardin 1968; Cheung 1970). The debate surrounding the commons centered on the institutional

mechanisms to avoid such tragedy (Ostrom 1990) and on the optimality of common versus private property regimes in managing physical nonexclusive resources.

More recently, scholars have turned their attention to anticommons phenomena. In this context, concurrent controls on entry over a common resource exercised by individual co-owners acting under conditions of individualistic competition and exclusion rights will be exercised even when the use of the common resource by one party could yield net social benefits.

While law and economics scholars have long recognized how multiple vetoes in contractual relations generate holdup problems which could produce inefficiency and underinvestment (Klein et al. 1978; Williamson 1979; Hart and Moore 1988), Heller (1998) has been the one who has initially popularized the definition of the anticommons and made more explicit this dilemma in property theory by drawing observation from the development in post-Communist Europe, where many buildings remain empty while numerous kiosks occupy the streets.

Introducing the anticommons has allowed to unveil similarities and peculiar symmetric relationships with the commons, which in turn helps analyzing the optimality of different property regimes.

According to formalized models presented by some authors (Buchanan and Yoon 2000; Schulz et al. 2002), commons and anticommons lead to symmetric tragic outcomes. Both the commons and the anticommons tragedies feature self-interested choices that are collectively sub-optimal. However, in common property regimes, too many use privileges lead to overexploitation of the common resource, while in the latter fragmentation in property and too many exclusion rights lead to underuse of the asset. In this perspective, Heller (1998, 1999) points out how commons and anticommons may be intended as property regimes which provide a useful framework to define the economic implications of property law for the management of resources. The choice of property regimes may in fact be seen as a continuum along commons, private property, and anticommons regimes. In this model the main

challenge is to set the boundaries of private property by taking into account two diverging forces. On the one hand, it is necessary to consider the trade-off that occurs in the so-called tragedy of the commons between the benefits derived from large-scale operations and the costs of overuse. On the other hand, property fragmentation may generate the familiar anticommons result of underexploitation.

While acknowledging a symmetric relationship between commons and anticommons, some scholars have nonetheless warned from regarding the anticommons as a new distinct type of property regime. For example, Lueck and Miceli (2007) argue that anticommons should more properly be seen as an “open-access” investment problem: a resource, such as land or apartments, can be overused by multiple owners, but the investment on the asset can be simultaneously suboptimal because of the lack of unanimous agreement by the multiple owners. In a similar vein and focusing on a game theoretical approach, Fennell (2011) suggests that the most promising and general approach to assess commons and anticommons is to consider the distinct strategic behavior underlying the two situations. Whereas the commons tragedy follows the strategic pattern of the prisoner’s dilemma, the anticommons often resembles the strategic game of chicken.

Commons and anticommons may be therefore seen as useful metaphors for understanding the misalignment of incentives of multiple owners who wish to use a common resource. Much of the confusion concerning the definition of commons and anticommons either as dilemmas in the management of common resources or as distinct property regimes stems from the fact that most of the studies on property regimes assume that each underlying resource possesses one unique attribute that allows for a single productive activity to take place. To the contrary, however, a resource system may be comprised of a bundle of useful attributes, which can be put to various productive uses and at different scales of operation (Barzel 1982; Lancaster 1966; Smith 2002). Considering the familiar example of the grazing field used to explain the tragedy of the commons, Fennell (2011) points out how the problem is not only

related to the fact that grazing is pursued at a large scale and on commonly owned ground, but also to the fact that the commonly owned attributes interact with other privately owned inputs, such as cattle. The overuse of the commons emerges from the appropriation of benefit to the privately owned attribute of the resource system. If only land were privatized, overgrazing would be avoided, but the misalignment of incentives and the overuse would affect other attributes or inputs still experienced in common, such as water. As a result, the intrinsic dilemma of the commons is more properly due to a problem of efficient specification of what attributes of a resource system are held in private or common ownership and how those property regimes interact in the productive use. In a similar vein, in the anticommons case, fragmentation of property over a resource may be seen as a private individual property arrangement held by distinct owners as long as some attributes of the resource fail to be available for potential productive use at a greater scale of operation due to assembly problem of property rights.

Introducing the Semicommons

The recognition that resources hold bundle of useful attributes implies that different property regimes may coexist to govern the simultaneous uses of the resource and add new insights in the analysis of the complexity of mixed ownership regimes.

In this context, Smith (2000) has developed the notion of the semicommons, a property regime in which one attribute of a resource is privately owned, while another is in common ownership, and the two potentially interact. The now classic example of the semicommons is taken from medieval open fields. In such fields, peasants had exclusive property rights in strips of land for purposes of growing grain but shared these land strips with other farmers in one large grazing commons during fallow seasons. A semicommons property regime enables the various right holders to benefit from the multiple uses of the resource. First, semicommons owners obtain the advantage of scale

economies. This is demonstrated in the medieval field example, as the peasants hold the resource in common for grazing. Secondly, semicommons harness private incentives by allocating privately owned rights on the resource. Interestingly, a semicommons property regime may induce problems of strategic behavior that go beyond the familiar incentives of overuse that exist in a common property regime.

Similar to the tragedy of the commons, semicommons members may overuse the common resource to the extent that they internalize merely a fraction of the cost of their actions. Additionally, however, because of the interaction between common and private uses, owners will attempt to distribute benefits to their own part of the commons and steer bads to the other parts of the assets. In the open field example, peasants have incentives to allocate the benefits of manure onto the part of the commons that they privately own during fallow season and to further steer the damage of trampling onto the land of others (Smith 2000). Institutional solutions are needed to moderate such strategic behavior in a semicommons setting. In the example of open fields, the scattering of privately owned strips reduces strategic behavior among farmers. By randomly dispersing the privately held parcels, the altered borders of the land make the costs of engaging in strategic behavior prohibitive.

Semicommons Between Commons and Anticommons

Because of its peculiar characteristics, the semicommons holds fundamental relationships with both commons and anticommons arrangements.

From an analytical viewpoint, semicommons has been related to the commons using the latter as a reference benchmark to analyze the economic benefits and cost of the two ownership structures. On one hand, semicommon property allows operation on a larger scale than common property by enabling both common and private uses. If adding one productive activity leads to a more efficient use of the resource (i.e., by increasing total output), the semicommons would be a superior

solution instead of a pure common or private property regime where only one productive use is performed. On the other hand, it poses strategic problems that may well go beyond the familiar incentive of overuse in a commons (Smith 2000). In particular the uncoordinated behavior by agents in a pure commons is likely to generate “equally shared” externalities among agents. By contrast, in a semicommons, actions taken while the resource is in common use generate positive or negative external effects on the privately owned parcels. For this reason, agents in a semicommons will try to act strategically by trying to distribute in unequal share to their privately owned attributes the negative or positive externalities arising from their common use activities. As a result, the interaction between private and common uses induces strategic behavior that may undermine the gains provided by multiple uses of the resource and potentially leads to a lower value of using the resource in a mixed property regime. Considering negative externalities generated by common use to the privately owned attributes of the resource, Bertacchini et al. (2009) have confirmed through a formalized model the more tragic outcome of semicommons property arrangements as compared to the stylized commons, with an increased overexploitation of the resource by the common use activity. However, it is not still clear whether the same conclusion applies in the presence of positive spillovers from common use to privately owned attributes use or vice versa.

As for the relationship between semicommons and anticommons, this may be found in the role of scattering strategy in the privately owned attribute of the resource.

According to Smith (2000), scattering reduces the opportunities for strategic behavior. However, it may also reduce the efficiency of crop production, since the privately owned attribute is fragmented to a scale of operation that might be below the efficiency standard. This highlights the trade-off that agents face when they choose to implement a scattering strategy. On the one hand, agents need to fragment the privately owned attribute enough so as to increase the costs of strategic behavior and maintain the

multiple use of the resource. On the other hand, agents seek to minimize the losses that derive from the efficiency reduction in the production activity pursued with private use.

As noted by Bertacchini et al. (2009), a rule of scattering allows agents to contract into an anticommons regime in order to create a mix of private and common property. The use of scattering illustrates in fact a potential virtuous linkage between commons and anticommons property regimes. A number of scholars have suggested that an anticommons regime is a desirable allocation of property rights when nonuse of the resource is the preferred equilibrium, such as in the context of conservation management or environmental preservation of resources (Mahoney 2002; Parisi et al. 2005). Analysis of semicommons ownership reveals an additional benefit of anticommons property arrangements. When the mix of private and common uses generates strategic behavior, fragmenting a resource for one activity can thus be useful to achieve efficiency with regard to other activities. For instance, a semicommons can introduce a benevolent “comedy” of the anticommons by fragmenting property rights in an attribute of the resource in order to realize scale benefits of the resource’s other attributes.

More generally, the introduction of the semicommons opens the door to a broader analysis of property regimes and resources that involve different scales of use.

According to Fennell (2011), the semicommons is thus less a distinctive property type than a frame through which to view existing or proposed arrangements that involve activities at different scales, whether simultaneously or over time. On this account, Smith (2000, 2005) recognizes that scattering strategies of privately owned attributes are quite common forms of boundary placement, such as proportionate holding schemes or *ex ante* uncertainty about the appropriation of benefits deriving from common use activities to one’s holdings. In this sense, scattering is a fragmentation strategy which enables alignment of incentives by providing a more cost-effective solution as compared to maintaining an attribute of a given resource in

common and establishing governance norms to monitor compliance.

Other Applications of the Semicommons

Other than the semicommons in the English open field, there is a growing literature which extends the use of this concept in other domains.

The interpretative framework of the semicommons is particularly suited to analyze property rights over fugitive resources, such as water (Smith 2008), or assets contributed to a joint venture (Smith 2005). In this latter case, an asset can be used for purposes of the joint venture, but the joint ventures may retain certain private uses and engage in strategic behavior to extract benefits for assets over which they retain some private uses and dump costs to the assets over which others have retained private uses.

More importantly, the notion of semicommons has been applied to intellectual resources, such as information and knowledge.

Because information is difficult to subject to exclusive rights and because multiple uses may derive from the non-rivalrous resource, intellectual property law is argued to establish a semicommon property regime in information goods (Heverly 2003; Frischmann 2007). In this case, intellectual property law devises a dynamic interaction between the common use and the protection of private production of intellectual resources. Although it is true that intellectual property rights partly enclose the information commons (public domain), it is equally true that they feed in many ways the information commons. This may occur, for instance, when intellectual property rights expire or in cases of research exemption in patents and fair use doctrine in copyright laws. Likewise, the information commons partly enclosed by the intellectual property rights provides information inputs for the private creation of new intellectual resources. Strategic behaviors arise also in the information semicommons and can be of two types. The illegal reproduction and distribution of privately owned information may be deemed as an improper expansion of common use because pirates

strategically distribute harms to the owners of the protected information.

To the same extent, the strengthening of exclusive rights that blocks the access to the common components of information may be seen as an expansion in protection of private use that restricts the use of the information in public domain.

A slightly different application of the semicommons in intellectual resources may be found in Reichman (2011), dealing with innovation and intellectual property regimes.

According to Reichman, the realm of industrial property law may be divided into three spheres: a commons, a semicommons, and exclusive rights (in the form of patents). The commons is characterized by free flow of scientific and technical information that are extensively government generated or funded in public research programs. At the other extreme, patents confer exclusive rights to innovators as reward for their investments in innovative endeavor and historically have been granted for truly nonobvious inventions. Between these extremes lies the main area of industrial innovative activity that does not rely on path-breaking and discontinuous inventions but rather on cumulative and routine applications of know-how to industry.

The cumulative development of know-how is seen as a semicommons because it reflects a community project that benefits from the small-scale contributions.

Indeed, small-scale innovators draw from the public domain to make improvements and enrich the public domain by generating new information that others in the technical community may exploit to their own advantage (dynamic interaction between private and common use). The cumulative additions are based on the private use of information inputs and reverse engineering practices.

Nevertheless, these additions do not attract exclusive property rights because they normally do not surpass the ability of the routine engineers who comprise the relevant technical communities.

Because the overall problem of innovators is to recoup their investment in innovation through the availability of lead time, the patent system grants legal lead time to nonobvious inventions. On the

contrary, in the semicommons of incremental applications, innovators have just a natural lead time. In this case, the lead time greatly depends either on the protection of trade secrecy from unlawful misappropriation of new industrial applications or on the technological conditions that allow competitors to lawfully reverse engineer the industrial application. In summary, under the Patents-Trade Secret system, the semicommons of innovative applications of know-how favors the spread of innovation in a healthy competitive environment. At the same time it impedes single small-scale innovators to strategically behave, removing their contributions from the semicommons by means of exclusive property rights.

References

- Barzel Y (1982) Measurement cost and the organization of markets. *J Law Econ* 25:27–48
- Bertacchini E, De Mot J, Depoorter B (2009) Never two without three: commons, anticommons and semicommons. *Rev Law Econ* 5(1):163–176
- Buchanan J, Yoon Y (2000) Symmetric tragedies: commons and anticommons property. *J Law Econ* 43:1–13
- Cheung S (1970) The structure of a contract and the theory of a non-exclusive resource. *Journal of Law and Economics* 13(1): 49–70.
- Fennell LA (2011) Commons, anticommons, semicommons. In: Ayotte K, Smith HE (eds) *Research handbook on the economics of property law*. Edward Elgar, Cheltenham
- Frischmann BM (2007) Evaluating the demsetzian trend in copyright law. *Rev Law Econ* 3(3):649–677
- Gordon S (1954) The economic theory of a common-property resource: the fishery. *J Polit Econ* 62(2):124–142
- Hardin G (1968) The tragedy of the commons. *Science* 162:1243–1248
- Hart O, Moore J (1988) Incomplete contracts and renegotiation. *Econometrica* 56(4):755–785
- Heller M (1998) The tragedy of the anticommons: property in the transition from Marx to markets. *Harv Law Rev* 111:621–687
- Heller M (1999) The boundaries of private property. *Yale Law J* 108:1163–1223
- Heverly RA (2003) The information semicommons. *Berkeley Technol Law J* 18:1127–1189
- Klein B, Crawford RG, Alchian AA (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J Law Econ* 21(2):297–326
- Lancaster KJ (1966) A new approach to consumer theory. *J Polit Econ* 74:132–156

- Lueck D, Miceli T (2007) Property law. In: Shavell S, Polinsky AM (eds) *Handbook of law & economics*. Elsevier, Amsterdam
- Mahoney JD (2002) Perpetual restrictions on land and the problem of the future. *Virginia Law Rev* 88:739–775
- Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK
- Parisi F, Depoorter B, Schulz N (2005) Duality in property: commons and anticommons. *Int Rev Law Econ* 25:578–591
- Reichman JH (2011) How trade secrecy law generates a natural semicommons of innovative know-how. In: Dreyfuss RC, Strandburg KJ (eds) *From the law and theory of trade secrecy: a handbook of contemporary research*. Edward Elgar, Cheltenham, pp 185–200
- Schulz N, Parisi F, Depoorter B (2002) Fragmentation in property: towards a general model. *J Inst Theor Econ* 158(4):594–613
- Smith HE (2000) Semicommon property rights and scattering in the open fields. *J Leg Stud* 29:131–169
- Smith HE (2002) Exclusion versus governance: two strategies for delineating property rights. *J Leg Stud* 31:453–487
- Smith HE (2005) Governing the tele-semicommons. *Yale Law J Regul* 22:289–314
- Smith HE (2008) Governing water: the semicommons of fluid property rights. *Ariz Law Rev* 50(2):445–478
- Williamson OE (1979) Transaction-cost economics: the governance of contractual relations. *J Law Econ* 22(2):233–261

Company Liability

► Organizational Liability

Competition Policy: France

Marc Deschamps
CRESE EA3190, Université Bourgogne Franche-Comté, Besançon, France

Abstract

From a law and economics perspective, competition policy has always been a natural connection between law, economics, and politics. Antitrust is at its heart, but competition policy is a much larger topic. In our entry, we will

shed some light on how competition policy is conceived and structured in modern-day France.

Introduction

The expression “competition policy” is sometimes misused as synonymous with “antitrust.” This is a mistake because competition policy is the articulation between politics, law, and economics, while antitrust is generally understood to be the totality of legal and economic provisions for dealing with infringements of the competition rules (i.e., essentially abuse of dominant position, cartel, concentration). In other words, antitrust, understood in its most common sense, evacuates the political dimension and concentrates on the technical aspects. This clarification is not merely a terminological one but also makes it clear that, in the field of competition policy, it is impossible not to make political choices and that, in order to be coherent, they must be in line with all the others policies. Moreover, this distinction makes it possible to understand why it is possible, for example, to observe that, apart from the case of a different technical assessment, two countries may consider similar practices of businesses differently or why they do not have the same procedural or substantive rules. Thus, far from following the same path, countries choose their policy according to their size, their degree of openness, their growth strategy, and so on.

Moreover, because France is a member of the European Union, France’s competition policy is politically and legally constructed on two bases: national law and European law (i.e., the law deriving from the Council of Europe and the law of the European Union). Under the European treaties, European Union law is an own legal order with primacy over national rights and direct effect. We will here concentrate on the law and actors at the national level.

In the remainder of this article, we will present, in turn, the essential elements relating to the genesis of French competition policy (1), its main players (2), and the structure of French competition law (3).

Origins and Development of French Competition Policy

The idea that free trade is the best way to promote prosperity emerged in France at least since Montesquieu. Yet what characterized prerevolutionary France is its corporatist organization and the existence of *jurandes*, that is to say, bodies of trade whose members held a monopoly. The spirit of competition was introduced to France in 1791 with two laws: first, that of March 2–17, known as the Allarde decree, which established the principle of freedom of commerce and industry. Further, the law of June 14–17, known as the Le Chapelier law, definitively abolished guilds. With the Penal Code of 1806, there is a first provision, Article 419, which prohibits the hoarding of commodities in order to raise prices. These remain the main elements that characterize French competition policy before the second half of the twentieth century.

France began to introduce a competition law with the ordinance of 30 June 1945 on prices. After there was: the decree of 9 August 1953 (Prohibition of Unlawful Settlement Agreements and Establishment of a Technical Commission on Cartels), the Act of 2 July 1963 (Prohibition of Abuses of Dominant Positions and Establishment of a Technical Commission on Restrictive Practices and Dominant Positions), and the Act of 19 July 1977 (creation of the Competition Commission with additional advisory functions on mergers and any competition issues).

A paradigm change took place with the ordinance of 1 December 1986, since it was on this basis that France established a genuine autonomous competition law based on the general principle of freedom of prices (Article L.410-1 and L.410-2 of the French Commercial Code). This empowerment with regard to political control is materialized by the creation of the *Conseil de la Concurrence* (Competition Council), an independent body henceforth given the power of sanctioning businesses in the event of anticompetitive practices (taking it over from the Minister in charge of the economy). The Council is controlled by the judiciary (the *Cour d'appel de Paris* and the *Cour de cassation*). The same decree extends the scope for referral and offers a procedure that

better guarantees the rights of the concerned persons. Subsequently, the law of 11 December 1992 empowers the Competition Council to apply European provisions, and the law of 15 May 2001 strengthens competition law by introducing leniency and settlement procedures, raising the maximum penalties, improving international cooperation, and introducing systematic monitoring of mergers. Under the European leadership, the decree of 4 November 2004 aligns the powers of the Competition Council with those of the other European competition authorities.

The last break took place in 2008. Indeed, the law of 4 August created the Competition Authority (*Autorité de la Concurrence, ADLC*) and affords this Competition Authority the ability to conduct investigations of its own, the ability to self-refer in advisory matters, and the power to decide in terms of control of concentrations. Until then the Competition Council was only responsible for the analysis of the merger, and the decision remained with the minister in charge of the economy. In addition, the decree of 13 November gave the Competition Authority an investigative unit under the responsibility of a general rapporteur, as well as the capacity to issue behavioral and structural injunctions.

Actors in French Competition Policy

The French competition policy is organized around three groups of actors: political actors, the Competition Authority, and the courts.

Except for sporadic laws and regulations, political actors now exercise only a residual role. Indeed, there are essentially only three options available to the government to intervene in matters of competition policy. The first is the Competition, Consumer Affairs, and Prevention of Fraud Directorate General (DGCCRF), which, with 3000 agents under the minister's authority, is responsible for the competition regulation of markets by fighting unfair commercial practices between traders and anticompetitive practices. On the understanding that the DGCCRF has only the power of injunction and transaction for local practices (less than 200 million euros for all responsible companies), the ADLC retains the

power to take preliminary action and ensures that the case is dealt with in the event of refusal or nonfulfillment of the undertakings. The second possibility for the minister responsible for the economy to intervene in competition policy concerns merger control (Article L.430-7-1 of the French Commercial Code). The minister may ask for a thorough examination, and he can also finally take a decision that is contrary to that of the ADLC on grounds of general interest other than the maintenance of competition (e.g., industrial development, creation or conservation of employment). Thirdly, the minister in charge of the economy has certain prerogatives and powers with regard to practices restricting competition.

Under Article L.461-1 of the French Commercial Code, the ADLC is an independent administrative authority responsible for ensuring the free movement of competition and for supporting the competitive functioning of markets at European and international levels. It has an annual budget of around 20 million euros and employs about 200 people. It has three main components:

1. A 17-member college with a president and four vice-presidents holding full-time positions (the other members are nonpermanent and hold positions in jurisdictions, associations, or businesses); this college publishes opinions (advisory function) and decisions (decision-making function and merger control).
2. Investigative services under the authority of the general rapporteur.
3. An auditor who acts as procedural mediator at the disposal of the involved parties.

It should therefore be noted that the ADLC clearly separates the investigative acts, which are the responsibility of the general rapporteur, and the opinions and decisions, which are the responsibility of the college. Since the law of August 6, 2015, the ADLC also offers a mapping proposal to the ministries of justice and the economy for notaries, bailiffs, and auctioneers.

There are two reasons that courts have a central role in the area of competition in France. First, it is they who are in charge of the control of the ADLC: (1) in the case of anticompetitive

practices, the *Cour d'appel de Paris* is the appellate court, and the *Cour de Cassation* acts as supreme court; and (2) the *Conseil d'Etat* has a decision function: in merger control, for the advisory function of ADLC, as well as for all administrative acts it takes in relation to it. Further, both the judicial and the administrative courts have to deal with certain disputes, such as those relating to restrictive practices, acts of unfair competition, and individual or collective actions for damages.

The Structure of French Competition Law

French competition law is sometimes split into two parts for pedagogical purposes: the “small competition law” and the “large competition law.” It is within the first that the essence of the national specificities lies.

“Small competition law” designates the provisions of the French Commercial Code and case law relating to tariff transparency, noncompetition clauses, restriction of competition, and unfair competition. We shall briefly mention only the latter two. As early as 1945, France chose to issue rules to protect the contractor without even having to prove a breach of the market. These rules constitute what are called restrictive practices. France does not have a law defining unfair competition, which is therefore purely based on case law. Legal doctrine classically distinguishes between acts of confusion, denigration, disorganization, and parasitism.

“Large competition law” corresponds to national provisions on anticompetitive practices (cartel and abuse of dominant position) and merger control, decisions of the ADLC, and the case law of the *Cour d'appel de Paris*, the *Cour de cassation*, and the *Conseil d'Etat*. To this the European provisions, decisions, and case law relating to these areas or concerning the control of State aid must be added.

Conclusion

French competition law has followed the European and international trends and today has

a good record since, for example, its antitrust appears to conform to the highest world standards as is proved by its five-stars rating by the Global Competition Review.

Nevertheless, according to us, there remain four main challenges for French competition policy. The first is to create a global and clear structure for the restrictive practices of competition and unfair competition. Second, it would be very useful for all shareholders to have a ranking of the ADLC's decisions and opinions as it is done by the *Cour de Cassation* or the *Conseil d'Etat*. Third, opinions and decisions would make more sense if the Competition Authority would develop economic models in line with economic academic standards. Last but not least, it would be more democratic and efficient if everybody could understand who really decides the French competition *policy* and what it will be for the next years, as it is the case, for example, at the federal level for the United States or for the European Union.

Cross-References

- Abuse of Dominance
- Act-Based Sanctions
- Cartels and Collusion
- Competition Policy: France
- Conflict of Interest
- European Law
- Leniency Programs
- Merger Control

References

- Autorité de la concurrence. see <http://www.autoritedelaconcurrence.fr/user/index.php?lang=en>. Accessed 1 Jan 2017
- Code de commerce. see <https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000005634379>. Accessed 1 Jan 2017
- DGCCRF. see <http://www.economie.gouv.fr/dgccrf/concurrence>. Accessed 1 Jan 2017
- Cour d'appel de Paris. see <http://www.ca-paris.justice.fr/>. Accessed 1 Jan 2017
- Cour de cassation. see <https://www.courdecassation.fr/>. Accessed 1 Jan 2017
- Conseil d'Etat. see <http://english.conseil-etat.fr/>. Accessed 1 Jan 2017

Competitive Neutrality

Régis Lanneau

Law school, CRDP, Université de Paris Nanterre, Nanterre, France

Definition

The concept of competitive neutrality is not yet a legal concept in most OECD countries; nevertheless, the OECD is pushing for its wider integration. Competitive neutrality is a “regulatory framework (i) within which public and private enterprises face the same set of rules and (ii) where no contact with the state brings competitive advantage to any market participant” (OECD 2009). From an economic point of view, it is providing a narrative through which most public law could be reinterpreted.

Introduction

The concept of competitive neutrality is relatively new. Its first legal occurrence dated back in 1995 in Australia's “Competition Principles Agreement.” In 1996, the Commonwealth of Australia released its own “Competitive Neutrality Policy Statement.” In the beginning of the 2000s, the Netherlands, Finland, and Sweden, while not necessarily using the word, developed regulations dealing with competitive neutrality problems. The European Union, through its treaties and their interpretation by the European Court of Justice, could also be considered as having developed a competitive neutrality framework. Since 2004 and a document entitled “Regulating Market Activities by Public Sector,” the OECD is producing guidelines and gathering good practices regarding competitive neutrality, no less than nine reports and organized conferences around this topic. Following OECD, the UNCTAD also introduced the concept in 2014 to inquire into the practices of some developing countries (including

China, India, or Malaysia). As the OECD recognized on its website, “While the principle of competitive neutrality is gaining wide support around the world, obtaining it in practice is a much more difficult question.”

Indeed, if the idea of competitive neutrality might appear fairly simple in theory, its practical aspects are far from being easy especially since the notion does not have a stable meaning and evolved through time. At first and in its narrowest conception, “Competitive neutrality (CN) requires that government business activities do not have net competitive advantages over their private sector competitors simply as a result of their public ownership” (Commonwealth of Australia 1998). In a broader approach, it is a “regulatory framework (i) within which public and private enterprises face the same set of rules and (ii) where no contact with the state brings competitive advantage to any market participant” (OECD 2009). From these definitions, this concept is addressing the question of the limits of state intervention in the economy especially when such intervention could distort competitive positions: “Competitive neutrality occurs where no entity operating in an economic market is subject to undue competitive advantages or disadvantages” (OECD 2012).

Since this concept cannot be understood without its economic rationale, it could be considered that law and economics considerations are constitutive of its foundations. From a more legal point of view, competitive neutrality could play a key function in reorganizing and unifying (through legal interpretation) different areas of law, from competition law to state aids, procurement, and tax law.

The Economic Rationale of Competitive Neutrality

For the OECD, “the main economic rationale for pursuing competitive neutrality is that it enhances allocative” (OECD 2012). Indeed, competitive neutrality is following the same logic as competition law, even if it is also possible to consider that its function is also to promote accountability of decision makers. This economic rationale is sometimes criticized for its probable

consequences, reducing the size of the public sector or the range of state interventions.

Promoting Efficiency and Accountability

Competitive neutrality appears to address the issue of competition between the public sector (and especially state-owned enterprise) and the private sector businesses. Indeed, both sectors are sometimes competing to provide the same goods or services to consumers. If public sector enterprises are providing these goods and services while benefiting from competitive advantages (resulting from different regulations – including exemption from competition law – interest rates, easier access to public data, procurement contract, or taxes to mention few sources of advantages), distortions of competition might occur leading to misallocation of resources. Indeed, according to mainstream economic theory, in such a system, there are no reasons why the goods and services will be produced and provided by those who can do it most efficiently. Competitive neutrality would then be a way to maintain the integrity of the markets and thus their allocative efficiency. From a dynamic point of view, insuring the integrity of the markets is also not without consequences regarding innovation and economic development (e.g., Graham and Richardson 1995); it should also logically push public entities to operate more efficiently through the discipline provided by the market.

The other idea behind competitive neutrality is that, in a neutral environment (or in the process of achieving such an environment), it would be possible to assess the true cost of the provision of public services when these services are provided by entity which are also competing with the private sector in some branch of their activities. Democratic debate could then be based on empirical data.

Promoting Efficiency or Reducing the Size of the Public Sector and the Range of State Intervention?

Since competitive neutrality is necessarily restricting the type of governmental intervention and forcing state-owned companies to adopt “private” behaviors, some might believe that the

concept's purpose is merely to reduce state intervention and the size of the public sector. It is difficult to address this question in this entry. However, three points should be mentioned.

First, when services of general interests are not provided adequately (or at all) by private agents in a market, public authorities are not hampered to provide these services through public entities or after call for tender. Nevertheless, in such a case, the entity providing the service, especially when it is operating also in a competitive domain, should neither be over nor under compensated to avoid competitive distortions.

Second, it should be remembered, at least in the context of the European Union, that some activities are not considered as “economic” like the army, air navigation, maritime traffic, antipollution surveillance, organization, financing, and operation of prison. Moreover, this list can include other activities when they are structured in a certain way like social security or health care based on solidarity principles or public education. Lastly, certain activities, because of their size, are often not considered as threat to competitive neutrality like local museum, local hospital, or swimming pool (European Union 2011).

Third, in situation of crisis, sticking to competitive neutrality could be unproductive. Bank bail-outs are often mentioned (OECD 2009).

Competitive Neutrality in Practice

Giving substance to competitive neutrality is far from being easy (OECD 2015). Indeed, this concept is much more an aspiration than a perfectly feasible reality. Two dimensions should be considered with attention: its scope and implementation principles.

The Scope of Competitive Neutrality and State's Margin of Appreciation

When addressing competitive neutrality, the first task is to identify its scope. At this level, two dimensions should be considered: its breadth and its depth.

Regarding the former, the problem is to define what economic activities are. As it has been

mentioned before, it is possible to adopt an extensive conception – providing goods and services – or a more restricted one considering that these goods and services are “economic,” thus reducing the scope of the market, or through the exemption of certain activities because of their size, defined by commercial turnover (Australia), market power, or capacity to distort markets. In any case, it will be required to justify exemptions from this principle. These justifications are often economic in their nature (the good or the service cannot be provided through markets), but political justifications are quite frequent (solidarity, services of general interests, sovereignty, independence, etc.).

Regarding the later, the problem is to identify both the entities subjected to competitive neutrality and the type of governmental interventions which could have impacts on competition. Indeed, according to the narrow definition of competitive neutrality, only state-owned enterprises are considered, and the purpose is merely to insure they are not benefiting from any competitive advantage because of their “public” nature. A broader conception will consider that the ownership structure should be irrelevant to competitive positions. An even broader conception will consider all economic entities and will only focus on state direct and indirect state interventions. It is at this level that the concept of competitive neutrality is the most interesting because it is forcing us to identify the type of intervention that could influence competitive positions. Not only are subsidies and compensations for public services' obligations concerned; governance structures, regulations (and especially administrative law, contract law, and labor law), taxes, cost of capital, access to public contracts, etc. are also relevant dimensions (OECD 2012; Lanneau 2016). Virtually all interventions could have consequences on competitive positions of economic agents. Only a case-by-case approach could allow the identification of these effects.

Implementation Principles

If competitive neutrality is difficult to achieve, some procedural requirements for its implementation are certainly required.

First and foremost, transparency is a key requirement. Indeed, to assess if some entities are not benefiting from a favorable treatment by the state, it is required that its intervention regarding these entities should be made in perfect transparency. This is especially true regarding the compensation for public service obligations but also for grants, loans, or other services provided by the state for the benefit of some entities. This transparency could also force some entities to maintain two separate accountabilities, one for commercial activities and the other for non-commercial activities, to facilitate the evaluation of the adequacy of compensations for public services' obligations. These principles are, for example, enacted in the transparency directive (80/723/EEC of 25 June 1980 amended several times and codified by the directive 2006/111/EC of 16 November 2006) of the European Commission. In any cases, the value of the advantage – if an advantage is identified – will be difficult to assess.

Second, some procedures should be developed to allow competitor to challenge some state's interventions, not only direct but also indirect. Indeed, since it is difficult to identify ex ante the interventions that could lead to offer some competitive advantages, it is required to decentralize the possibility to identify “problematic” interventions. Of course, some exemptions by categories could be enacted to reduce enforcement costs, but most of these categories cannot be ascertained ex ante. For some specific acts, some ex ante procedure could also be created to avoid the consequences of competitive distortion.

Conclusion

Competitive neutrality could offer a real opportunity to address, under a common framework, different areas of law or regulations which are often considered as separated. Moreover, it would offer the opportunity to compare national practices because of its functional dimension. If the concept has not yet entered in many national legislations, its potentiality is difficult to deny, and law and

economics could find, in this topic, a fantastic field of study.

Cross-References

- [Political Competition](#)
- [Public Choice: The Virginia School](#)
- [Public Goods](#)
- [Public Interest](#)
- [Regulatory Impact Assessment](#)
- [State Aids and Subsidies](#)
- [State-Owned Enterprises](#)

References

- Commonwealth of Australia (1998) Commonwealth competitive neutrality guidelines for managers. Available online: <http://archive.treasury.gov.au/documents/274/PDF/cnguide.pdf>
- European Union (2011) Communication from the Commission on the application of the European Union State aid rules to compensation granted for the provision of services of general economic interest. OJEU C 8 of 11.1.2012, <http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:008:0004:0014:EN:PDF>
- Graham E, Richardson D (1995) Competition policies for the global economy. Columbia University Press, New York
- Lanneau R (2016) La Neutralité Concurrentielle, Nouvelle Boussole du Droit (Public) Economique. Droit administratif, 2016 no. 1
- OECD (2004) Regulating market activities by public sector, DAF/COMP(2004)36. Available online: [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DAF/COMP\(2004\)36&docLanguage=En](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DAF/COMP(2004)36&docLanguage=En)
- OECD (2009) Policy roundtables, state owned enterprises and the principle of competitive neutrality. Available online: <http://www.oecd.org/daf/ca/corporategovernanceofstate-ownedenterprises/50251005.pdf>
- OECD (2012) Competitive neutrality: maintaining a level playing field between public and private business. Available online: <http://www.oecd.org/daf/ca/corporategovernanceofstate-ownedenterprises/50302961.pdf>
- OECD (2015) Roundtable on competitive neutrality in competition enforcement – note by the European Union, DAF/COMP/WD(2015)31. Available online: http://ec.europa.eu/competition/international/multilateral/2015_june_competitive_neutrality_en.pdf
- UNCTAD (2014) UNCTAD research partnership platform: competitive neutrality and its application in selected developing countries. Available online: http://unctad.org/en/PublicationsLibrary/ditcclpmisc2014d1_en.pdf

Concentrated Ownership

Caspar Rose
Copenhagen Business School, Copenhagen,
Denmark

Abstract

This entry summarizes the main theoretical contributions and empirical findings in relation to concentrated ownership from a law and economics perspective. The various forms of concentrated ownership are described as well as analyzed from the perspective of the legal protection of investors, especially minority shareholders. Concentrated ownership is associated with benefits and costs. Concentrated ownership may reduce agency costs by increased monitoring of top management. However, concentrated ownership may also provide dominating owners with private benefits of control.

Synonyms

[Blockholders](#); [Majority control](#)

Mitigating Agency Costs Through Concentrated Ownership

Modern listed firms are characterized by having a diffuse ownership structure with a profound separation between ownership and control. This is contrary to the heavily concentrated ownership structure that dominated business life some centuries ago where a patriarchal management structure – often relying on a family that maintained control over the firm's business and operations. However, as production became more capital intense due to the technological expansion in the eighteenth century, there was a need to expand ownership in order to attract sufficient capital. In some countries, e.g., in Northern Europe, the founder or his/her descendants often maintained control of the firm, since they held the shares with superior voting rights, whereas

outside suppliers of equity held shares with inferior voting rights.

The separation between management and ownership in modern firms creates some potential conflicts of interests between management and shareholders. Agency costs are generated due to the separation of ownership as investors are not able to monitor management without incurring costs (cf. the free rider problem). This is foreseen by the external providers of capital, so even if management has identified projects with positive net present value, it may find it difficult to convince capital suppliers to invest their wealth in the firm (see Monks and Minow 2001).

In case management promises not to exploit the firm's resources is in the terminology of game theory, simply not credible. Moreover, agency costs may be reduced if a blockholder exercises active ownership, which may benefit all the other minority shareholders. However, whether concentrated ownership is beneficial for all the shareholders or even society is still a theoretical and empirical unsettled question.

In the following concentrated ownership and investor protection are discussed which is followed by different ownership types, i.e., family and foundation ownership as well as managerial ownership. The entry ends with discussion of the increasing role of institutional investors, which also includes key empirical contributions.

Concentrated Ownership and Investor Protection

In the last decade there has been an increasing attention to the issue of investor protection, in particular how investor protection influences the development and functioning of capital markets globally. Shleifer and Vishny (1997) argue that concentrated ownership and investor protection may serve as important mechanisms to reduce agency costs due to the separation of ownership and control.

Law and economics is the study of how legal rules and court practice affect parties' incentives. The legal setup that regulates investor protection such as in company law, security regulation, and

accounting standards is enormous. However, the important issue here is that changes in the legal setup change the balance of power between management and investors, i.e., shareholders and creditors. If investor protection is increased, the development of financial market is increased as investors are more likely not to be exploited by top management. In relation to concentrated ownership, investor protection concerns how minority shareholders are formally protected.

In a study by La Porte et al. (1998), the authors relate legal families to the level of investor protection and ownership concentration. They find concentration of ownership of shares in the largest public companies is negatively associated with investor protection, consistent with the hypothesis that small diversified shareholders are unlikely to be important in countries that fail to protect their rights. Their cross-sectional regression model is estimated under the assumption that ownership depends on the rule of law, so that the rule of law is treated as an exogenous variable. However, it is not implausible that the causation may run in the opposite direction; hence both rule of law and ownership are considered as endogenous variables. The reason is that powerful owners might seek to influence the degree of investor protection by lobbying the political process. For instance, large shareholders with sufficient political power may try to evade the protection of minority shareholders stipulated in the company laws. It may be plausible that in countries with high ownership concentration, majority shareholders such as influential families would try to undermine investor protection through the political process or by influencing the court system.

Ownership by Families and Foundations

Family ownership plays a major role, not only in privately held firms but also in listed firms where families such as the founder or his decedents hold major ownership stakes in listed firms. Sometimes they maintain control holding shares with superior voting rights, whereas the shares with less voting power are held by other investors,

e.g., institutional investors (see Rose 2008). Family ownership may facilitate long-term stable investors, but there is also a risk that the family may seek to enjoy private benefits at the expense of all the minority shareholders.

The relation between ownership and investor protection has also been analyzed in La Porta et al. (2002) who relate concentrated ownership with family ownership. They argue that resistance against the introduction of strong investor protection laws in some countries comes from families that control large corporations. They mention that "From the point of view of these families, an improvement in the rights of outside investors is first and foremost a reduction in the value of control due to the deterioration opportunities." Obviously, one cannot neglect the risk that, e.g., an incumbent family would seek control by occupying the board deriving private benefits of control. However, whether there exist severe agency costs associated with family control/control is in essence an empirical question. And, from this perspective, the recent evidence suggests the contrary. Anderson and Reeb (2003) conduct a substantial study of the firms in the S&P 500 where they show that family ownership is both prevalent and substantial, as family ownership is present in one third of the firms and accounts for 18% of outstanding equity. Specifically, they document that family firms perform better compared to non-family firms (also indication that the relationship is nonlinear). Moreover, the authors show that when family members serve as CEO, performance is improved compared to outside CEOs suggesting that family ownership may be an effective organization structure.

A new study by Isakov and Weisskopf (2014) also shows that family firms are more profitable than companies that are widely held or have a nonfamily blockholder. Their sample covers Swiss listed firms from 2003 to 2010, and the authors measure performance by Tobin's *q*. Specifically, they document that the generation of the family and the active involvement of the family play an important role for market valuation. Their sample covers Swiss firms and a country where investor protection is well developed. However, one cannot reject that in countries with

weaker investor protection, family ownership is not by per se positive for performance.

Foundation or trust ownership is widespread in several countries especially in Northern Europe, but it is also present in the Netherlands and Germany. Recent evidence does not associate foundation ownership with significant agency costs, even though the concept of foundation ownership seems to violate classical principal-agent theory (see, e.g., Thomsen and Rose 2004). A company founder may create a separate legal entity, i.e., a trust of a foundation, instead of transferring his shareholdings to his descendants. The foundation is governed by the board, but it has no owners, and the main purpose is to maintain control with the listed firm. The dividend proceedings are very often allocated to charitable purposes and/or the founder's descendants. The principal-agent model assumes that a principal hires an agent to conduct a task. Due to imperfect information, moral hazard problems may be created as the principal is not able to monitor the agent perfectly. Therefore, in order to align the interests between the principal, i.e., the owner and the agent, the latter is offered incentive contracts. However, in a foundation there are not any principals, since a foundation does not have any owners.

Managerial Ownership

Managerial control with the firm may be initiated by an MBO (management buyout). In such a situation the existing management acquires the shares which is financed by a high degree of debt. The literature concerning the impact of managerial ownership on firm performance does not offer a picture of unanimity.

Instead two divergence hypotheses exist. The so-called convergence of interest's hypothesis states a positive relationship between managerial ownership and firm performance. The underlying idea is to let a manager's remuneration depend more on the total wealth creation in the company by making him a residual claimant. As managers' stake rise, managers pay a larger share of the costs from activities that reduce firm value and therefore agency problems become less likely.

This stands in contrast to the *entrenchment hypothesis* that predicts a negative relationship between managerial ownership and firm performance. A manager who controls a substantial part of the firm's equity may be able to have sufficient influence to secure the most favorable employment conditions, including an attractive salary. One may argue that such benefits may also be obtained by his reputation, superior qualifications, or personality, independently of how much voting power a manager controls from his equity stake in the firm (see, e.g., Rose 2005).

Thus, even if managerial equity stake in a firm is low, there might be other forces to discipline managers away from opportunistic behavior such as competition in product markets, the managerial labor market. But at higher levels of managerial ownership, managerial entrenchment blocks takeovers making them more costly, which eventually decrease firm value since the probability of a successful tender offer decreases.

The Increasing Role of Institutional Investors

The increase in institutional funds in the industrialized countries has been tremendously rapid within the last 30 years, where institutional investors are the largest owners managing an enormous amount of capital. As a result, institutional investors are as a group considered the most influential actor on the scenes of capital markets. In the debate over corporate governance and, in particular, the role of institutional investors, more pressure has been put on institutional investors to be more active in their ownership, e.g., to exercise their proxies and their voting rights at the firm's general meeting.

Institutional investors cover a wide group of heterogeneous investors, which are all subjected to different legislations. They include pension funds, banks, insurance companies, mutual funds, mutual companies, and investment funds/foundations. Some of the largest institutional investors have all been very active in exercising their rights as owners, e.g., CalPERS, New York City pension fund, and TIAA-CREF. They have

sought, e.g., to challenge excessive executive compensation, the adoption of takeover defenses, to split the roles of chairman and CEO, and to ensure enough independent directors. However, on overall, institutional activism has been limited. The reason is that regulation often puts various restrictions on the ownership by institutional investors, such as requiring them not to have a dominant stockholding in given firm.

Moreover, when institutional investors hold a substantial proportion of shares, this might discipline management, since the free rider problem associated with dispersed ownership would be alleviated. Contrary to small investors, institutional investors are more able to absorb the costs from monitoring management and engaging in active ownership. Specifically, institutional investors may reduce the free rider problem caused by dispersed ownership and therefore avoid managerial focus on short termism. However, one may argue that if all small investors believe that institutional investors will undertake the monitoring role, the free rider problem may be enhanced. The reason is that this would destroy the incentives for small investors to play any active role at all.

Proponents of institutional activism argue that strengthening institutional investor ownership would benefit society as a whole, because they would be able to influence managerial actions, so that the interests of the society and the company more coincide; see, e.g., Monks and Minow (2001) for this view. Moreover, it has been advocated that institutional investors may facilitate the promise of “relationship investing” (see, e.g., Blair (1995)), who describe a situation where they are engaged in overseeing management over the long term instead of being detached or passive.

Furthermore, one may hypothesize institutional investors could influence management, only to take the interests of shareholders into account but also to serve the interests of other stakeholders; see Rose (2007). To illustrate, consider, e.g., a pension fund, where all the members have strong preferences against firms that directly or indirectly use child labor. Even if such firms may earn a higher profit due to lower costs, it might be reasonable for a pension fund not to invest in such firms, due to the members’ strong

preferences against the use of child labor. In other words, institutional investors may consider a broader view, trying to get management not only to care about the shareholders’ interests, which to some extent can be justified, since shareholders are usually considered residual claimants (see, e.g., Fama and French 1983).

Moreover, one could argue that enhanced ownership by institutional investors does not necessarily influence performance positively. Specifically, it is doubtful that institutional investors act, as they have a long investment horizon. The reason is that a portfolio manager employed by institutional investors is evaluated yearly, by comparing each portfolio manager’s performance, with a selected peer group or benchmark. As a consequence, the portfolio manager might care less about the return from their investments in the future 30 years from now, i.e., when the proceeds are repaid to the pension customers. Put differently, there is an embedded agency problem within institutional investors.

Furthermore, it is also questionable whether institutional investors would always act in a way that benefits all investor groups. Naturally, management in listed firms needs to care about the preferences of the large shareholders, since they are the owners and could replace incumbent management at the forthcoming general meeting. If institutional investors hold a high stake in a company, there is an inherent risk that institutional investors might seek to derive private benefits on behalf of all the other minority shareholders. For instance, institutional investors might get inside information, when management holds investor meetings or is in contact with the dominant owners. Even though this is prohibited by law, it is still quite difficult to prove afterwards by the authorities. Collusion between large blockholders and management may be sanctioned by the law, since this could violate the principle of the equal treatment of shareholders that prevails in most countries’ legislation.

Some Key Empirical Contributions

Hartzell and Starks (2003) argue that institutional investors serve a monitoring role in mitigating the

agency problem between managers and shareholders. Specifically, they find that institutional ownership is positively related to the pay for performance sensitivity of executive compensation and negatively related to the level of compensations. The result is robust when they control for firm size, industry, and investment opportunities.

Duggal, R. and J. Millar (1999) empirically challenge the ability of institutional investors to monitor management. Based on takeover decisions in 1985–1990, they examine the impact of institutional ownership and performance, but they do not find evidence that active institutional investors, as a group, enhance efficiency in the market for corporate control. Duggal and Millar also identify a number of institutional investors that have a reputation for exercising an active ownership, but regressing bidder returns against active institutional investors only result in an insignificant relationship.

Wahal and McConnell (2000) find no support for the contention that institutional investors cause managers to behave myopically. Based on a large sample from 1988 to 1994 of US firms, they document a positive relation between the industry-adjusted capital expenditures, as well as research and development, and the proportion of shares held by institutional investors. Both are proxies for management's degree of long-term orientation.

Prevost and Rao (2000) study whether institutional investor activism benefits shareholders, using an event study of shareholder proposals surrounding proxy mailing dates. Contrary to earlier studies they find a strong negative wealth effect surrounding the proxy mailing dates of firms targeted by two very visible, publicity-seeking types of sponsors: CalPERS and coalitions of public funds sponsoring or cosponsoring one or more proposals on the same proxy. Prevost and Rao argue that the results are consistent with the hypothesis that a formal proposal submission signals a breakdown in the negotiation process between the funds and management.

Louis, Chan, and Lakonishok (1993) examine the price effect of institutional stock trading and they find that the average effect is small. They also

document market asymmetry between price impact of buys versus sells, which is related to various hypotheses on the elasticity of demand for stocks, the costs of executing transactions, and the determinants of market impact. For instance, they argue that institutional purchase might be a stronger signal of favorable information, whereas there are many liquidity-motivated reasons to dispose a stock; see also Sias and Starks (1997) for an analysis of return autocorrelation and institutional investors.

Bhagat and Black and Blair (2004) conduct a large study of ownership and performance over a 13-year period, focusing on whether relationship investing has a positive impact on firm performance. They document a significant secular increase in large-block shareholding with sharp percentage increase in these holdings by mutual funds, partnerships, investment advisors, and employee pension plans. However, most institutional investors, when they purchase large blocks, sell the blocks relatively quickly afterwards. Bhagat, Black, and Blair provide a mixed result of whether relational investing affects firm performance. In the late 1980s where there was a high takeover wave, there is a significant relation between relational investing and firm performance, but this pattern was not found in the other periods. In essence, they do not find any persistent and sustainable effect of relational investing on firm performance. Thus, they argue that the idea of relational investing must be more carefully specified in theory.

Ackert and Athanassakos (2001) focus on agency considerations within institutional investors. They show that market frictions are important concerns for institutional investors, when they make portfolio allocation decisions. The availability of information about a firm is a significant friction, so that institutional holding increases with market value and firm's visibility, as proxied by the number of analysts following the firm. They also show that institutions adjust their portfolios away from highly visible firms at the beginning of the year but increase their holdings in these firms as the year-end approaches, which is as they argue consistent with the gamesmanship hypothesis.

In contrast to several other studies that focus on firm-level effects of institutional ownership, Davis (2002) examines how institutional shareholding in the largest countries on aggregate level impacts macroeconomies. Specifically, Davis links the development of institutional investors to important indicators of corporate sector performance, such as increasing dividend distribution, less fixed investment, and higher productivity growth.

Anand et al. (2013) examine the impact of institutional trading on stock resiliency during the financial crisis of 2007–2009. A resilient market is defined as one where prices recover quickly after a liquidity shock. The authors focus on why financial markets stayed illiquid over an extended period during the 2007–2009 crisis. They show that liquidity suppliers withdraw from risky securities during the crisis, and their participation does not recover for an extended period of time. Moreover, the illiquidity of specific stocks is significantly affected by institutional trading patterns.

Institutional shareholders exercise their formal power by voting at the AGM. The agenda on the AGM consists mostly by the board's own proposals. However, there has been an increasing tendency of shareholder-initiated proxy proposals. Renneboog and Szilagyi (2011) study the role of shareholder proposals in corporate governance. They find that target firms tend to underperform and have generally poor governance structures with little indication of systematic agenda setting by the proposal sponsors. The authors also find that proposal implementation is largely a function of voting success but is affected by managerial entrenchment and rent seeking. According to the authors, their results imply that shareholder proposals are a useful device of external control which should not be legally restricted.

References

- Ackert LF, Athanassakos G (2001) Visibility, institutional preferences and agency considerations. *J Psychol Financ Markets* 2:201–209
- Anand A, Irvine P, Puckett A, Venkataraman K (2013) Institutional trading and stock resiliency: evidence from the 2007–2009 financial crisis. *J Financ Econ* 108:773–793
- Anderson RC, Reeb DM (2003) Founding-family ownership and firm performance. Evidence from the S&P 500. *J Financ* LVIII(3):1301–1328
- Bhagat S, Black B, Blair M (2004) Relational investing and firm performance. *J Financ Res* 27:1–30
- Blair M (1995) Ownership and control. Rethinking corporate governance for the twenty first century. The Brookings Institution, Washington, DC
- Davis EP (2002) Institutional investors, corporate governance and the performance of the corporate sector. *Econ Sys* 26:202–229
- Duggal R, Millar JA (1999) Institutional ownership and firm performance: the case of bidder returns. *J Corp Financ* 5:103–117
- Fama E, French K (1983) Agency problems and residual claims. *J Law Econ* 26:327–349
- Hartzell JC, Starks LT (2003) Institutional investors and executive compensation. *J Financ* 58: 2351–2374
- Isakov D, Wisskopf J-P (2014) Are founding families special blockholders? An investigation of controlling shareholder influence on firm performance. *J Bank Financ* 41:1–16
- La Porta R, de-Silanes L, Schleifer A, Vishny RW (1997) Legal determinants of external finance. *J Financ* 52:1131–1150
- Louis K, Chan C, Lokonishok J (1993) Institutional trades and intraday stock price behavior. *J Financ Econ* 33:173–199
- Monks RAG, Minow N (2001) Corporate governance, 2nd edn. Blackwell Business, Malden
- Porta L, Rafael FL-d-S, Shleifer A, Vishny RW (1998) Law and finance. *J Polit Econ* 61: 1113–1155
- Porta L, Rafael FL-d-S, Shleifer A, Vishny RW (2002) Investor protection and corporate valuation. *J Financ* 57:1147
- Prevost AK, Roa RP (2000) Of what value are shareholder proposals sponsored by public pension funds? *J Bus* 73:177–204
- Renneboog L, Szilagyi PG (2011) The role of shareholder proposals in corporate governance. *J Financ Econ* 17:167–188
- Rose C (2005) Managerial ownership and firm performance of listed Danish firms – in search of the missing link. *Eur Manag J* 23(5):542–553
- Rose C (2007) Can institutional investors fix the corporate governance problem? – Some Danish evidence. *J Manag Gov* 11:405–428
- Rose C (2008) A critical analysis of the one share – one vote controversy. *Int J Disclos Gov* 5(2): 126–139
- Shleifer A, Vishny RW (1997) A survey of corporate governance. *J Financ* 52:737–783
- Thomsen S, Rose C (2004) Do companies need owners? Foundation ownership and firm performance. *Eur J Law Econ* 18:343–363
- Wahal S, McConell JJ (2000) Do institutional investors exacerbate managerial myopia? *J Corp Financ* 6: 307–329

Confiscation Orders and Judicial Cooperation in the EU

Barbara Piattoli Girard

Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università degli Studi del Piemonte Orientale “Amedeo Avogadro”, Alessandria, Italy

Abstract

Following ongoing reflection and experience at European level, it is possible and necessary to reason about the main features of the EU's legal strategy in building a simplified procedure for the recognition of confiscation orders among EU countries, in order to avoid the different barriers to the effectiveness of the EU's regime on the confiscation of proceeds of crime. It's significant in this context to focus on the consequences of the principle of mutual recognition on the rights of individuals. The proposal for a new regulation on the mutual recognition of freezing and confiscation orders aims to amend the EU's regime and eliminate gaps, uncertainties that legal rules still present; however, its adoption might significantly improve effectiveness of the EU's action if the emphasis on legal solutions doesn't come at the expense of broader questions concerning the safeguards applicable to domestic criminal proceedings which are crucial to ensuring effective cooperation between Member States in recovery action.

EU Legal Strategy on Mutual Recognition of Confiscation Orders: Improving an Effective Cooperation Between Member States?

The confiscation and recovery of criminal assets have assumed a prominent position in the fight against organized and other profit-driven crime in the EU, and it is also an important tool to combat terrorist financing (COM(2016)50 final, Chapter 1.3). The terrorist attacks in 2015 and 2016 in the European Union and beyond

underlined the urgent need to prevent and fight terrorism through a rapid and cohesive action to disrupt terrorist financing and its close link with organized crime networks. This has reinforced the issue of a better coordination and cooperation at EU level in order to give effectiveness of the EU's legal strategy in building a genuine area of justice, which led to the adoption of several measures designed to improve confiscation across the EU Member States with the most recent instruments of mutual recognition (Lelieur 2015).

Through the Council Framework Decision 2006/783/JHA of 6 October 2006 concerning the application of the principle of mutual recognition of confiscation orders (together with the Framework Decision 2003/577/JHA of 22 July 2003 on the execution in the European Union of orders freezing property or evidence), the EU is intended to strengthen the direct execution of confiscation (and freezing) orders for proceeds of crime (Brown 1996) by establishing simplified procedures for recognition among EU countries and rules for dividing confiscated property between the country issuing the confiscation order and the one executing it. In addition the recent proposal for a regulation of the European Parliament and of the Council on the mutual recognition of freezing and confiscation orders (COM(2016)819 final), which will replace these two framework decisions, builds an existing EU legislation on mutual recognition considering the need of putting in place a comprehensive system for freezing and confiscation of proceeds and instrumentalities of crime in the EU (Kingah 2015). It addresses the fact that legislation on mutual recognition of freezing and confiscation orders leaves *lacunae*, and Member States have developed new forms of freezing and confiscation of criminal assets (Fazekas and Nanopoulos 2016).

The European Council, meeting in Tampere on 15 and 16 October 1999, stressed that the principle of mutual recognition should become the cornerstone of judicial cooperation in both civil and criminal matters within the union (Flore 2014; Vermeulen 2014). After the entry into force of the Lisbon Treaty, confiscation has been given strategic priority at EU level as an effective instrument to fight organized crime (Feraldo Cabana

2014). The purpose of the actual European rules is to facilitate cooperation between Member States as regards the mutual recognition and execution of orders to confiscate property so as to oblige a Member State to recognize and execute in its territory confiscation orders issued by a court competent in criminal matters of another Member State.

Mutual recognition instruments are linked to harmonization measures (Council Framework Decision 2005/212/JHA of 24 February 2005 on confiscation of crime-related proceeds, instrumentalities, and property and Directive 2014/42/EU of 3 April 2014 on the freezing and confiscation of instrumentalities and proceeds of crime in the EU) (Simonato 2015). Both types of instruments (harmonization measures and mutual recognition instruments) are necessary in order to have a functioning regime of recovery of criminal assets, and they complement each other. In parallel, efforts were made to strengthen the identification and tracing of the proceeds and instrumentalities of crime. Council Decision 2007/845/JHA provides for the establishment of Asset Recovery Offices in all Member States.

In particular, the Directive 2014/42/EU, which replaces certain provisions of Council Framework Decision 2005/212/JHA, provides uniform rules for domestic possibilities to freeze and confiscate assets. Whereas the Framework Decision 2005/212/JHA continues to apply to all criminal offenses punishable by detention of at least 1 year, regarding ordinary confiscation, the Directive could only cover the so-called Eurocrimes and set minimum rules for national freezing and confiscation regimes: it requires ordinary and value confiscation for Eurocrimes, including where the conviction results from proceedings in absentia. It provides rules for extended confiscation subject to certain conditions (Boucht 2013). It also enables confiscation where a conviction is not possible because the suspect or accused person is ill or has absconded (Alagna 2015). The Directive also enables the confiscation of assets in the possession of third parties. Finally, the Directive introduces a number of procedural safeguards, such as the right to be informed of the execution of the freezing order including, at least briefly, on

the reason or reasons, the effective possibility to challenge the freezing order before a court, the right of access to a lawyer throughout the confiscation proceedings, the effective possibility to claim title of ownership or other property rights, and the right to be informed of the reasons for a confiscation order and to challenge it before a court.

Playing field of the Framework Decision 2006/783/JHA is mutual recognition and direct execution of confiscation orders. It acknowledges the need to modernize cross-border cooperation legislation within the EU and at an international level providing Member States with a set of procedural rules for a coordinate action (for the execution in the European Union of orders freezing property or evidence, see Council Framework Decision 2003/577/JHA of 22 July 2003. Both instruments are based on the principle of mutual recognition and work in a similar way). The Framework Decision 2006/783/JHA requires confiscation orders issued in one Member State to be recognized and executed in another Member State. The orders are transmitted alongside a certificate to the competent authorities in the executing State which must recognize them without further formalities and take the measures necessary for their execution.

Like it happens for the European arrest warrant, in order to facilitate the procedure, the double criminality check has been abolished in relation to a list of 32 offenses, provided that these offenses are punishable in the issuing country by a custodial sentence of at least 3 years. For all types of crime other than those listed in the Framework Decision 2006/783/JHA, the executing country can continue to apply the principle of double criminality – that is, it can make recognition and execution of the order dependent on the condition that the facts giving rise to the confiscation order constitute an offense according to its law. In limited cases the executing country may refuse to recognize and execute the order if the certificate is missing or incomplete or does not correspond to the order (in accordance with the *ne bis in idem* principle, the same person has already been the subject of a confiscation order for the same facts, for immunities or privileges that prevent execution, and for the rights of the parties

concerned and third persons acting in good faith); if the judgment was given in the absence of the person concerned, unless he/she was informed of the date and place of the trial and that an order may be handed down regardless of his/her presence; if he/she was represented by a legal counsel; or if he/she did not contest the judgment nor request a retrial or an appeal within the set time limit, for the principle of territoriality.

According to the procedural rules, the confiscation order, together with a certificate of which a copy is annexed to the framework decision and must be translated into the official language of the executing country, or another official language of the EU as indicated by that country, will be sent directly to the competent authority of the EU country where the natural or legal person concerned has property or income, is normally a resident, or has its registered seat. If the issuing authority cannot identify the authority in the executing country that is competent to recognize and execute the order, it will make inquiries, including through the European Judicial Network. A written record of the transmission of the order must be available to the executing country, which checks that it is genuine.

The transmission of a confiscation order does not restrict the right of the issuing country to execute the order itself. Where appropriate, the competent authority in the executing country must be informed.

The executing country recognizes and executes the order forthwith and without requiring the completion of any further formalities. The order is executed in accordance with the law of the executing country and in a manner decided upon by its authorities. The amounts confiscated are disposed of by the executing country as follows: if the amount is below EUR 10000, it accrues to the executing country; if it is above that amount, 50% of it is transferred to the issuing country. Both the executing and the issuing country can grant a pardon or amnesty, while the issuing country alone is responsible for appeals on the substance of the case lodged against the order.

This EU's intervention based on the mutual recognition aims to give more effectiveness of the EU's asset confiscation regime. The

experience has shown that not all the Member States have transposed the framework decisions on mutual recognition of freezing and confiscation orders until now, and the level of transposition of these framework decisions into the national legal systems of Member States was not satisfactory. It has been complained about the underuse of confiscation in cross-border situations at judicial level with the current system (commission staff working paper accompanying document to the proposal for a Directive of the European Parliament and the Council of the freezing and confiscation for proceeds of crime in the European Union, SWD(2012) 31 final; Report from the Commission to the European Parliament and the Council of 22.12.2008 [COM(2008) 885 final; Report from the Commission to the European Parliament and the Council of 23.8.2010 [COM(2010) 428 final].

It appears the need of a more strategic action at EU level to improve mutual recognition of freezing and confiscation orders which is also relevant to address terrorist financing in a more effective manner. The European initiative COM(2016)819 of 21 December 2016 is a response to these calls, and it addresses the fact that Member States have developed new forms of freezing and confiscation of criminal assets. It also takes into account developments at EU level, including the minimum standards set out in Directive 2014/42/EU. Whereas the directive improves the domestic possibilities to freeze and confiscate assets, the proposal aims to improve the cross-border enforcement of freezing and confiscation orders. Together, both instruments should contribute to effective asset recovery in the European Union.

Opportunity and Possible Benefits of the New Proposal for an EU Regulation on Mutual Recognition of Freezing and Confiscation Orders

The proposed regulation, once adopted, will be directly applicable in the Member States. This provides greater legal certainty and avoids the transposition problems that the framework decisions on mutual recognition of freezing and

confiscation orders were subject to. A regulation does not leave a margin to Member States to transpose such rules with the result that orders issued by other Member States will have to be executed like domestic ones, without the need to modify their internal legal system and their way of working.

Another important issue is the extended scope of the instrument compared to Directive 2014/42/EU as the proposal improves provisions that ensure a wider circulation of freezing and confiscation orders imposed by a court following proceedings in relation to a criminal offense including non-conviction-based confiscation orders. So the regulation applies to all types of orders covered by Directive 2014/42/EU (including extended confiscation and third-party confiscation orders), and in addition, it will also cover orders for non-conviction-based confiscation issued within the framework of criminal proceedings: the cases of death of a person, immunity, prescription, and cases where the perpetrator of an offense cannot be identified, or other cases where a criminal court can confiscate an asset without conviction when the court has decided that such asset is the proceeds of crime. This requires the court to establish that an advantage was derived from a criminal offense. In order to be included in the scope of the regulation, these types of confiscation orders must be issued within the framework of criminal proceedings. Therefore all safeguards applicable to such proceedings will have to be fulfilled in the issuing State. This perspective could generate several problems as sometimes orders can be issued without all the safeguards applicable to criminal proceedings (European Court of Human Rights, 22 October 2009, *Paraponiaris c. Greece*, n. 42,132/06; see, *infra*, § 3). However, the regulation does not cover civil and administrative orders. For what concerns its object area, it is not limited to particularly serious crime with a cross-border dimension so-called “Eurocrimes” (unlike Directive 2014/42/EU which is based on 83 TFEU) as Article 82 TFEU (on which this proposal is based on) does not require such a limitation for mutual recognition of judgments in criminal matters.

Fundamental Rights

Asset recovery is assumed to have positive impacts to enhance redistributive justice for victims of crime, and the victim’s right to compensation and restitution has been duly taken into account in the proposal. It is ensured that in cases where the issuing State confiscates property, the victim’s right to compensation and restitution has priority over the executing and issuing States’ interest.

This issue underlined that freezing and confiscation measures may interfere with fundamental rights protected by the EU Charter of Fundamental Rights (the Charter) and the European Convention on Human Rights (ECHR). In particular, the European Court of Human Rights (ECtHR) has repeatedly considered non-conviction-based confiscation, including civil and administrative forms, and extended confiscation to be consistent with Article 6 ECHR and Article 1 of Protocol 1, if effective procedural safeguards are respected. Shifts of the burden of proof concerning the legitimacy of assets have not been found in violation of fundamental rights by the ECtHR, as long as they were applied in the particular case with adequate safeguards in place to allow the affected person to challenge these rebuttable presumptions (ECtHR 29 ottobre 2013, *Varvara c. Italia*; ECtHR 20 gennaio 2009, *Sud Fondi et al. c. Italy*; ECtHR 10 maggio 2012, *Sud Fondi e altri c. Italy*). Recently, ECtHR 23 February 2017, *De Tommaso c. Italia* focused on preventive measures, and the European Court reiterates its settled case law, according to which the expression “in accordance with law” not only requires that the impugned measure should have some basis in domestic law but also refers to the quality of the law in question, requiring that it should be accessible to the persons concerned and foreseeable as to its effects. “One of the requirements flowing from the expression ‘in accordance with law’ is foreseeability. Thus, a norm cannot be regarded as a ‘law’ unless it is formulated with sufficient precision to enable citizens to regulate their conduct; they must be able – if need be with appropriate advice – to foresee, to a degree that is reasonable in the circumstances, the

consequences which a given action may entail.” The Court reiterates that a rule is “foreseeable” when it affords a measure of protection against arbitrary interferences by the public authorities. A law which confers a discretion must indicate the scope of that discretion, although the detailed procedures and conditions to be observed do not necessarily have to be incorporated in rules of substantive law (see *Silver and Others v. the United Kingdom*, 25 March 1983, § 88, Series A no. 61).

This statement can also involve the determination of preventive measure concerning real property (“in rem”); they apply – if requested by the prosecutor and ordered by the judge – before, the conviction becomes final under special circumstances, and one of the requirements flowing from the expression “in accordance with law” is foreseeability. According to this perspective, the domestic law cannot be vague and excessively broad terms. The individuals to whom preventive measures are applicable and the content of these measures must be defined by law with sufficient precision and clarity.

Anyway, some important safeguards are included in the proposed regulation of the European Parliament and of the Council: the principle of proportionality needs to be respected; there are grounds for refusal based on the non-respect of the principle of “ne bis in idem” and the rules on “in absentia” proceedings. Moreover, the rights of bona fide third parties have to be respected; there is an obligation to inform interested parties of the execution of a freezing order including the reasons thereof and the legal remedies available, and there is an obligation for Member States to provide for legal remedies in the executing State. Furthermore, Article 8 of Directive 2014/42/EU includes a list of safeguards that need to be ensured by the Member States for those orders falling within the scope of the directive.

Finally, all criminal law procedural safeguards are applicable. This includes in particular the right to a fair trial enshrined in Article 6 ECHR and Articles 47 and 48 of the Charter. It also includes the relevant legislation at EU level on procedural rights in criminal proceedings:

- Directive 2010/64/EU on the right to interpretation and translation in criminal proceedings
- Directive 2012/13/EU on the right to information about rights and charges and access to the case file
- Directive 2013/48/EU on the right of access to a lawyer and communication with relatives when arrested and detained
- Directive 2016/343 on the strengthening of certain aspects of the presumption of innocence and the right to be present at one’s trial
- Directive 2016/800 on the procedural safeguards for children
- Directive 2016/1919 on legal aid for suspects and accused persons in criminal proceedings and for requested persons in European arrest warrant proceedings

However, Member States often have regulation which does not assure a coherent exercise of the right of defense, in particular if the confiscation order is issued in preliminary steps of the proceedings such as a summary judgment or a dismissal of the case (e.g., or an acquittal anticipated *ex art. 129 c.p.p.* and *469 c.p.p.* in the Italian procedural system). That lack of safeguards could be exploited for a refusal in certain procedural situations of a cohesive cooperation system for freezing and confiscation of proceeds and instrumentalities of crime.

The New Rules for the Procedure of Mutual Recognition

The procedure for recognition and execution of freezing and confiscation orders is regulated separately in the proposal to simplify direct application by competent national authorities. As it regards in particular confiscation orders, it provides for a direct transmission of a confiscation order between competent national authorities but also allows for the possibility of assistance by central authorities. The confiscation order must be accompanied by a standard certificate annexed to this proposal. The executing authority must recognize the confiscation order without further formalities and must take the necessary measures

for its execution in the same way as for a confiscation order made by an authority of the executing State unless it invokes one of the grounds for refusal or postponement. A list of offenses for which the mutual recognition and execution of freezing and confiscation orders cannot be refused based on dual criminality is the same as the list contained in other mutual recognition instruments with one exception only: point (y) of the list now reflects the existence of common minimum standards for combating fraud and counterfeiting of noncash means of payment (Framework Decision 2001/413/JHA). Dual criminality cannot be invoked for a list of offenses punishable by at least 3 years of imprisonment in the issuing State. In cases of offenses not included in this list, recognition can be refused if the crime to which the freezing or confiscation order relates is not a criminal offense under the laws of the executing State.

A list of grounds for nonrecognition and non-execution of confiscation orders on which basis the executing authority may refuse the recognition and execution of the confiscation order is laid down in Article 9. The list differs significantly from the list contained in the 2006 Framework Decision. Some grounds for refusal remain the same, e.g., the ground based on the principle *ne bis in idem* or the ground based on immunity or privilege. However, the grounds for refusal linked to the type of the confiscation order (e.g., extended confiscation) have not been included in the proposal thus considerably broadening and strengthening the mutual recognition framework.

Regarding the ground for refusal based on the right to be present at the trial, it only applies to trials resulting in confiscation orders linked to a final conviction and not to proceedings resulting in non-conviction-based confiscation orders. Detailed rules for a possibility to confiscate different type of property than the one specified in the confiscation order are laid down. The instrument lays down conditions for issuing and transmitting a freezing order to align the proposal with Article 6 of Directive 2014/41/EU, thereby ensuring that the same conditions apply to freezing for evidence and freezing for subsequent confiscation.

Conclusion

The proposal for a new regulation aims to amend the EU regime and to eliminate gaps, uncertainties that legal rules still present; however, its adoption might significantly improve effectiveness of the EU's action if the emphasis on legal solutions does not come at the expense of broader questions concerning the safeguards applicable to criminal proceedings which are crucial to ensuring effective cooperation between Member States in recovery action. It is significant in this context to focus on the consequences of the principle of mutual recognition on the rights of individuals and to consider the need, in particular in relation to non-conviction-based confiscation, of effective procedural safeguards within all the proceedings steps according to the ECtHR case law. This perspective will oblige Member States to rethink and to restyle certain domestic procedural rules to give real effectiveness of the EU's action, even if the proposed regulation, once adopted, will be directly applicable in the Member States.

Reference

- Alagna F (2015) Non-conviction based confiscation: why the EU directive is a missed opportunity. *European Journal on Criminal Policy and Research* 21(4):447
- Boucht J (2013) Extended confiscation and the proposed directive on freezing and confiscation of criminal proceeds in the EU: on striking a balance between efficiency, fairness and legal certainty. *European Journal of Crime Criminal Law and Criminal Justice* 21(2):127
- Brown AN (1996) *Proceeds of crime, money laundering, confiscation and forfeiture*. W. Green/Sweet&Maxwell, Edinburgh
- Fazekas M, Nanopoulos E (2016) The effectiveness of EU law: insights from the EU legal framework on asset confiscation. *European Journal of Crime, Criminal Law and Criminal Justice* 24(1):39
- Feraldo Cabana P (2014) Improving the recovery of assets resulting from organised crime. *European Journal of Crime, Criminal Law and Criminal Justice* 22:13
- Flore D (2014) *Droit pénal européen: les enjeux d'une justice pénale européenne*. Larcier
- Kingah S (2015) Measures for asset recovery: a multiactor global fund for recovered stolen assets. *World Bank Legal Review* 6:457
- Lelieur J (2015) Freezing and confiscating criminal assets. *European Union in Criminal Law Review* 5(3):279

- Simonato M (2015) Directive 2014/42/EU and social reuse of confiscated assets in the EU: advancing a culture of legality. *New Journal of European Criminal Law* 2:195
- Vermeulen G (2014) *Essential texts on international and European criminal law*. Maklu, Antwerp-Apeldoorn

Conflict of Interest

Remus Valsan

Edinburgh Law School, University of Edinburgh,
Edinburgh, Scotland, UK

Definition

Fundamental to the notion of conflict of interest is the idea that someone's ability to exercise proper judgment is at risk of being affected by a personal interest or by a competing duty. These extraneous factors interfere with judgment not as ends that a decision-maker has in view but as factors that tend to influence the ends in view. The presence of such factors puts at risk the decision-maker's ability to evaluate the weight to be given to the relevant considerations on which the decision is based, irrespective of his desire to resist the temptation of self-interest. The main danger in a conflict of interest situation is the risk of unreliable judgment rather than corruption.

Introduction

A conflict of interest situation arises when a person who has a preexisting obligation to exercise judgment over the interests of another has a personal interest or duty that tends to interfere with the proper exercise of his judgment (Davis 1982). Conflicts of interest can arise in various situations where a preexisting duty of proper judgment exists. The legal relations that can create a situation of conflict of interest are not limited to positions with respect to which there are established rules against conflicts, such as per se fiduciaries, members of professions, or public officials (Norman and MacDonald 2010). The rules against conflicts

of interest applicable to these roles are more visible because, on the one hand, such roles involve exercise of professional judgment or official discretion and, on the other hand, the maintenance of a good public image of such office holders is essential (Valsan 2016).

A situation of conflict of interest should be distinguished from a situation of conflicting interests. The former is a conflict between interest and judgment duty, whereas the latter is an opposition between the individual interests of parties to a legal relation. The two categories have distinct legal regimes. In private law, conflict of interest situations are governed by the law of fiduciary duties. Conflicting interest situations are pervasive, but the law leaves it largely to the parties to adjust their idiosyncratic interests through bargaining. The law protects the effectiveness of the bargaining process through several doctrines, such as good faith, undue influence, unconscionability, or disclosure obligations. The fundamental aims of the legal doctrines governing the two types of conflict are, thus, significantly different. While in a conflict of interest situation the law aims to protect the unencumbered judgment of the decision-maker, in a conflicting interest situation, the law aims to establish a certain level playing field between contracting parties.

Actual, Potential and Apparent Conflicts of Interest

A conflict of interest is actual if the decision-maker has a conflicting interest or duty with respect to a certain judgment that he must make. A conflict of interest is potential if the decision-maker has a conflict of interest with respect to a certain judgment but is not yet required to make that particular judgment (Davis 1998) or if a real sensible possibility of conflict exists (Valsan 2016).

Actual or potential conflicts of interest should be distinguished from situations that only give the appearance of a conflict of interest. A conflict of interest is apparent if there is no actual or potential conflict, but a third party may erroneously believe that a conflict exists. Appearances of conflict cannot, by themselves, indicate the existence of

a conflict. The outward impressions or indications that a person's actions produce are often a matter of the beholder's subjective perception (Davis 1998).

The distinction between apparent conflicts, on the one hand, and actual and potential conflicts of interest, on the other, is important as concerns the actions that decision-makers must take when faced with these situations. Apparent conflicts, although posing no actual or potential threat to the decision-maker's judgment, should nevertheless be clarified, for the same reason for which any apparent wrongdoing is objectionable. If the decision-maker becomes aware of appearances of conflict of interest relating to his activity, he must eliminate them by making available enough information to show that there is no actual or potential conflict. In the case of professionals or public officials, the obligation to dissipate appearances of conflict is justified by the damage that such appearances cause to public confidence in the profession as a whole (Davis 2001).

The Risk of Impaired Judgment

The legal rules governing conflicts of interest aim to eliminate the risk of unreliable judgment. In certain cases, such as public officials or members of a profession, the rules serve the added purpose of maintaining the appearance of propriety of judgment, in order to preserve the public confidence in the respective offices or professions (Valsan 2016).

The concept of judgment or discretion denotes the absence of a predefined algorithm based on which a decision can be modelled. In a situation requiring the exercise of judgment, the specification of the problem to be solved or the ends to be achieved are contested or open to interpretation. In contrast, decisions that do not require judgment are routine, mechanical, or ministerial. They require only technical rationality, in the sense of applying specific techniques or theories to achieve predefined unambiguous goals. Given the absence of a predefined pattern regarding the ends to be attained and the means to achieve them, the exercise of judgment goes beyond mechanical rule-following. When a decision requires judgment, different decision-makers may disagree on the ends to be pursued or on the optimal course of

action, without anyone being wrong in an objective, measurable sense (Davis 1993, 2001).

When a person has an obligation to exercise judgment over the interests of another, the risk of an impaired decision-making process is not only pervasive, but also difficult to detect or correct. The literature on cognitive and motivational biases shows that the influence of an extraneous interest over the decision process is seldom a matter of deliberate choice. Information processing relating to self-interest is relatively effortless and unconscious, whereas the processes governing professional judgment are more analytical and more effortful. When professional judgment and self-interest point in opposite directions, self-interest often prevails, even when decision-makers consciously attempt to comply with their professional requirements (Moore and Loewenstein 2004). Since self-interest considerations are governed by automatic processes that tend to occur outside of conscious awareness, conflict of interest situations pose a problem even when the decision-maker does not exploit them in corrupt ways. Conflicting personal or professional interests can impair the judgment of even the most conscientious and devoted expert by influencing the way in which a decision-maker evaluates the seriousness of various risks, the desirability of certain outcomes, or the perception of connections between cause and effect (Norman and MacDonald 2010). Consequently, conflicts of interest are reprehensible not so much because they create a measurable bias but because they create an unusual risk of error of judgment, which cannot be adequately measured and corrected (Davis 1998).

Since perturbing interests affect the decision-making process as factors that tend to influence the ends in view, the extent of the effect of such interests on one's judgment cannot be assessed based on the actual decision taken. The decision-maker has discretion, in these sense of authority to decide the appropriate course of action, so one cannot simply measure the influence of the perturbing interest by comparing the decision adopted with an objectively right decision. Since the effect of a conflict of interest cannot be assessed based on results, the legal rules governing conflict of interest focus on certain kinds of identifiable interests that are particularly

threatening to the exercise of judgment, such as material interests or family ties. The categories of interfering interests, however, are not closed (Davis 2001).

Not all factors that might compromise one's judgment can be regarded as interfering interests. First, factors that affect a decision-makers' competence, such as the depth and accuracy of information used to adopt a decision, may be more relevant to the duty to exercise due skill and care than to conflicts of interest. Second, exercise of discretion allows, by definition, a certain degree of subjectivity in the decision-making process. The combination of personal characteristics that is specific for each decision-maker accounts for the diversity of equally valid results that can occur in a situation involving discretion. The line between legitimate factors that influence the decider's judgment and factors that have the ability to create a conflict of interest is sometimes blurry. Ultimately, what constitutes a conflict of interest in a particular situation is an empirical question (Stark 2000).

Managing Conflicts of Interest

People have an imperfect understanding of the effect of self-interest on their judgment and of the optimal way of correcting this influence. Individuals have little insight into their cognitive processes and may thus be unable to detect if, and to what extent, their judgment reflects self-interested motivations (Nisbett and Wilson 1977).

People tend to either underestimate or overestimate the biasing effect of self-interest on themselves. Because they cannot have an objective understanding of the effect of self-interest on their decision-making, people may tend to think that they can resist the effects of self-interest without their judgment being affected. They may be inclined to discount self-interest as their own motivation and overestimate the role of self-interest in motivating other people (Pronin 2006). The opposite is also possible. When people are aware of a situation where self-interest could plausibly intrude on their judgment, they may assume that the bias exists and is influencing them. In such cases, the more committed a decision-maker is to fairness and objectivity, the

more likely he is to over-compensate for the presumed bias of self-interest, thus undermining the reliability of his decision (Kahneman et al. 1982).

Given the inability to measure accurately the effects of self-interest on judgment, the need to devise an effective response to a conflict of interest situation arises. Avoidance of conflicts is one potential solution. Persons having a duty to exercise judgment in the interest of another must avoid situations in which their interests pose an actual or potential threat to the reliability of their judgment. Although avoidance of conflict situations is an important duty of decision-makers, a flat prescription to avoid all conflicts of interest is undesirable. On the one hand, not all conflicts of interest are avoidable. Some conflict situations are embedded in the relation, while others arise independently of decision-maker's will. On the other hand, the mere fact of being in a situation of conflict is not always wrong from a legal or ethical perspective. Failure to address the conflict situation, however, is often reprehensible (Davis 1982).

Another potential strategy is to disclose the conflict to those relying on the conflicted person's judgment. Common sense suggests that complete disclosure will give the beneficiary of disclosure the opportunity to give informed consent to the situation of conflict, to adjust reliance accordingly, or to replace the decision-maker. Except the latter scenario, disclosure is an effective response if it does not affect the decision-maker's judgment process, or, alternatively, if the beneficiary is able correctly to adjust to the risk of impaired judgment. Psychological research shows that neither of these conditions may be met. Sometimes both parties may be worse off following disclosure (Cain et al. 2005).

Disclosure may have the unintended consequence of liberating a decision-maker from concerns about ethicality and give him a moral license to incorporate the conflicting interest into the decision-making process. Moreover, knowing that the beneficiary is likely to discount the decision to correct for the self-interest, the decision-maker may be tempted to counteract this adjustment by allowing self-interest to influence their decision even further. In addition, beneficiaries of disclosure do not always adjust to counteract the self-interest. In some cases, they see disclosure as

a sign of the decision-maker's trustworthiness and may increase their confidence in the latter's judgment (Cain et al. 2005).

Another potential response to the problem posed by a conflict of interest is to escape the situation. The decision-maker can escape a conflict by redefining the scope of the relationship, so that the scope of the judgment is restricted, by divesting himself of the interest creating the conflict or, where possible, by withdrawal from the relationship (Davis 1982). Divestment of the conflicting interest, however, may not be entirely effective. A person who in the past had an interest in the outcome of a decision cannot be said to be in the same psychological position as someone who was never interested in that matter (Miller 2005).

All potential responses to a situation of conflict of interest have their specific costs and benefits, and the law does prescribe an optimal response. In devising a strategy to manage conflicts, two guiding principles should be observed. First, a mere instruction to abstain from the temptation of self-interest is a mistaken response to a conflict of interest. Second, disclosure and consent do not put the parties in the same situation as when no conflict exists (Valsan 2016).

Conclusion

Personal interests or duties can affect the judgment of even the most honorable and disciplined decision-maker. A person has a conflict of interest on the basis of being in a conflicted situation, irrespective of that person's belief that he is capable of resisting the temptation or corrupting influence of the interest that could interfere with his judgment. Consequently, prescribing ethical self-restraint is a misguided solution to the conflict. Decision-makers must take active steps to steer clear of situations of conflict, to manage unavoidable ones, and to dissipate the mere appearances of conflict.

Cross-References

► [Fiduciary Duties](#)

References

- Cain DM et al (2005) The dirt on coming clean: perverse effects of disclosing conflicts of interest. *J Leg Stud* 34:1
- Davis M (1982) Conflict of interest. *Bus Prof Ethics J* 1:17
- Davis M (1993) Conflict of interest revisited. *Bus Prof Ethics J* 12:21
- Davis M (1998) Conflict of interest. In: Chadwick R (ed) *Encyclopedia of applied ethics*, vol 1. Academic Press, London, p 589
- Davis M (2001) Introduction. In: Davis M, Stark A (eds) *Conflict of interest in the professions*. Oxford University Press, Oxford, p 3
- Kahneman D et al (eds) (1982) *Judgment under uncertainty: heuristics and biases*. Cambridge University Press, Cambridge
- Miller DT (2005) Psychologically naïve assumptions about the perils of self-interest. In: Moore DA et al (eds) *Conflicts of interest: challenges and solutions in business, law, medicine, and public policy*. Cambridge University Press, New York, p 126
- Moore DA, Loewenstein G (2004) Self-interest, automaticity, and the psychology of conflict of interest. *Soc Justice Res* 17:189
- Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. *Psychol Rev* 84:231
- Norman W, MacDonald C (2010) Conflicts of interest. In: Brenkert G, Beauchamp T (eds) *The Oxford handbook of business ethics*. Oxford University Press, Oxford, p 441
- Pronin E (2006) Perception and misperception of bias in human judgment. *Trends Cogn Sci* 11:37
- Stark A (2000) *Conflicts of interest in American public life*. Harvard University Press, Cambridge
- Valsan R (2016) Fiduciary duties, conflict of interest and proper exercise of judgment. *McGill Law J* 62:1

Conflict of Laws

Amit M. Sachdeva
Ernst & Young LLP, Houston, TX, USA
New York University School of Law (NYU),
New York, USA
London School of Economics (LSE), London, UK
The Hague Academy of International Law,
The Hague, The Netherlands

Definition

The *Conflict of Laws*, also known as *Private International Law*, is that branch of the domestic

law of a State which deals with cases having a foreign element. (The word “State” has been used in the sense of a political system or a political subdivision having a separate and distinct legal system.)

All elements of any dispute can broadly be classified into those related to the parties and those related to the subject matter. When operating in a purely domestic scenario, i.e., when both/all the parties to the dispute hail from the same State and where the subject matter arises entirely within the territorial limits of the State, legal rights and obligations of the parties are required to be determined on the basis of substantive and procedural standards provided under the law of that State.

The position is, however, fundamentally different where, at least, one of the parties to the dispute is a foreign party or where the subject matter of the dispute arises, at least, partly outside the territory of the State whose courts are called upon to adjudicate it. In such situations, before proceeding to resolve the dispute and determine the rights and obligations of the parties, the court must first determine two important questions: (a) whether it has jurisdiction to resolve the dispute (“international jurisdiction question”) and, (b) if so, by reference to the law of which State must the substantive rights and obligations be determined (“applicable law question”).

These two questions, together with the question whether a judgment rendered by a foreign court may be recognized and/or enforced by the domestic court to which it is presented (“recognition question”), constitute, in very broad and general terms, the subject matter of the *Conflict of Laws*.

Scope, Purpose, and History

Narrower and Broader Scope

According to some scholars, the narrower understanding of the term “Conflict of Laws” dictates its use as limited only to the applicable law question inasmuch as while answering the applicable law question, the courts come across and resolve the “conflict” between “laws” of different States which may potentially govern the issue in dispute.

Parties to a transnational dispute, in most cases, stand to gain from whether a court exercises or declines to exercise jurisdiction and whether the substantive law of one State is preferred to the other. Similarly, a party with a favorable judgment handed down by a foreign court seeks to enforce its rights by means of recognition and enforcement of the (favorable) foreign judgment in other States than assuming the risk of litigating in each State. Thus, while “laws” of different States do not per se conflict – each being a comprehensive and complete system and regime in itself – the possibility of succeeding in a dispute induces significant disagreements between the parties about each of these three questions. The branch of domestic law of a State that concerns itself with answering any or all of these questions is, thus, called the *Conflict of Laws*. It is in this broader sense that the term is used here.

Need, Object, and Purpose

Since the earliest times, changes in kingship and integration of smaller kingdoms into empires resulted in (peaceful) coexistence of the newer with the older legal systems. Further, commercial dealings by men resulted in inevitable interaction between various legal systems which oftentimes gave rise to mutually incompatible rights and obligations.

In more recent times, with the growth in globalization and international trade, movement of persons, capital, labor, and resources has increased between States, resulting in courts being required to deal with matters involving foreign elements more often than not. This has necessitated the growth of *Conflict of Laws*.

The purpose of the conflicts rules is threefold:

- (a) To define and confine the extent to which the laws of one State may extend extraterritorially.
- (b) To resolve the disputes with a foreign element by exercise of international jurisdiction and by application of that law which results in substantial justice between the parties.
- (c) To ensure that the litigation outcome is least distorted regardless of the State where the litigation is brought and that the substantive rights and obligations of the parties are adjudicated in the same manner.

In short, therefore, the object of the court in applying the conflicts rules is to reach at the outcome of a dispute which would have been reached in the jurisdiction and legal system to which it naturally and properly belonged.

History and Development

The history of *Conflict of Laws* goes back to a few centuries. The earliest history goes to the thirteenth-century universities in Italy during the times of the glossators and commentators, in particular Bartolus and Baldus.

A largely unbroken history of this branch dates back to the contributions of Charles Dumoulin's party autonomy and Bertrand d'Argentré's territorialism theories in France in the sixteenth century and Ulrich Huber in the Netherlands in the seventeenth century. The writings of the Dutch Huber were in the background of the "territorial sovereignty" thesis of the French Jean Bodin, whose thesis was strongly endorsed by the Dutch Hugo Grotius. Much of Huber's efforts and work focused on reconciling the application of foreign law with this overarching principle of territorial sovereignty, much in the interest of the Dutch being one of the leading international traders of that time.

The Dutch, unlike the Italians and the French scholars in the past, viewed the necessity of conflict arising out of the potential application of more than one competition legal system and are attributed the name "conflictus legum," directly translated as "Conflict of Laws." Huber, like Paul Voet, explained the resolution of this conflict sometimes resulting in the application of foreign law on the principle of "comity," a principle that apparently did not, in any manner, subdue the principle of territorial sovereignty. Huber laid the following three propositions which had unparalleled influence on the development of this branch of law in the common law:

- (a) The laws of each State have effect within the territory of the sovereign, but not beyond.
- (b) The laws of a State are binding on every person who is within, or enters, whether permanently or intermittently, the territory of the sovereign.

- (c) On the ground of comity, a sovereign will allow rights acquired under the laws promulgated by other sovereigns to retain their force, so long as they do not prejudice the rights or powers of the sovereign or of its subjects.

These three propositions and, in particular, the third expatiated the rationale behind the application of foreign law by States as also carved out two important exceptions, which later more formally came to be referred as public policy and mandatory rules of the forum.

Subsequently, Friedrich Carl von Savigny, a German scholar, proposed to scientifically deal with the problem of Conflict of Laws and develop a scheme of yielding rules of universal application. He focused on the legal relationships and sought to seat each legal relationship to a jurisdiction to which it related. This resulted in neutral, ready-to-apply, and common system of conflicts rules that would potentially yield identical results regardless of the forum, forming the present-day system of *Conflict of Laws*.

Relationship with Public International Law

Public International Law constitutes the body of rules which govern the relationship between different States and normally sets the standard by which the validity of the conduct of a State is evaluated. This body of rules is a legal system in itself, independent of the municipal legal systems of the States. It may be noted however that there is a difference between the monist and the dualist perception of single or separate legal systems. Public International Law is embodied either in treaties or as customary international law.

On the other hand, *Conflict of Laws* is a part of the domestic law of the State that merely facilitates the domestic courts to resolve a particular category of disputes, i.e., those involving a foreign element. There is thus nothing more "international" about *Conflict of Laws* or *Private International Law* than that it governs issues of international jurisdiction, applicable law, and enforceability of foreign judgments in situations

that are not purely domestic. The conflicts rules are embodied in domestic statutes, court decisions, or jurisprudence and sometimes in treaties.

The Method of Conflict of Laws

The approaches and answers to the three questions forming the scope of the *Conflict of Laws* vary significantly. The following part offers a slightly detailed insight into the questions and offers an overview of some of the solutions that different States have to these questions.

International Jurisdiction Question

This question assumes particular significance because in an international context the party proposing to sue will, invariably, have at least two (and sometimes many) jurisdictions to choose from.

Since the law of the forum where the dispute is litigated determines its procedures as well as the applicability of mandatory rules and public policy, this question can at least in some cases be outcome determinative. The forum State's treaty network or reciprocity relations also have bearing on the enforceability of its judgments. Its importance can be gauged from the fact that in practice, the process of choosing the forum for litigation involves substantial time for a transnational lawyer.

While rules of international jurisdiction vary between States, there are some that are more general and common. Often, domestic rules on the point confer jurisdiction on their courts on the basis of the "link" that the defendant – the party being sued – has with the State. Some States accept "cause of action" as a basis for jurisdiction. In addition, most courts, in the present-day world, assume jurisdiction where conferred by the parties to the dispute.

Thus, where a defendant is domiciled or has the principal place of business in a State or the defendant has "minimum contacts" with the State, courts of that State have jurisdiction. This type of jurisdiction is called as "personal jurisdiction." Some States also assume jurisdiction on the basis of the nationality of the plaintiff, service of

court process/writ, and the situation of defendant's property.

The court must also have "subject matter" jurisdiction that constitutes in some States an independent basis of jurisdiction and in some complements personal jurisdiction. Courts in several States entertain matters where the "cause of action" either wholly or partly arose within their territory.

Lastly, many courts, accepting the political and economic realities of this globalizing world, respect "consent-based" jurisdiction, i.e., where the parties either have a specific agreement that a chosen court shall deal with the dispute – called a "choice of court clause" – or, absent such agreement, the defendant does not, within a certain time, dispute or object to the jurisdiction of the court where the plaintiff has brought an action.

Common law countries significantly differ from the civil law countries in terms of "when" to exercise or decline jurisdiction. Common law courts have traditionally enjoyed much wider discretion in declining jurisdiction, even when all the prescribed requirements are otherwise satisfied. This discretion to decline the exercise of jurisdiction is exercised under different principles such as *forum non conveniens*, *lis alibi pendens*, foreign choice of court clauses, arbitration clauses, or abuse of court process (forum shopping). Apart from this, some common law courts are also known to issue "anti-suit injunctions" directed at the defendant (as opposed to the foreign State) preventing him from initiating or pursuing litigation in a foreign court, where the court is convinced that by resorting to another court, the defendant is likely to cause serious prejudice and manifest injustice to the plaintiff.

Applicable Law Question

The process of choosing the applicable law generally involves steps that are intricate and complex.

A typical choice of law analysis begins with the identification and characterization of the issue leading to the dispute. This requires the courts to identify, as a matter of domestic law, as narrowly and precisely as possible, one or more issues involved in the dispute. Once this is done, each

issue is “characterized” into one of the several known (or proposed) categories to which the choice of law rule must be applied. Thereafter, by applying the “connecting factor” prescribed by the concerned conflicts rule following from the characterization, the applicable law is determined.

Connecting factors or *points de rattachement* are those facts pertaining either to the parties or to the subject matter of dispute that tend to create nexus between the parties or dispute on one hand and a legal system on the other. In other words, connecting factors are those aspects which localize a controversy to one or more legal systems. As a generally accepted rule, the choice and prescription of connecting factors is done as a matter of the law of the forum, *lex fori*. Like any set of facts in all legal disputes, each fact carries different normative weight and depending upon the nature of the issue, different connecting factors may have lesser or greater “connecting” strength. For example, the domicile or residence of a contracting party is likely to be far less relevant while determining the governing law of contract than while determining the law governing the succession of movable property of a deceased.

While by no stretch of imagination exhaustive, the following is a list of conflicts rules relevant to different connecting factors that are most often-times used for determining the applicable law:

- *Lex fori*: law of the State where the forum/court is situated
- *Lex domicilii*: law of the State where the party is domiciled
- *Lex patriae*: law of the State of which a party is a national/citizen
- *Lex loci contractus*: law of the State where the contract was made
- *Lex loci solutionis*: law of the State where the contract is required to be performed
- *Lex delicti*: law of the State where a tort was committed
- *Lex loci celebrationis*: law of the State where a marriage was celebrated
- *Lex situs*: law of the State where a property is situated

An illustration would bring out the discussion more succinctly: Assume that XYZ A.G., a German holding company of a worldwide group, enters into a contract in Munich, Germany, with ABC Inc., a US corporation for sale and supply of x units of widgets, to its UK subsidiary, XYZ SubCo plc, for ultimate sale to consumers by the subsidiary in the UK. Now suppose that the goods are of an inferior quality and the action is brought in Germany. The rules by reference to which the German court in such a situation would determine the applicability of the law of one State in preference to the other is the subject matter of the applicable law question. The German court would first identify the issue between the parties: a contractual claim for sale of non-conforming goods. The court would then apply the conflicts rule appropriate to the concerned connecting factor. For example, if the German conflicts rule for determining the governing law in cases of contractual claims is that a contract is governed by the law of the place where it is made, the UK court would apply the German law.

In arriving at the applicable law, one comes across many difficult situations. Some of them are briefly highlighted here.

Characterization

Characterization of an issue into one as opposed to another may significantly change conflicts rule (and, consequently, the law) to be applied.

For instance, assume that under a legal system, two conflicts rules are known: (a) a contract is governed by the law of the State with which it is most closely connected, and (b) a tort claim is governed by the law where the tort is committed. Now suppose that Mr. X, a German national and domicile, purchased, while in Paris, a laptop from a French company, ABC, SA, for which Mr. X paid ABC in euros. Assume further that Mr. X carried the laptop with him to a seminar in India where the laptop short-circuited resulting in destruction of his research thesis. Mr. X now brings a claim against ABC in a court in Paris in France. If the short circuit is characterized as a contractual claim, then the French law (being the law to which the contract *ex facie* appears to be most closely connected) would apply. On the

other hand, if the French court were to characterize the claim as one arising in tort, it must apply the Indian law, being the law of the place where the tort was committed.

A related question which arises is: by reference to which law must the characterization of the issue be determined? Scholarship is fairly equally divided between those who advocate that characterization must be governed as a matter of domestic law of the State in which the court/forum is located (*lex fori*) and those who support the view that characterization itself must be governed by *lex causae*. Some recent thinkers urge for an internationalist approach to characterization.

Applicable Law: Internal or Conflicts Rules

The function of the *Conflict of Laws* however does not stop at identifying the “applicable law.” Before the rights and obligations of the parties are determined by reference to the substantive and procedural rules prescribed in the municipal system of a State, the court must also determine the further question whether the applicable law identified by employing the appropriate conflicts rule refers directly to the internal law of that State or does it refer to the internal law as well as the conflicts rules of that State.

Counterintuitively, most legal scholarship and court decisions endorse the view that the applicable law includes not only the domestic law but also the conflicts rules of that State. This is founded on the reasoning that the dispute between the parties must be determined as if it was presented to the court whose law is determined to be the applicable law and then by resolving the dispute in the same manner as that court applying its own law (including the conflicts rules) would resolve.

Renvoi

Since the applicable law includes also the conflicts rules under the foreign law, it may potentially lead to a situation where on application by the court of the conflicts rules under the foreign law, it is determined that the dispute must be resolved by reference to the “law” of a third State or, in some cases, the law of the forum. “Law” of the third State or of the forum itself would constitute of domestic law as well as

conflicts rules. The tendency of the court to apply the conflicts rule at this (and each subsequent) stage may result in either a ping-pong ball game or a relay race.

These “ping-pong ball” and “relay race” games are popularly known as *renvoi* and can potentially be avoided in many ways. One of the ways is by directly applying the internal law of the State whose law is determined to be the applicable law, i.e., applicable law may be understood as “applicable internal law.” Another possibility could be to stop the reference to “law” as inclusive of conflicts rules after the first level, i.e., reference to “law” under the conflicts rules of the “applicable law” may be understood as the internal law, exclusive of the conflicts rules. Lastly, the court may consider the treatment of *renvoi* under the applicable law determined by applying the conflicts rules of the forum and act accordingly. *Renvoi* presents one of the most complicated and difficult issues in the area of *Conflict of Laws* and is laced with lack of predictability and judicial discretion.

Incidental Questions and Time Factor

Complications in determining the applicable law compound many folds when there is more than one interconnected issue. Referred to as the “incidental question” or “question préalable,” it deals with ascertaining the legal system whose conflicts rules must determine the law governing the incidental question(s), i.e., whether the incidental question(s) must be determined by applying the conflicts rule of the law governing the main question or must the incidental question be determined by independently applying the conflicts rules of the forum.

Similarly, where either the conflicts rule or the content of the connecting factors or the applicable internal law undergoes change, serious difficulties arise about applying the altered rules to facts and circumstances prevailing before the change. Rare though, these questions have no easy answers.

Exclusion of Foreign Law

Courts however do not always apply the laws of other States, even when their conflicts rules lead to their application.

Called by different captions, a court would resist any temptation to apply foreign law, when doing so would either offend its public policy or conflict with its mandatory rules. Public policy or *ordre public* and mandatory rules represent those rules or values which, in a legal system, enjoy very high normative strength and, thus, have an overriding effect over any rule or norm to the contrary, including therefore those set by or under foreign law. This exception is broad based and controversial and one on which consensus is often difficult. It is pregnant with uncertainty, subjectivity, and judicial discretion.

Another exception that is widely accepted is enforcement of foreign penal and revenue laws. The rationale behind this exception is that penal and revenue laws give rise to “public,” rather than “private,” rights and liabilities and involve the assertion of sovereignty by a (foreign) sovereign over its subjects.

Recognition and Enforcement of Foreign Judgments

Under the traditional understanding of territorial sovereignty, a judgment rendered by a court has force within the territory of the State, but not beyond.

The “recognition” question constitutes the third and last scope of *Conflict of Laws*. Jurisprudentially, “recognition” and “enforcement” are different concepts. Enforcement requires positive action on behalf of a State to implement a foreign judgment, whereas recognition contemplates mere respect. A sword is to enforcement, what a shield is to recognition.

The recognition rules provide for the degree of finality a decision rendered by a court abroad enjoys in the State. Whether a decision rendered by a foreign court is conclusive of the rights and obligations of the parties and whether the parties are eligible, under certain circumstances, to re-litigate the issue before the courts of the concerned State are concerns that are addressed very differently by different States. The rules of different States are too disparate to admit any generalization. Suffice to say, absent an international treaty, the domestic rules that occupy the field reflect the peculiar historical development

and geopolitical relations that the State whose court is petitioned to recognize a foreign judgment has with the State whose court has rendered that judgment.

The Hague Convention on the Recognition and Enforcement of Foreign Judgments in Civil and Commercial Matters, 1971, made some unsuccessful efforts to harmonize the law on the point and has entered into force for only five States. Efforts revived in the last decade of the previous millennium but failed to achieve an overambitious harmonization and settled with a far more humble treaty on choice of courts and recognition and enforcement of judgments rendered by the chosen court(s). In terms of this convention, a judgment rendered by a competent foreign court and which is subject to no further review is required to be recognized and enforced, unless the case falls under one of the exceptions stated in Article 5 of that convention.

Harmonization Efforts

As highlighted above, since the difference in the contents of laws of different States lies at the bottom of the Conflict of Laws, States have undertaken harmonization efforts in order to mitigate the difference(s) in the outcome.

General

These efforts can be divided in two categories: those seeking to harmonize the contents of the substantive internal law of the countries and those endeavoring synchronization of the conflicts rules across States. The latter is, relatively speaking, simpler since it involves synergizing only one branch among many branches of the domestic laws of different States. Several multilateral treaties have been entered into that deal with very specific aspects, such as carriage of goods by rail, carriage of passengers and luggage by road, marriage, divorce, adoption, minors’ rights, nuclear energy, etc. Their formation and enforceability are, of course, matters of *Public International Law*.

European Union

Because of the peculiar supranational structure that the European Union is and because the EU

has the power to pass “regulations” in the area of *Conflict of Laws* that assume direct force as to both the text and effect in the Member States (with two exceptions), much of the conflicts rules have been harmonized by regulations at the EU level that have obliterated, as between the Member States, the operation of State conflicts rules.

In the EU, questions of jurisdiction pertaining to “civil and commercial matters” are governed by EC Regulation 44 of 2001, commonly called as “Brussels I Regulation.” The EC Regulation contains well-defined and clear rules of jurisdiction, targeted to result in uniformity and predictability. The former is buttressed by the mandatory reference procedure to the European Court of Justice, which issues binding rulings interpreting the Regulation. Similarly, Brussels II bis Regulation, EC Regulation 2201 of 2003, provides for rules of jurisdiction in matrimonial matters and matters of parental responsibility.

The EU has also been largely successful in evolving a uniform system for determining the applicable law. The law on the point is basically contained in two EC Regulations, popularly called Rome I Regulation (EC Regulation 593 of 2008) and Rome II Regulation (EC Regulation 864 of 2007) providing for rules for determining the law applicable to contractual and non-contractual disputes, respectively.

A judgment rendered by one Member State of the EU is recognized and enforced by other Member States under the rules embodied in the two Brussels regulations and several other EC Regulations concerning recognition of judgments in specific cases, such as those pertaining to uncontested claims (EC Regulation 805 of 2004), insolvency (EC Regulation 1346 of 2000), etc.

The harmonization (being) brought among the Member States of the EU is, without doubt, the best model of harmonization yet known.

Conclusion

The conflicts inhering from the differences in legal systems have historically impeded businesses. The *Conflict of Laws* reflects one of the

many where law, rather than facilitating economic activity, obstructs it. The impact (and sometimes deterrence) that the differences in legal systems on the economic efficiency is glaring and defies the most basic principles of *economics* and dictates further concerted action for harmonization.

References

- Bariatti S (2011) Cases and materials on EU private international law. Hart, Oxford
- Boele-Woelki K (2010) Unifying and harmonizing substantive law and the role of conflict of laws. Martinus Nijhoff, Boston
- Bogdan M (2006) Concise introduction to EU private international law. Europa Law Publishing, Groningen
- Briggs A (2008) The conflict of laws, 2nd edn. Oxford University Press, Oxford
- Collins L (ed) (2012) Dicey, Morris and Collins on the conflict of laws, 15th edn. Sweet & Maxwell/Thomson Reuters, London
- Esplugues C et al (2011) Application of foreign law. Sellier, Munich
- Fawcett J (ed) (1995) Declining jurisdiction in private international law. Oxford University Press, Oxford
- Hay P et al (2009) Comparative conflict of laws: conventions, regulations and codes. Thomson Reuters/Foundation Press, New York
- Stone P (2010) EU private international law, 2nd edn. Edward Elgar, Northampton
- Symeonides SC, Perdue WC (2012) Conflict of laws: American, comparative, international: cases and materials. West Thomson/Reuters, St. Paul

Consensus

Pavel Kuchař
Department of Economics and Finance,
University of Guanajuato, DCEA-Sede Marfil,
Guanajuato, GTO, Mexico
Facultad de Economía-División de Estudios de
Posgrado, Universidad Nacional Autónoma de
México, Ciudad de México, Mexico

Abstract

How does the process of consensus formation affect the accuracy and reliability of our knowledge? Cognitive and epistemic division of labor creates a problem of trust in the use

and application of knowledge. Consequently, the reliability of scientific consensus depends on whether the incentives, which the self-interested members of scientific communities face, are aligned in the right way.

Definition

Consensus is a conventional source of justified beliefs. In modern democratic societies, rational consensus, formed by means of free and open discussion, is a criterion of political legitimacy.

Introduction

Modern democracy as a form of government is a unique phenomenon, existing for only a short period of time in the history of our civilization. In this contribution, the role of consensus in directing the course of modern democracies is addressed. Public choice scholars have extensively studied the problem of amalgamating individual beliefs into aggregate social estimates for the purposes of legitimizing political authority. The aggregation of individual beliefs, however, is a specific example of a general question about the creation and use of knowledge. How does the process of consensus formation affect the accuracy and reliability of our knowledge? Without understanding how the division of labor brings together scientific communities and without understanding how these communities produce expert consensus, the criteria according to which we assess the accuracy and reliability of our warranted beliefs remain unclear.

According to literature, the reliability and accuracy of expert consensus depend on the nature of scientific institutions. In general, institutions are understood to be the rules of the game that determine the structure of payoffs for the agents involved. Given the cognitive and epistemic division of labor, expert consensus introduces a problem of trust. This problem is documented by empirical evidence that shows gaps between expert and nonexpert beliefs. If expert consensus results from the coordination of self-interested

individuals, its reliability is dependent on whether the incentives, which the self-interested members of scientific communities face, are aligned in the right way.

Determining the right way to align incentives is ultimately an empirical problem. Therefore, the study of the economics of scientific knowledge through a comparative institutional analysis should be of interest to anyone who is curious about the role institutions play in the creation and use of knowledge in modern society.

Is There a Consensus Among Economic Experts?

If we want to know how the process of consensus formation influences the accuracy of our beliefs, we should perhaps first look into the state of expert consensus today. Some claim that the scope of consensus among economists is overwhelming; others, however, find this contention questionable. Generally, there seems to be a consensus regarding mainstream economic theory. With regard to economic policy, however, the attitudes among economic experts often differ.

According to Mark Blaug, “nothing like an overwhelming consensus has emerged from . . . postwar economic methodology. But despite some blurring around the edges, it is possible to discern something like a mainstream view” (1992, p. 110). Today, this mainstream expert agreement can be summarized as follows: “Economists share the view that individuals are utility-maximizers, human wants are unlimited, and that mathematical modeling should be an important part of economic modeling” (May et al. 2013, p. 25). These are the basic building blocks of mainstream economic theory.

The scope of agreement among economists has been examined from the 1970s (Brittan 1973), and the surveys and expert polls seem to have established consensus on a range of economic questions. This consensus can be generally summarized with the following statement: “Price system or market is taken to be an effective and desirable social choice mechanism” (Frey et al. 1984, p. 994). In fact, Gordon and Dahl

(2013) found “no detectable systematic differences in views across departments, or across school of PhD.” Moreover, they found “no evidence to support a conservative versus liberal divide” (p. 635). This would mean that regardless of where economists complete their education and regardless of their priors, they tend to agree on the core points of their discipline.

Expert consensus is, and has often been, challenged, however. Attitudes toward economic policy differ among male and female economists (May et al. 2013), among Democrat and Republican economists (Klein et al. 2013), and, in 1984, it was shown that expert opinion differed even among economists based in different countries: “The American, German, and Swiss economists tend to support more strongly the market and competition than their Austrian and French colleagues, who rather tend to view government interventions into the economy more favorably” (Frey et al. 1984, p. 994). These findings suggest that gender, political affiliation, and cultural and historical circumstances influence what kind of questions economists ask and what kind of answers they tend to give and agree on.

But why should we care if economists agree on anything? Presumably, the answer lies in the fact that economists are trained to take things which are not seen, at least not immediately, into account. The qualified point of view is then needed to uncover common fallacies. “What economists think, and whether there is consensus among economists, would not be a matter of concern if *beliefs* do not have a very strong effect on economic policy decisions and on the state of the economy” (Frey et al. 1984, p. 986 emphasis original). Expert opinion is a consequence of a particular kind of division of labor that takes place in modern democratic societies. As such, expert opinion is a source of warranted beliefs. The division of labor, however, introduces a problem of trust.

Can the Consensus Be Trusted?

A division of cognitive and epistemic labor has fostered the creation of knowledge by means of

specialization and expertise. This division of labor, however, introduces problems of trust among the producers and consumers of expert knowledge. If the consumers of expert opinion do not trust its reliability, the expert consensus becomes inconsequential. Evidence shows that there is indeed a gap between expert and non-expert opinion.

In *The Republic of Science*, Michael Polanyi (1962, p. 471) argued,

Scientific opinion is an opinion not held by any single human mind, but one which, split into thousands of fragments, is held by a multitude of individuals, each of whom endorses the others’ opinion at second hand, by relying on the consensual chains which link him to all the others through a sequence of overlapping neighbourhoods.

There is a vast division of cognitive labor among scientists. Given his or her specialization, each expert looks at the world from a particular perspective. Each specialist, in turn, observes different aspects of reality. The deeper the specialization, the more diverse the observed aspects become. The essential difficulty consists in making sure these diverse observations are reliable and, above all, reconcilable with observations established by experts in distant neighborhoods of science.

Because of the division of cognitive labor, “nobody knows more than a tiny fragment of science well enough to judge its validity and value at first hand”; scientists have to “rely on views accepted at second hand on the authority of a community of people accredited as scientists” (Polanyi 1974, p. 173). Experts must mutually rely on the accuracy of peer review in scientific communities distant from their own. “If the aim of scientists is to maximize their knowledge of the world,” writes Jesús Zamora Bonilla (2008, p. 4), “they need trust in the word of their colleagues, making science a collective enterprise.”

On the one hand, the cognitive division of labor makes each scientist specialize in his or her area of expertise and build on the knowledge gained from exchanges and interactions with other experts, past and present. On the other hand, there is also epistemic division of labor.

Unlike the cognitive division of labor which results in specialization among scientists, epistemic division of labor separates the scientist from a nonscientist. At some point, the scientist had to make a decision to enter into the business of curiosity in the first place; it is a consequence of the epistemic division of labor that some people get to contribute to the production of expert consensus, while others become its mere consumers.

A pragmatist conception of democracy takes “the epistemic division of labor as one of the central features of effective and informed public deliberation”; there is a caveat, however: “Any such division of labor which produces epistemic gains will also produce deep asymmetries in the social distribution of knowledge” (Bohman 1999, p. 591). As a result, the body of scientific knowledge “can be received only when one person places an exceptional degree of confidence in another, the apprentice in the master, the student in the teacher, and popular audiences in distinguished speakers or famous writers” (Polanyi 1974, p. 220). Unless such a degree of confidence takes place, expert agreement loses its effect on public opinion.

Expert consensus has a strong effect on public opinion as long as it is relevant and credible. There are several reasons why expert consensus might be inconsequential, however. First, aspiring for abstract rigor, economic consensus might simply not be relevant for any of the problems non-experts perceive as pressing (Mayer 1993; Šťastný 2010). Second, as Bryan Caplan (2007, p. 53) points out, even if generally relevant, the consensus may be perceived as biased. Caplan identified two main challenges to the objectivity of expert consensus:

The first is *self-serving bias*. A large literature claims that human beings gravitate toward selfishly convenient beliefs. Since economists have high incomes and secure jobs, perhaps they are biased to believe that whatever benefits them, benefits all ... The second doubt about economists’ objectivity is less sordid but equally damaging: *ideological bias*. ... A consensus of fundamentalists hardly inspires confidence. It sounds like an intellectual chain letter: Maybe each batch of graduate students was brainwashed by the previous generation of ideologues.

In his analysis, Caplan found a gap between what economists and laymen believe, suggesting that the exceptional degree of confidence in expert consensus is indeed missing. When identifying the sources of the gap, however, it became clear that the “self-serving and ideological bias *combined* cannot account for more than 20% of the lay/expert belief gap. The remaining 80% should be attributed to the experts’ greater knowledge” (2007, p. 56 emphasis in original). The persistence of the gap must therefore be a consequence of some other cause. As Caplan hypothesizes, we “*turn off* our rational faculties on subjects where we don’t care about the truth” (2007, p. 2 emphasis in original). Consequently, given the persistent ignorance, the expert consensus – even if generally unbiased and reliable – fails to have a strong impact on public opinion.

Another explanation of the gap suggests that although ignorance may be convenient, “the problem, it seems, is not that members of the public are unexposed or indifferent to what scientists say, but rather that they disagree about what scientists are telling them” (Kahan et al. 2011, p. 148). Even if the expert consensus was perceived as generally reliable, it could still fail to have a strong effect on public opinion as a result of a cultural cognition effect according to which “individuals systematically overestimate the degree of scientific support for positions they are culturally predisposed to accept” (pp. 166–167).

Indeed, as Gordon Gauchat (2012) shows, a “growing distrust in science in the United States has been driven by a group-specific decline among conservatives” (p. 179). Such a decline may reflect particular cultural, political, and ideological characteristics of the research agenda that the conservative group of expert-opinion consumers is not willing to take in. This explanation seems plausible, given the mostly progressive composition of the US expert group (Klein et al. (2013) shows that the proportion of Democrat economists to Republican economists is approximately three to one).

Note, for example, that unlike in the United States, there is a “a general tendency to the right among the Swedish social scientists” (Berggren et al. 2009, p. 2). If the political composition of

academia influences the approach to research, the situation Gauchat observed in the United States may actually run in reverse in countries like Sweden, where academia tends to be rather conservative as compared to the general public. In these cases, the progressive public may, in fact, display decreased perceived credibility of the scientific consensus arrived at by a more conservative cohort.

In short, there are two conditions needed for the scientific consensus to make a difference in the process of public deliberation. It must be reliable, and it must be credible. As long as it is reliable, the expert consensus has the potential to improve our knowledge of the world. As long as it is credible, the expert consensus has the power to influence public opinion. These two conditions are implied by the process of consensus formation which is inherently social and which therefore crucially depends on the nature of the institutional framework that supports it.

The Nature of Scientific Institutions

Institutions are the rules of the game; they determine the structure of payoffs for the agents involved in the game. If consensus formation is a social process and if expert consensus results from the coordination of self-interested members of the expert communities, the reliability of scientific consensus will depend on whether the incentives which the self-interested members of scientific communities face are aligned in the right way. It is then the analysis of the nature of scientific institutions that sets the stage for our understanding of the creation and use of knowledge in modern democratic societies.

Keeping in mind the cognition effect, let us assume that the influence of expert consensus is a function of its reliability: the more reliable and accurate the scientific consensus gets, the more credible it becomes. There are two essential notions to consider when determining the reliability of scientific consensus. First, scientists follow their self-interest; in this, they are no different from any other subject of economic analysis. Second, expert consensus, as a heuristic source of

knowledge, induces conformity. Taking these assumptions into account implies that the reliability of expert opinion depends on particular properties of the scientific network that produced it. Let us look into these points further.

“Could it turn out,” asks Philip Kitcher (1990, p. 6), “that high-minded inquirers, following principles of individual rationality, should do a poor job of promoting the epistemic projects of the community that they constitute?” The epistemic division of labor presumably sorts out people who are better at scientific research from people who are better equipped, for example to, say, start and run a business. This does not imply, however, that there is a hardwired feature in each and every scientist forcing him or her to pursue the discovery of truth at all costs.

It might not be in the best interest of a scientist to advance the stock of reliable knowledge. Brock and Durlauf (1999) emphasize this point: “Whereas the predominant themes in the philosophy of science . . . presumed . . . identical desire to find ‘truth’ . . . recent trends . . . have been concerned with the social context in which research is conducted” (p. 114). The literature has discerned that the social context of research produces motivations, which may compete with the presumed scientific urge to find truth.

Payoff structures determine the best course of actions for every scientist. When social context determines the structure of payoffs, it is reputation that scientists value highly. As Paul David (1998) argues, “A scientist working in a collegiate reputational reward system will consider the nearer-term reputational consequences of current actions (including expressions of scientific opinion), as well as considering long-term payoffs possibilities in the form of lasting fame for having gotten it right.” In such a case, conformity may push against refining the body of reliable knowledge.

The success – or even survival – of scientists is, to a considerable extent, determined by their ability to publish in peer-reviewed journals. If getting it right coincides with getting it published, then “the implication is that referees act in the interests of science as a whole,” as Frey (2003, p. 208) pointed out. But can the epistemic division of labor rely on the inherent goodwill of referees

involved in the peer review? Frey (2003, pp. 208–209) explains:

Personal interests must also be expected to play a role. Many referees will be tempted to judge papers according to whether their own contributions are sufficiently appreciated and their own publications quoted. They carry, for instance, no costs when they advise rejection of a paper they dislike (e.g., because it criticizes their own work), even if they expect that it would be beneficial for economics as a discipline.

Economists may not question the generally followed assumptions of their discipline. In a situation when the problem at hand resembles Newtonian celestial dynamics, conforming to a mainstream consensus saves methodological effort. If, however, the sort of problems economists look into turns out to be better explained by some variant of generalized Darwinism, then following others in following Newton might lead the economist off the cliff – and into the land of the inconsequential.

According to Cass Sunstein, “following others can itself be seen as a heuristic, one that usually works well, but that also misfires in some cases” (2002, p. 38). Given the vast range of modern scientific knowledge, the consumers of expert consensus cannot help but follow expert opinion and trust that beliefs coming from the fields of science they are not acquainted with are well substantiated and properly justified. Expert consensus is a mental shortcut, a heuristic of consolidated opinion which, as J. S. Mill anticipated, “is salutary in the case of true opinions, as it is dangerous and noxious when the opinions are erroneous” (1859).

When conformity constitutes a rational course of action, “society can end up making large mistakes. . . . Social influences . . . threaten, much of the time, to lead individuals and institutions in the wrong directions.” In such cases, “dissent can be an important corrective” (Sunstein 2002, p. 40). Dissenters, however, are often ignored and even ostracized, accused of destroying peaceful agreement for their own selfish motives. Furthermore, the dissenting opinion is often embedded within a system of language and reasoning that, from the perspective of the prevailing expert consensus, appears to lack rigor and preciseness. New ideas

are substitutes for current prevailing thought. If reputation matters, and if the leading researchers are heavily invested in the prevailing consensus, plain dismissal of any dissent may occur.

When dissent does not pay off, originality does not take place. Dissenters, the individuals questioning the prevailing consensus, provide a valuable service that has a public-good character. By sharing their personal knowledge and pointing out where widely shared expert agreement runs short in terms of reliability, they contribute to the refinement of the body of scientific knowledge, benefiting all. Originality and dissent, however, are not absolutely valuable in themselves. They matter on the margin. Some questions of marginal analysis of dissent seem to follow: How much dissent is too much? And when does consensus become noxious?

The answer to these questions lies in the process of emergent consensus formation. An indispensable condition of producing reliable knowledge is an institutional mechanism, which ensures that the incentives scientists face align their self-interest with the general purpose of the epistemic division of labor. Through discovery and experimentation, an incentive-compatible institutional structure produces an emergent ratio of agreement and dissent. At this point, the production of scientific knowledge becomes a subject of comparative institutional analysis or, more broadly, economic analysis of scientific knowledge. “To the extent that we can make realistic presuppositions about human cognitive capacities and about the social relations found in actual communities of inquirers, we can explain, appraise and *in principle* improve our collective epistemic performance” (Kitcher 2002, p. 441 emphasis in original).

Given the agreement among economists about the price system being a desirable mechanism of social choice, it should be no surprise that one of the institutional arrangements suggested to improve the accuracy of scientific consensus is the prediction market, in other words, betting on beliefs. Prediction markets “doubtless have their limitations but they may be useful as a *supplement* to the other relatively primitive mechanisms for predicting the future like opinion

surveys, politically appointed panels of experts, hiring consultants or holding committee meetings” (Wolfers and Zitzewitz 2004, p. 125 emphasis added).

Robin Hanson, an advocate of prediction markets as an incentive-compatible institution of knowledge creation, claims that “betting prices ... are a robustly accurate public institution estimating policy-relevant topics” (Hanson 2013, p. 156). The system of prediction markets supports dissent by rewarding out-of-favor beliefs, which turn out to be true, more than true beliefs that everybody supports. Such a system also seems to transparently draw the line between desirable and undesirable dissent because it encourages the well-informed to step forward and speak up and the ill-informed or dishonest to stay silent.

The scientific wager of Julian L. Simon and Paul Ehrlich on the implications of price theory is, after all, a well-known affair. The practice of betting on beliefs, however, does not come without problems. Today, prediction markets that aggregate information on questions of science or policy are disparaged in the same way life insurance – another form of betting on beliefs – used to be constrained by repugnance. With the exception of British prediction markets (generating odds on matters such as the likelihood of secession from the Eurozone or on the prognosis of the 2015 UK unemployment rate), betting on beliefs is mostly illegal.

Prediction market is but one of the institutional arrangements aggregating individual estimates into a composite indicator. Such an arrangement may provide some benefits, but does it outperform the current organization of science in producing justified beliefs? “Whether the rules according to which scientists compete for recognition among each other, and the rules that govern their competition for resources, are well aligned and whether they support or inhibit each other in promoting the growth of knowledge is an empirical matter” (Vanberg 2010, p. 43). Eventually, a comparative analysis should provide insights into the effects of diverse mutations of scientific institutions on the process of consensus formation.

Concluding Remarks

The epistemic performance of modern democratic societies depends on the division of labor in the creation of knowledge. At the same time, however, there is a gap between expert and nonexpert opinion, suggesting that the cognitive and epistemic division of labor creates a problem of trust in the use and application of knowledge. There are, in fact, several reasons why expert consensus may fail to impact public opinion: expert consensus may be irrelevant, biased, or simply opposed. In general, expert opinion fails to make a difference in the process of public deliberation when it is unreliable.

The reliability of expert consensus depends on the nature of scientific institutions. In other words, the reliability of scientific consensus depends on whether the incentives, which the self-interested members of scientific communities face, are aligned in the right way. The analysis of scientific institutions sets the stage for our understanding of the creation and use of knowledge in modern democratic societies. It is the empirical assessment of how different institutional arrangements perform in the social process of consensus formation that future research will have to address.

References

- Berggren N, Jordahl H, Stern C (2009) The political opinions of Swedish social scientists. *Finn Econ Pap* 22(2):75–88
- Blaug M (1992) *The methodology of economics, or, how economists explain*. Cambridge University Press, Cambridge/New York
- Bohman J (1999) Democracy as inquiry, inquiry as democratic: pragmatism, social science, and the cognitive division of labor. *Am J Polit Sci* 43(2):590–607. <https://doi.org/10.2307/2991808>
- Brittan S (1973) *Is there an economic consensus?: an attitude survey*. Macmillan, London
- Brock WA, Durlauf SN (1999) A formal model of theory choice in science. *Econ Theory* 14(1):113–130
- Caplan B (2007) *The Myth of the rational voter: why democracies choose bad policies*. Princeton University Press, Princeton
- David PA (1998) Communication norms and the collective cognitive performance of “invisible colleges.” In: *Creation and transfer of knowledge*. Springer, Heidelberg,

- pp 115–163. Retrieved from http://link.springer.com/chapter/10.1007/978-3-662-03738-6_7
- Frey BS (2003) Publishing as prostitution? – choosing between one's own ideas and academic success. *Public Choice* 116(1–2):205–223. <https://doi.org/10.1023/A:1024208701874>
- Frey BS, Pommerehne WW, Schneider F, Gilbert G (1984) Consensus and dissension among economists: an empirical inquiry. *Am Econ Rev* 74(5):986–994. <https://doi.org/10.2307/557>
- Gauchat G (2012) Politicization of science in the public sphere a study of public trust in the United States, 1974 to 2010. *Am Sociol Rev* 77(2):167–187. <https://doi.org/10.1177/0003122412438225>
- Gordon R, Dahl GB (2013) Views among economists: professional consensus or point-counterpoint? *Am Econ Rev* 103(3):629–635. <https://doi.org/10.1257/aer.103.3.629>
- Hanson R (2013) Shall we vote on values, but bet on beliefs? *J Polit Philos* 21(2):151–178. <https://doi.org/10.1111/jopp.12008>
- Kahan DM, Jenkins-Smith H, Braman D (2011) Cultural cognition of scientific consensus. *J Risk Res* 14(2):147–174. <https://doi.org/10.1080/13669877.2010.511246>
- Kitcher P (1990) The division of cognitive labor. *J Philos* 87(1):5–22. <https://doi.org/10.2307/2026796>
- Kitcher P (2002) Contrasting conceptions of social epistemology. In: Brad Wray K (ed) *Knowledge and inquiry: readings in epistemology*. Broadview Press, Peterborough
- Klein DB, Davis WL, Hedengren D (2013) Economics professors' voting, policy views, favorite economists, and frequent lack of consensus. *Econ J Watch* 10(1):116–125
- May A, McGarvey MG, Whaples R (2013) Are disagreements among male and female economists marginal at best?: a survey of AEA members and their views on economic policy. *Contemp Econ Policy*. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/coep.12004/full>
- Mayer T (1993) *Truth versus precision in economics*. E. Elgar, Aldershot/Brookfield
- Mill JS (1859) On liberty. In: *Essays on politics and society*, vol XVIII. University of Toronto Press, Toronto
- Polanyi M (1962) The republic of science: its political and economic theory. *Minerva* 38(1):1–21
- Polanyi M (1974) *Personal knowledge: towards a post-critical philosophy*. University of Chicago Press, Chicago
- Šťastný D (2010) *The economics of economics: why economists aren't as important as garbagemen (but they might be)*. Instituto Bruno Leoni/CEVRO Institute and Wolters Kluwer, Turin/Prague
- Sunstein CR (2002) *Conformity and dissent*. Law School, University of Chicago, Chicago. Retrieved from http://www.law.uchicago.edu/files/files/34.crs_conformity.pdf
- Vanberg VJ (2010) The “science-as-market” analogy: a constitutional economics perspective. *Constit Polit Econ* 21(1):28–49. <https://doi.org/10.1007/s10602-008-9061-5>
- Wolfers J, Zitzewitz E (2004) Prediction markets. *J Econ Perspect* 18(2):107–126. <https://doi.org/10.2307/3216893>
- Zamora Bonilla JP (2008) The elementary economics of scientific consensus. *Theor Rev Teoria Hist Fundam Cienc* 14(3):461–488

Consequentialism

David Moroz

Champagne School of Management, Groupe ESC
Troyes, France

Abstract

Adopting a consequentialist approach requires knowing the whole set of consequences of an action. If we assume the future is radically uncertain, which may be argued for several different reasons, this approach can be used only in an explicative way and not in a predictive one.

Definition

Consequentialism can be defined as an instrumentalist approach to ethics that consists in evaluating a given system through its resulting effects (Pettit 2003; Blackburn 2008), i.e., through the maximization of gains and the minimization of losses it enables (Baggini and Fosl 2007).

Several forms of consequentialism can be identified (Baggini and Fosl 2007; Thiroux and Krasemann 2012), among which are the different kinds of utilitarianism. What enables the comparison between the different kinds of consequentialism can be said to be (i) the set of principles that are adopted to define positive as well as negative consequences and (ii) the choice of the agents, in a given system, that should benefit from positive consequences.

Consequentialism and Knowledge

From a methodological point of view, and if we want to embrace wider issues than the mere sphere of economic science, we cannot disconnect the

issues of consequentialism from the issues of knowledge. Indeed, outside of its ethical dimension, consequentialism is related to the ability to define causal relationships to enable the prediction of the results emerging from a given system (Thiroux and Krasemann 2012): choosing between alternative systems on the basis of expected results requires evaluating *ex ante* the relevance of said results. Such an issue is problematic for any consequentialist approach, even in the case where one individual determines for herself what positive and negative consequences are, i.e. even in the case where the subjectivity of individual preferences is respected.

To be able to predict with certainty the results of a given system, three conditions are necessary (Faber and Proops 1993): (i) the knowledge of its initial conditions; (ii) the absence of novelty; and (iii) the knowledge of its laws. Yet, three arguments enable us to explain that the future is radically uncertain: (i) the impossibility of predicting the evolution of knowledge; (ii) the cognitive limits of the analysis of complexity; and (iii) the impossibility of defining the limits of our ignorance.

Concerning the first argument, Popper (1957) explains that “if there is such a thing as growing human knowledge, then we cannot anticipate today what we shall know only tomorrow.” So it is impossible to assert that the last two conditions necessary for the prediction of a system – absence of novelty and knowledge of its laws – can be met.

The second argument relies on the limits of the analysis of so-called complex phenomena. As explained by Hayek, our ability to analyze complex phenomena is necessarily limited because of the impossibility for the human brain, as for any “apparatus of classification” (Hayek 1952), to analyze a system more complex than itself. Consequently, the theories we build can only be *ceteris paribus* constructions (Lachmann 1978) that cannot take account of the totality of variables that may influence a phenomenon.

Concerning the third argument, it can be found partly in Kirzner (1992) with the concept of limit of ignorance. Let us suppose it is possible to

differentiate among the set of future events, the predictable events and the unpredictable events. Drawing the frontier between these two sets of events would require knowing the limit of our ignorance, which is antithetical.

The concept of radical uncertainty does not mean that the future cannot be imagined, but that it cannot be known before its time (Lachmann 1978). In the opposite situation, the notion of novelty would be meaningless: novelty or surprise cannot exist in a world where the future is already known (see Shackle 1972, 1979).

This impossibility of building a causal relationship concerning the future does not, however, prevent building any causal relationship. It is here that the concept of time relativity defined by Hicks (1979), which enables us to differentiate a fixed past time and an evolving future time, takes on its full meaning. Past time is a closed set of events that can be analyzed following a deterministic approach (De Uriarte 1990), whereas future time is an open set of events, the result of which cannot be predicted.

Conclusion

As Hodgson (1996) explains, even if the world is determinist, we should consider it unpredictable. Indeed, if we adopt the deterministic approach, it is impossible to explain either the possibility of changing the future (Lawson 1988; Davidson 1996), of creating it (Shackle 1972; Lachmann 1977; Loasby 1991, 1999), or the existence of choice and action (Von Mises 1963; Shackle 1972; Hodgson 1996), i.e., free will (De Uriarte 1990; Hodgson 1996).

The rub is that if we assume the world as unpredictable, so as to consider the possibility for individuals to choose and to create, then resorting to any consequentialist approach to establish normative prescriptions becomes highly problematic.

Cross-References

► [Choice Under Risk and Uncertainty](#)

References

- Baggini J, Fosl PS (2007) *The ethics toolkit: a compendium of ethical concepts and methods*. Blackwell, Oxford
- Blackburn S (2008) *The Oxford dictionary of philosophy*, 2nd edn. Oxford University Press, Oxford
- Davidson P (1996) Reality and economic theory. *J Post Keynes Econ* 18(4):479–508
- De Uriarte B (1990) On the free will of rational agents in neoclassical economics. *J Post Keynes Econ* 12(4):605–617
- Faber M, Proops JLR (1993) *Evolution, time, production and the environment*, 2nd edn. Springer Verlag, Berlin Heidelberg
- Hayek FA (1952) *The sensory order: an inquiry into the foundations of theoretical psychology*. University of Chicago Press, Chicago
- Hicks J (1979) *Causality in economics*. Basil Blackwell, Oxford
- Hodgson GM (1996) *Economics and evolution: bringing life back into economics*. University of Michigan Press, Ann Arbor
- Kirzner IM (1992) *The meaning of market process: essays in the development of modern Austrian economics*. Routledge, London
- Lachmann LM (1977) *Capital, expectations, and the market process: essays on the theory of the market economy*. Subsidiary of Universal Press Syndicate, Kansas City
- Lachmann LM (1978) An Austrian stocktaking: unsettled questions and tentative answers. In: Spadaro LM (ed) *New directions in Austrian economics*. Sheed Andrew and McMeel, Kansas City, pp 1–18
- Lawson T (1988) Probability and uncertainty in economic analysis. *J Post Keynes Econ* 11(1):38–65
- Loasby BJ (1991) *Equilibrium and evolution: an exploration of connecting principles in economics*. Manchester University Press, Manchester
- Loasby BJ (1999) *Knowledge, institutions and evolution in economics*. Routledge, London
- Pettit P (2003) Consequentialism. In: Darwall S - (ed) *Consequentialism*. Blackwell, Oxford, pp 95–107
- Popper KR (1957) *The poverty of historicism*. Routledge & Kegan Paul, London
- Shackle GLS (1972) *Epistemics and economics: a critique of economic doctrines*. Cambridge University Press, Cambridge
- Shackle GLS (1979) *Imagination and the nature of choice*. Edinburgh University Press, Edinburgh
- Thiroux JP, Krasemann KW (2012) *Ethics: theory and practice*, 11th edn. Pearson Prentice Hall, Upper Saddle River
- Von Mises L (1963) *Human action: a treatise on economics*, 4th edn. Fox & Wilkes, San Francisco
- Hayek FA (1978) The pretence of knowledge. In: Hayek FA (ed) *New studies in philosophy, politics, and economics, and the history of ideas*. Routledge & Kegan Paul, London, pp 23–24
- Hodgson GM (2004) Darwinism, causality and the social sciences. *J Econ Methodol* 11(2):175–194
- Lawson T (1997) *Economics and reality*. Routledge, London
- Salmon WC (1998) *Causality and explanation*. Oxford University Press, New York
- Shackle GLS (1967) *Décision, Déterminisme et Temps*. Dunod, Paris

Constitutional Economics

► Constitutional Political Economy

Constitutional Evolution in Ancient Athens

George Tridimas

Department of Accounting, Finance and Economics, Ulster University Business School, Belfast, Northern Ireland

Definition

The present essay reviews recent political economy research in the policy-making institutions of the direct democracy of ancient Athens, 508–322 (all dates BCE), their origins, and evolution. The historical events, especially tensions between rich and poor and existential external threats, which led to the emergence of the Athenian democracy, are first described. The institutions of decision-making, assembly, magistrates, and courts, are then discussed along with various reforms in response to changing circumstances. The essay ends with some reflections on the nature of the direct democracy, its legitimatization, and the internal consistency of its institutions in comparison to its modern representative analog.

The Birth of Democracy

Constitutional economics raises time-invariant questions. Inquiring how historical societies

Further Reading

- Auyang SY (1998) *Foundations of complex-system theories in economics, evolutionary biology, and statistical physics*. Cambridge University Press, Cambridge

addressed constitutional building, architecture of governance, and rights of governed, not only allows to make sense of the past but also helps to understand present arrangements, as well as showing the strengths and limitations of theoretical models. This is even more so for democracy with its key elements of freedom and political equality. It is this realization which has inspired a vibrant branch of research by both political economists and historians into the political institutions of ancient Athens, their origins, evolution, and performance. The present essay surveys this work and reflects on the constitutional evolution of ancient Athens.

The constitutional development of ancient Athens is the story of the emergence of direct participatory democracy (albeit for adult men only) where political events and institutional building were inextricably linked (Aristotle (1984) for the original account; Raaflaub et al. (2007) for debate on the origins of the Athenian democracy; Cartledge (2016) for birth and tribulations of ancient democracy compared to modern). Table 1 summarizes the main developments.

In the Archaic Times (750–500), Athens was governed by the nine archons appointed from the members of the noble, landed, elite, who were responsible for religious, military, civic affairs, and recording laws. After serving one-year terms, the archons were appointed for life as members of the Council of *Areopagus* with the authority to oversee laws and magistrates and conduct trials. In 594, following a period of internal social conflicts, Solon, an aristocrat, was appointed as lawgiver and introduced a series of institutional and economic changes. Inscribed and publically displayed, the laws keynoted that governance secretly managed by the aristocracy had ended. Among other issues, Solon conditioned appointment to public office on wealth (valued in annual agricultural production) rather than hereditary birthright, with the highest income classes enjoying access to more powerful offices and the lowest (the landless) altogether excluded. He also granted all citizens the right to participate in the assembly (which was more like a consultative body as it was not empowered to make binding decisions) and act as prosecutors in criminal

Constitutional Evolution in Ancient Athens, Table 1 Timeline of the Athenian democracy

750–500	Archaic Athens Main government bodies: Nine archons selected from the aristocracy; ex-archons appointed to the Areopagus overseeing laws and magistrates and conducted trials
594	Solon, the lawgiver, introduced a wealth-based political dispensation
546–510	Tyranny of Peisistratus and his son Hippias
510	Hippias expelled
508–404	Classical Athens. Fifth Century Democracy
508–507	Democracy established: Cleisthenes reforms of citizenship and council of five hundred. Ostracism introduced
490–479	Persian wars. 480: Athenian victory at Marathon; 490: Athenian victory at Salamis
487	Selection of nine archons by lot
462	Powers of Areopagus removed. Introduction of pay for court service
451	Pericles' law restricts citizenship to those whose both parents were Athenians
431–404	<i>Peloponnesian War, Athens V Sparta</i>
415–413	Sicilian expedition of Athenian navy. Syracuse and Sparta defeat Athens
411	Democracy overthrown by oligarchic coup
410	Democracy restored by the Athenian navy
405	Defeat of Athens at Aegospotami
404	Athenian defeat and surrender; Tyranny of the Thirty
404–322	Classical Athens. Fourth Century Democracy
403	Democracy restored
403/402	Introduction of pay for attending the assembly
ca 402	The assembly voted that new laws would be made by boards of legislators appointed by lot
ca 355	<i>Theoric</i> (festival money) fund formalized
338	Athens & Thebes defeated in Chaeronea by Philip of Macedon
322	End of the Athenian Democracy after defeats by Macedon in the Lamian War

trials, and introduced accountability of magistrates. In 546, Peisistratus established tyranny, meaning extra-constitutional dispensation rather than oppressive government, which lasted until 510 (see Fleck and Hanssen (2013) on the transition from tyranny to democracy). In the ensuing contest for power between two aristocrats

Isagoras and Cleisthenes, the former prevailed. Then, in an unprecedented move, Cleisthenes allied himself to the *demos*, the common people, proposing reforms that would extend political rights. Isagoras turned to oligarchic Sparta, the strongest military power at the time for help. Spartan forces occupied Athens and expelled Cleisthenes. However, when they tried to dissolve the governing council and impose Isagoras, the Athenian *demos* rose up forcing the Spartans to leave. Cleisthenes was recalled in 508 and introduced a series of fundamental constitutional reforms 508/7 that gave birth to the direct participatory democracy.

The Institutions of the Athenian Direct Democracy

In the first instance, the reforms reconstituted the rules for citizenship, the idea that all locally born free men within a city-state had equal political rights and enjoyed legal protections, regardless of wealth, birth, education, or any other factor (Hansen (1999) offers an extensive treatment of the operation of the Athenian democracy; Ober (2008) focuses on aggregation of private knowledge through democratic institutions; Tridimas (2011) and Lyttkens (2013) analyze the institutions of democratic governance in the light of political economy). Citizenship rights were extended to all adult resident males, a move equivalent to egalitarian enfranchisement (citizenship rights were, however, limited by *Pericles* in 451 to those whose both parents were Athenians). Cleisthenes then divided the citizens into three geographical sections, Urban, Inland, and Coast, and further divided each geographical section to ten parts and a total of 139 *demes* (local communities) whose membership was hereditary. A lottery allocated the resulting 30 groups to 10 new tribes ("*phylae*"), so that each new tribe included groups from each one of the three geographical sections with their different traditions and economic bases. Each tribe was a microcosm of the citizenry and shared the same interests with the rest of the tribes. The new structure translated into a more effective and successful Athenian

military (see Pritchard 2015 on the contribution of democracy to the military success of Athens).

Next, the reforms established the assembly of (male) Athenian citizens as the principal decision-making body, responsible for all domestic and external policy issues, including military issues, foreign alliances, tax, expenditure, infrastructure works, and public festivals. All adult male Athenians had the right to vote and address the assembly. Out of a population of perhaps 60,000 male citizens in the fifth century, the quorum stood at 6000. The quorum remained the same during the fourth century when as a result of disease, war losses, and famine population fell to 30,000. In the fifth century, the assembly met 10 times a year, but gradually the number increased to 40 in the second half of the fourth century (Tridimas (2017) on the choice of frequency of voting). Contrary to modern practice, policy measures were proposed by individuals in their capacity as citizens rather than office holders. Neither political parties nor executive offices as known today existed. It is for this reason that the ancient authors never refer to anything resembling the office of president or prime minister. Although all Athenians had the right to propose policy measures and address the assembly, in practice a small number of political leaders, statesmen, dominated its debates. Decisions were taken by majority (see Pitsoulis (2011) for the origins of majority voting). Voting took place by show of hands; if the show was unclear, then the chairing officers counted hands.

Cleisthenes set up a new body the Council of Five Hundred (fifty from each tribe) selected annually by lot, with responsibilities to prepare the agenda of the assembly and carry out the day-to-day administration of Athens. Members of each tribe chaired the administration of Athens for one-tenth of the year. Every day at sunset, the 50 councilors of the presiding tribe selected by lot a chairman, who was the head of the Athenian state for a night and a day. The members of the Council met every working day and received a compensation for their services, so that poorer citizens could afford to take time off their daily work and serve in public office. Councilors too voted by show of hands. A man could serve in the Council only

twice in his lifetime, implying a considerable turnaround in the office and, as a consequence, a large number of Athenians with some experience of the affairs of the state.

In 501, the board of the Ten Generals was introduced to serve as commanders of the army and navy for a year. In a special assembly meeting, candidates were proposed from each tribe, but had to be voted in by a majority of the voters of all tribes, so a General was not the political representative of his tribe; from 440 at least one General was elected from all tribes. In contrast to other office-holders who were subject to term limits, generals could be reelected.

The *Heliaia* Court of 6000, or “People’s Court,” first set up by Solon to hear appeals against the decisions of the officials, had its powers enlarged in 462 and became the most important court with wide responsibilities. Every year 6000 citizens (600 from each tribe) not in debt to the state were selected by lot to serve as jurors. Each day jurors were allocated to cases by lot and sat in sessions with a jury size of 501 or bigger as the case may be (201 minimum). There was no public prosecutor all parties appearing before the court, citizens who brought a charge, the magistrates preparing and presiding over a case and the jurors who decided were ordinary citizens without any legal training. Contrary to assembly votes, jury decisions were taken by secret ballot. Pay for jurors was introduced most probably in 462 thought to be equal to half the average wage. From the fifth century, the court tried both civil and penal cases, checked the eligibility and conduct of public officers (but not competence), and conducted trials for treason and corruption (Karayiannis and Hatzis (2012) for social norms and the rule of law; Carugati et al. (2015) for the decentralized legal order of Athens).

In addition, another six hundred, or so, magistrates were appointed annually by lot to serve the polis. They typically worked in boards of ten members and carried specific administrative tasks, including inspection of markets, supervision of public works, judicial administration, collection of state revenues, etc.

Cleisthenes also introduced ostracism (banishment) whereby the demos in a secret ballot

decided whether to banish a leading individual for a period of 10 years, but without any further criminal or financial penalties. It was conceived as a mechanism to stop destructive violent conflict for power by political leaders and to defend the democracy. It was used sparingly with ten attested ostracisms during the period 507–416 and none afterward, although the procedure was not abolished (see Tridimas (2016) for a game theoretic analysis).

Military success followed setting Athens in a glorious trajectory. In 490, the hoplites army of landowner–farmer citizens pushed back the invading Persians in the battle of Marathon. In 483/2, a rich silver vein was discovered in Southern Attica. The demos voted to use the windfall to build a navy instead of distributing it equally among the citizenry, which was a common practice at the time (Tridimas 2013). The fleet triumphed against the Persians in the 480 sea battle of Salamis. Athens then transformed to a sea power precipitating a shift in the internal balance of political power against the large and mid-size landowners who were the backbone of the land forces, and in favor of the poor class of the landless. They had found gainful employment as rowers in the newly built fleet, and on the basis of their new strength, they gained access to all political offices. By the mid-fifth century, direct democracy for the Athenian male citizens was fully functioning.

After the final defeat of the Persians in 479, Athens pursued an offensive war heading the Delian League of Greek islands and coastal city-states. In 478, the League was transformed into the Athenian hegemony, where the allies were paying tribute to Athens for protection by her navy (see Rhodes (2013) for the Athenian public finances). Athens reached the peak of its power, excelling in public monument building and as the intellectual and cultural center of the time.

War, Constitutional Reforms, Recovery, and End of Democracy

Intense rivalry between Athens and Sparta, the traditional Greek military power, led to the

Peloponnesian War, 431–404, spreading to the rest of Greece, Asia Minor, and the Greek colonies of Southern Italy and Sicily. Convulsions followed the 413 destruction of the Athenian fleet in Sicily. In 411, the democracy was overthrown by an oligarchic regime restricting full political rights to 5000 men only. Four months later, in 410, after defeating the Spartan fleet the navy reinstated the democracy. In 404, Athens was finally defeated. With Spartan backing, Athens was ruled by a cruel 30-member strong oligarchic commission, known as the “Thirty Tyrants.” The democrats regrouped, and in 403 after some fierce fighting, they expelled the Tyrants and restored democracy.

A number of institutional changes followed. Blaming the “demagogues” for misleading the demos in the Peloponnesian War, the powers of the assembly were curtailed. Instead of the assembly, a special board of legislators chosen by lot from the same panel of 6000 jurors of the People’s Court was appointed to pass laws describing “general norms without limit of duration.” The assembly still voted decrees and decided foreign policy (Schwartzberg (2004) on the sovereignty of the demos and legal change). In what can be thought as an early type of judicial constitutional review, the courts acquired more powers through deciding a lawsuit alleging an unconstitutional proposal. Accordingly, they could nullify assembly measures deemed contrary to the laws, unfavorable to the interests of the people, or procedurally invalid, and punish their proposers (Lyttkens et al. 2017). Nevertheless, it differs from modern judicial review for the latter is carried out by professional judges, while the Athenian one was entrusted to ordinary folk, amateurs who had no formal legal training. Thus, the demos exercised its rule in both the assembly and the courts (Hansen 1999).

Funding of a variety of state functions was fixed by law diminishing the discretionary power of the Assembly to allocate public expenditures. The new elected offices of the treasurer of the military fund, the board of the *theoric* fund to manage festival money, and the controller of the finances were also established. Supervision of the administration of the laws was transferred to the

Areopagus. Finally, pay for the first 6000 citizens attending the Assembly at the average daily wage was introduced. A further development of the fourth century was the relative decline of Generals as political leaders in comparison to the fifth century when Generals dominated assembly deliberations and provided policy direction. In the fourth century, without holding any formal office, a number of self-selected orators rose to prominence debating in the assembly and the courts, while the Generals focused on more narrow military matters.

Like all democracies, Athens redistributed away from the rich by taxing them and mandating them to finance various public services, known as liturgies (Kaiser 2007; Lyttkens 2013; Tridimas (2015b) on rent-seeking). Although not manifested through political parties, for there were none in the direct democracy, the division between the rich and the poor was clear and was evident in foreign policy. The rich, who carried the burden of war finance, and the farmers anxious when leaving their lands to fight, favored peace. But the poor with less property, the prospect of gainful employment in the fleet and land allotments abroad if victorious, favored war, while also worried about the abolition of democracy if Athens’ enemies prevailed.

Over the fourth century, the Athenian economy recovered after the defeat, and eventually thrived based on international trade buttressed by non-discriminatory laws that focused on transactions rather the origin (Athenian or foreign) of the transactors and a stable regulatory framework (Ober (2015) and Bresson (2016) for the performance of the Athenian economy; Bitros and Karayiannis (2010) on moral norms of Athens, Bergh and Lyttkens (2014) on institutional quality; Economou and Kyriazis (2017) on the evolution of private property rights). Tax revenues rose substantially (Kyriazis 2009). New alliances were formed and several wars were fought against rival Greek city-states, but Athens failed to achieve the earlier supremacy. In the 338 battle of Chaeronea, Philip of Macedon inflicted a heavy defeat on an Athenian–Theban alliance forcing them to join the Macedonian alliance that led the invasion of Persia under Alexander the Great.

Following the death of Alexander the Great in 322, Athens fought to break away from the Macedon-led alliance. It suffered a double defeat in the sea battle of Amorgos and the land battle of Crannon. An oligarchy was then established with Macedonian support. The democracy ended as the oligarchy abolished the courts, disenfranchised the poor (limiting the franchise to 9000 property owners), stopped pay for public service and assembly participation, abolished large numbers of liturgies, and reduced assembly meetings to ceremonial functions (Tridimas 2015a).

Reflections on the Nature of the Athenian Democracy

Democracy means rule of the demos, the many. Nevertheless, its Athenian essence was not just numerical, the majority versus the minority of citizens, but also economic, that is, democracy connoted rule by the “middling” and the “poor,” meaning all those who unlike the “rich” had to work for a living (Patriquin 2015). In explaining the emergence of the Athenian democracy, ancient historians focus on identifying “historically contingent factors,” the weight of important events as in 594, 508, and 462. On the contrary, economists search for reasons why the enfranchised elite may accept democracy which works against their privileged access to power. Fleck and Hanssen (2006) show that democratic institutions resolve time inconsistency problems, and thus it increases investment. In the broken territory of Athens, more suitable for olive oil than wheat production, monitoring the input of captive labor was next to impossible. The only way for the elite to ensure that nonelite engage in productive investment, so that it generates the necessary surplus to finance defense, is when the nonelite enjoy long-run security of their properties. The latter is achieved when the nonelite have full political rights so that they can vote for policies conducive to their interests. McCannon (2012) explains that democracy was beneficial to both the elite and the nonelite of Athens. The Athenian elite experienced significant wealth volatility, which implied that the children of the current generation of

the elite faced a considerable risk of suffering wealth losses and losing their privileged position. Democracy with its egalitarian principle of access to power not conditional on wealth and its redistributive impact provided an insurance mechanism against the risk of loss of wealth and political standing.

The historical narrative indicates that the establishment of the Athenian democracy was a gradual and cumulative process of institutional building over a long period in response to changing political, military, and economic circumstances. Starting with the reforms of Solon, which based eligibility for public office on wealth holdings rather than aristocratic birthright, and including an interval of tyranny; the process moved on with the reforms of Cleisthenes, which enfranchised all Athenian males, introduced ostracism as a peaceful way to resolve conflicts via electoral means, then opened up public office to the poorer citizens. Further, democracy-building encompassed appointment to public office by lot, payment of a service fee, so that citizens could afford time off their production activities for public service, and further changes in the fourth century which included payment for assembly attendance, conceptualization of the difference between permanent laws and assembly decrees, elevation the courts to an effective veto player by granting them the power to check the decisions of the assembly, and introduction of new treasury offices. The process bears remarkable similarities to Congleton’s (2011) theory of the emergence of western democracy and representative government from the industrial revolution to the twentieth century, that is, slow and piecemeal, building on established arrangements and more peaceful than violent. More importantly, there was nothing teleological about the establishment of the democracy, it was not inevitable, and as described despite its comparative longevity 507–322, it did not survive.

Table 2 compares the institutions of the Athenian democracy with those of a stylized modern democracy.

The idea of citizenship was central to the polis. Citizenship conferred political and civic rights and economic benefits, and was strictly regulated.

Constitutional Evolution in Ancient Athens, Table 2 Comparison of Athenian and modern democracies

	Athenian democracy	Modern democracy ^a
Policy making	Direct and participatory: The assembly chose policy	Indirect: Citizens vote for party candidates, who then choose policy
Legitimacy ^b of government	Equality of opportunity that citizens may occupy public office	Citizens consent to be governed by those who win an electoral majority
Political parties	Absent	Fundamental players in electoral competition
Chief executive	Not important	President or Prime Minister
Frequency of decision-making	Frequent assembly meetings per year	Periodic elections
Voting rule	Simple majority	Varies from majoritarian to proportional representation
Appointment to public office	By lot Election limited to a few offices	Election
Justice and administration of the state	Ordinary citizens took turns to serve as jurors and carry out administrative tasks	Professional judiciary Professional bureaucracy
Policy review and accountability of officials	Popular court – juries of ordinary citizens	Courts – professional judges/legal experts

^aStylized description of modern representative government

^bLegitimacy: the decision taker is recognized to have the right to do what he does

Citizens were expected to be active participants in the polis from fighting wars to deciding public policy and holding public office regularly. It marked a clear break with the eastern empires preceding the Greek city-states, where ordinary people were subjects of kings and obliged to pay tribute to them. On the other side of the spectrum, the premise of the ancient democracy was that all citizens have equal opportunities to serve in public office. This is unlike modern representative democracies which are predicated on the principle that citizens have equal opportunities to consent to how they are governed (Manin 1997).

The true hallmark of democracy was appointment to office by lot (also called sortition), rather than voting. The luck of the draw eliminated the advantages that the rich elite had in contesting elections for office (Taylor 2007; Tridimas 2012; Lyttkens 2013). The lot nullified the policy-making privileges of public offices formerly controlled by the aristocratic class, and by randomizing appointment to office, it also randomized the distribution of rents from holding office. Of course, sortition cannot select political leaders with the qualifications to govern, nor can exclude incompetent individuals from office. The

Athenians were aware that not all citizens were suitable for positions of responsibility. Thus, only offices that did not require specialized expertise were filled by lot. To put it another way, the Athenians accepted that the acts of boards filled by sortitioned members were invariant to the composition of those boards. On the other hand, offices requiring leadership skills, like the position of the general who had to be trusted in leading to battle, were filled by elections. The sortitioned public postholders served annual terms implying that considerable rotation took place.

The Athenian institutions of direct democracy, decision-making in the assembly by majority voting, appointment to office by lot, accountability to the courts consisting of ordinary members of the demos instead of professional legal experts, absence of political parties, and lack of professional expertise in public administration, comprised an integral structure, consistent with each other and mutually reinforcing. No part of the constitutional package could operate independently of the rest. When voters decide directly on every single policy issue which can be introduced by any willing citizen (a) simple majority suffices, (b) there is no need for political

parties to bundle issues in party platforms, and (c) the search for voting formulas other than simple majority to aggregate votes is negated. Further, in the absence of elections for candidates to office, the courts with juries randomly drawn from the citizenry ensured accountability. Equally, when direct democracy empowers citizens to occupy public office through sortition, implying that any citizen might hold office, and annual rotation, implying that every citizen might hold office at some time, (a) political parties can no longer secure rents for their members, (b) no classes of professional politicians or bureaucrats emerge, and (c) citizens have an incentive to learn about the running of the state and “all things political.”

Conclusion

It is clear from the above brief account that ancient and modern democracy share the basic principle of people's rule and equality in the sense of one man one vote, but in truth little else. Assembly debate to decide public policy and sortition to administer the state, rather than voting for candidates and party political intermediation, were the fundamentals of the former. Undoubtedly, the direct and the representative democracies share common values and principles, but conceptually and practically they are apart.

Cross-References

- [Constitutional Political Economy](#)
- [Voting Power Indices](#)

References

- Aristotle (1984). *The Athenian constitution*. London: Penguin Classics. (translated by P. J. Rhodes).
- Bergh A, Lyttkens CH (2014) Measuring institutional quality in ancient Athens. *J Inst Econ* 10:279–310
- Bitros GC, Karayiannis AD (2010) Morality, institutions and the wealth of nations: some lessons from ancient Greece. *Eur J Polit Econ* 26:68–81
- Bresson A (2016) The making of the ancient Greek economy: institutions, markets and growth in the city-states. Princeton University Press, Princeton. (translated by S Rendall)
- Cartledge P (2016) *Democracy. A life*. Oxford University Press, Oxford
- Carugati F, Hadfield G, Weingast B (2015) Building legal order in ancient Athens. *J Legal Anal* 7:291–324
- Congleton RD (2011) *Perfecting parliament: Constitutional reform and the origins of western democracy*. Cambridge University Press, New York
- Economou EML, Kyriazis N (2017) The emergence and the evolution of property rights in ancient Greece. *J Inst Econ* 13:53–77
- Fleck RK, Hanssen FA (2006) The origins of democracy: a model with applications to Ancient Greece. *J Law Econ* 49:115–146
- Fleck RK, Hanssen FA (2013) How tyranny paved the way to democracy: the democratic transition in ancient Greece. *J Law Econ* 56:389–416
- Hansen MH (1999) *The Athenian democracy in the age of demosthenes. Structure, principles and ideology*. Bristol Classical Press, London
- Kaiser BA (2007) The Athenian trierarchy: mechanism design for the private provision of public goods. *J Econ Hist* 67:445–480
- Karayiannis AD, Hatzis AN (2012) Morality, social norms and the rule of law as transaction cost-saving devices: the case of ancient Athens. *Eur J Law Econ* 33:621–643
- Kyriazis NK (2009) Financing the Athenian state: public choice in the age of Demosthenes. *European Journal of Law and Economics* 27:109–27
- Lyttkens CH (2013) *Economic analysis of institutional change in Ancient Greece: politics, taxation and rational behaviour*. Routledge, Abingdon
- Lyttkens CH, Tridimas G, Lindgren A (2017) Making direct democracy work. An economic perspective on the *graphe paranomon* in ancient Athens. Available at http://swopec.hhs.se/lunewp/abs/lunewp2017_010.htm
- Manin B (1997) *The principles of representative government*. Cambridge University Press, Cambridge
- McCannon BC (2012) The origin of democracy in Athens. *Rev Law Econ* 8:531–562
- Ober J (2008) *Democracy and knowledge*. Princeton University Press, Princeton and Oxford
- Ober J (2015) *The rise and fall of classical Greece*. Princeton University Press, Princeton and Oxford
- Patriquin L (2015) *Economic equality and direct democracy in ancient Athens*. Palgrave MacMillan, New York
- Pitsoulis A (2011) The egalitarian battlefield: Reflections on the origins of majority rule in archaic Greece. *European Journal of Political Economy* 27:87–103
- Pritchard DM (2015) Democracy and war in ancient Athens and today. *Greece and Rome* 62:140–154
- Raaflaub KA, Ober J, Wallace RW (eds) (2007) *Origins of democracy in Ancient Greece*. University of California Press, Berkeley
- Rhodes PJ (2013) The organization of Athenian public finance. *Greece and Rome* 60:203–231

- Schwartzberg M (2004) Athenian democracy and legal change. *Am Polit Sci Rev* 98:311–325
- Taylor C (2007) From the Whole Citizen Body? The Sociology of Election and Lot in the Athenian Democracy. *Hesperia* 76:323–346
- Tridimas G (2011) A political economy perspective of direct democracy in ancient Athens. *Constit Polit Econ* 22:58–72
- Tridimas G (2012) Constitutional choice in ancient Athens: the rationality of selection to office by lot. *Constit Polit Econ* 23:1–21
- Tridimas G (2013) Homo oeconomicus in ancient Athens. Silver bonanza and the choice to build a navy. *Homo Oeconomicus* 30:14–162
- Tridimas G (2015a) War, disenfranchisement and the fall of the ancient Athenian democracy. *Eur J Polit Econ* 31:102–117
- Tridimas G (2015b) Rent seeking in the democracy of ancient Greece. In: Hillman AL, Congleton RD (eds) *The Elgar companion to the political economy of rent seeking*. Edward Elgar Publishing, Cheltenham, pp 444–469
- Tridimas G (2016) Conflict, democracy and voter choice: a public choice analysis of the Athenian ostracism. *Public Choice* 169:137–159
- Tridimas G (2017) Constitutional choice in Ancient Athens: the evolution of the frequency of decision making. *Constit Polit Econ* 28:209–230

Constitutional Political Economy

Stefan Voigt
Institute of Law and Economics, University of
Hamburg, Hamburg, Germany

Abstract

Economists used to be interested in analyzing decisions assuming the rules to be given. Scholars of Constitutional Political Economy (CPE) or constitutional economics have broadened the scope of economic research by analyzing both the choice of basic rule systems (constitutions) as well as their effects using the standard method of economics, i.e., rational choice.

Synonyms

Constitutional economics

Definition

Constitutional Political Economy analyzes the choice of constitutions as well as their effects by drawing on rational choice.

Introduction

Buchanan and Tullock (1962, p. vii) define a constitution as "... a set of rules that is agreed upon in advance and within which subsequent action will be conducted." Although quite a few rule systems could be analyzed as constitutions under this definition, the most frequently analyzed rule system remains the constitution of the nation state. Two broad avenues in the economic analysis of constitutions can be distinguished: (1) the normative branch of the research program, which is interested in legitimizing the state and the actions of its representatives, and (2) the positive branch of the research program, which is interested in explaining (a) the outcomes that are the consequence of (alternative) constitutional rules and (b) the emergence and modification of constitutional rules.

Normative Constitutional Political Economy

Methodological Foundations

Normative CPE deals with a variety of questions, such as: (1) How should societies proceed in order to bring about constitutional rules that fulfill certain criteria, like being "just"? (2) What contents should the constitutional rules have? (3) Which issues should be dealt with in the constitution – and which should be left to post-constitutional choice? (4) What characteristics should constitutional rules have?, and many more. James Buchanan, one of the founders of CPE and the best-known representative of its normative branch, answers none of these questions directly but offers a conceptual frame that would make them answerable. The frame is based on social contract theory as developed most prominently by Hobbes. According to Buchanan (1987,

p. 249), the purpose of this approach is justificatory in the sense that “it offers a basis for normative evaluation. Could the observed rules that constrain the activity of ordinary politics have emerged from agreement in constitutional contract? To the extent that this question can be affirmatively answered, we have established a legitimating linkage between the individual and the state.”

The value judgment that a priori nobody’s goals and values should be more important than anybody else’s is the basis of Buchanan’s entire model. One implication of this norm is that societal goals cannot exist. According to this view, every individual has the right to pursue her own ends within the frame of collectively agreed upon rules. Accordingly, there can be no collective evaluation criterion that compares the societal “is” with some “ought” since there is no such thing as a societal “ought.” But it is possible to derive a procedural norm from the value judgment just described. Buchanan borrowed this idea from Knut Wicksell (1896): Agreements to exchange private goods are judged as advantageous if the involved parties agree voluntarily. The agreement is supposed to be advantageous because the involved parties expect to be better off with the agreement than without it. Buchanan follows Wicksell who had demanded the same evaluation criterion for decisions that affect more than two parties, at the extreme an entire society. Rules that have consequences for all members of society can only be looked upon as advantageous if every member of society has voluntarily agreed to them. This is the Pareto criterion applied to collectivities. Deviations from the unanimity principle could occur during a decision process on the production of collective goods, but this would only be within the realm of the Buchanan model as long as the constitution itself provides for a decision rule below unanimity. Deviations from the unanimity rule would have to be based on a provision that was established unanimously.

Giving Efficiency Another Meaning

Normative CPE thus reinterprets the Pareto criterion in a twofold way: It is not outcomes but rules that, in turn, lead to outcomes which are evaluated

using the criterion. The evaluation is not carried out by an omniscient scientist or politician but by the concerned individuals themselves. To find out what people want, Buchanan proposes to carry out a consensus test. The specification of this test is crucial as to which rules can be considered legitimate. In 1959, Buchanan had factual unanimity in mind and those citizens who expect to be adversely affected by some rule changes would have to be factually compensated. Later, Buchanan seems to have changed his position: Hypothetical consent deduced by an economist will do to legitimize some rule (see, e.g., Buchanan 1977, 1978, 1986). This position can be criticized because a large variety of rules seem to be legitimizable depending on the assumptions of the academic who does the evaluation. Scholars arguing in favor of an extensive welfare state will most likely assume risk-averse individuals, while scholars who argue for cuts in the welfare budgets will assume that people are risk neutral.

Buchanan and Tullock (1962, p. 78) introduced the veil of uncertainty, under which the individual cannot make any long-term predictions as to her future socioeconomic position. As a result, unanimous agreement becomes more likely. John Rawls’ (1971) veil of ignorance is more radical because the consenting individuals are asked to decide *as if* they did not have any knowledge on who they are. Both veils assume a rather curious asymmetry concerning certain kinds of knowledge: On the one hand, the citizens are supposed to know very little about their own socioeconomic position, but on the other, they are supposed to have detailed knowledge concerning the working properties of alternative constitutional rules (Voigt (2013) summarizes veilonomics).

Every society must decide on which actions are to be carried out on the individual level and which ones are to be carried out on the collective one. To conceptualize this decision problem, Buchanan and Tullock introduce three cost categories: *External costs* are those costs that the individual expects to bear as a result of the actions of others over which she has no direct control. *Decision-making costs* are those which the individual expects to incur as a result of her own participation

in an organized activity. The sum of these two cost categories is called *interdependence costs*. Only if their minimum is lower than the costs under private action will society opt in favor of collective action.

After having decided that a certain activity is to be carried out collectively, societies need to choose decision-making rules that are to be used to make concrete policy choices. The more inclusive the decision-making rule, the lower the external costs an individual can expect to be exposed to. Under unanimity, they are zero. Decision-making costs tend, however, to increase dramatically the larger the number of individuals required to take collective action. Buchanan and Tullock claim that “for a given activity the fully rational individual, at the time of constitutional choice, will try to choose that decision-making rule which will *minimize* the present value of the expected costs that he must suffer” (1962, p. 70).

Positive Constitutional Economics

Positive CPE can be divided into two parts: On the one hand, it is interested in explaining the outcomes that result from alternative rule sets. On the other, it is interested in explaining the emergence and modification of constitutional rules. Following Buchanan and Tullock (1962), the genuine contribution of CPE to economics should consist in endogenizing constitutional rules. To date, analysis of the effects of constitutional rules has, however, played a more prominent role. This is why we begin by summarizing the research on the effects of constitutions.

The Effects of Constitutional Rules

Quite a few empirical studies confirm that constitutional rules can cause important differences, only some of which can be mentioned here (Voigt (2011) is a more complete survey). Conceptually, it makes sense to distinguish between the catalogue of basic rights on the one hand and the organization of the various state representatives (*Staatsorganisationsrecht*) on the other. Here, we confine ourselves to the latter and begin by looking at the effects of electoral rules and the form of government as these two aspects

of constitutions have been analyzed the most. We continue with federalism and direct democracy. Other organizational aspects that deserve to be analyzed include uni- versus bicameral legislatures and the independence that so-called non-majoritarian institutions such as central banks enjoy. Further, analysis is confined to democracies, as analyzing the effects of electoral rules makes little sense with regard to autocracies.

Electoral Rules

According to Elkins et al. (2009), only some 20% of all constitutions contain explicit provisions regarding the electoral system to be used for the choice of parliamentarians. For two reasons, we nevertheless begin this brief survey with electoral systems: (1) The definition of constitutions introduced above does not imply that only constitutions “with a capital C” can be recognized and (2) the effects of these rules seem to be more significant and robust than those of other constitutional traits.

A distinction is sometimes made between electoral rules and electoral systems. *Electoral rules* refer to the way votes translate into parliamentary seats. The two most prominent rules are majority rule (MR) and proportional representation (PR). *Electoral systems* include additional dimensions such as district size and ballot structure. Although theoretically distinct, these dimensions are highly correlated empirically: Countries relying on MR often have a minimum district size (single-member districts) and allow voting for individual candidates. Countries that favor PR often have large districts and restrict the possibility of deviating from party lists.

Duverger’s (1954) observations that MR is conducive to two-party systems whereas under PR more parties are apt to arise have been called “Duverger’s law” and “Duverger’s hypothesis,” respectively, documenting the general validity ascribed to them. Research into the economic consequences of electoral systems began much later. It has been argued (Austen-Smith 2000) that since coalition governments are more likely under PR than under MR, a common pool problem among governing parties will emerge. Parties participating in the coalition will want to please

different constituencies, which explains why both government spending and tax rates are higher under PR than under MR.

In their survey of the economic effects of electoral *systems*, Persson and Tabellini (2003) also investigate district size and ballot structure. Suppose single-member districts are combined with MR, which is often the case empirically. In this situation, a party needs only 25% of the national vote to win the elections (50% of half of the districts; Buchanan and Tullock 1962). Contrast this with a single national district that is combined with PR. Here, a party needs 50% of the national vote to win. Persson and Tabellini (2000) argue that this gives parties under a PR system a strong incentive to offer general public goods, whereas parties under MR have an incentive to focus on the swing states and promise policies that are specifically targeted at the constituents' preferences.

The effects of differences in ballot structure are the last aspect of electoral systems to be considered. Often, MR systems rely on individual candidates, whereas proportional systems rely on party lists. Party lists can be interpreted as a common pool, which means that individual candidates can be expected to invest less in their campaigns under PR than under MR. Persson and Tabellini (2000) argue that corruption and political rents should be higher, the lower the ratio between individually elected legislators and legislators delegated by their parties.

Persson and Tabellini (2003) put these conjectures to an empirical test on the basis of up to 85 countries and a period of almost four decades (1960–1998). They find the following effects: (1) In MR systems, central government expenditure is some 3% of GDP lower than in PR systems. (2) Expenditures for social services are some 2–3% lower in MR systems. (3) The budget deficit in MR systems is some 1–2% below that of systems with PR. (4) A higher proportion of individually elected candidates is associated with lower levels of perceived corruption. (5) Countries with smaller electoral districts tend to have less corruption. (6) A larger proportion of individually elected candidates is correlated with higher output per worker. (7) Countries with smaller electoral districts tend to have lower output per worker.

Blume et al. (2009a) replicate and extend Persson and Tabellini's analysis, finding that with regard to various dependent variables, district magnitude and the proportion of individually elected candidates is more significant – both substantially and statistically – than the electoral rule itself.

Form of Government: Presidential Versus Parliamentary Systems

In parliamentary systems, the head of government is subject to a vote of no confidence by parliament and can, hence, be easily replaced by the legislature. In presidential systems, the head of government is elected for a fixed term and cannot be easily replaced. Conventional wisdom holds that the degree of separation of powers is greater in presidential than in parliamentary systems as the president does not depend on the confidence of the legislature. Persson et al. (1997, 2000) argue that it is easier for legislatures to collude with the executive in parliamentary systems, which is why they expect more corruption and higher taxes in those systems than in presidential systems. They further argue that the majority (of both voters and legislators) in parliamentary systems can pass spending programs whose benefits are clearly targeted at themselves, implying that they are able to advance their interests to the detriment of the minority. This is why they predict both taxes and government expenditures to be higher in parliamentary than in presidential systems.

Empirically, Persson and Tabellini (2003) find that (1) government spending is some 6% of GDP lower in presidential systems. (2) The size of the welfare state is some 2–3% lower in presidential systems. (3) Presidential systems seem to have lower levels of corruption. (4) There are no significant differences in the level of government efficiency between the two forms of government. (5) Presidential systems appear to be a hindrance to increased productivity, but this result is significant at the 10% level only.

These results are impressive and intriguing. Although presidential systems exhibit lower government spending and suffer less from corruption than parliamentary systems, the latter enjoy

greater productivity. In a replication study, Blume et al. (2009a) find that the results are not robust, even to minor modifications. Increasing the number of observations from 80 to 92 makes the presidential dummy insignificant in explaining variation in central government expenditure. This is also the case as soon as a slightly different delineation of presidentialism is used. If the dependent variable is changed to total (instead of central) government expenditure, the dummy also becomes insignificant.

The differences in these results clarify an important point: To date, many of the effects supposedly induced by constitutional rules are not very robust; they crucially hinge upon the exact specification of the variables, the sample chosen, the control variables included, and so on. This suggests that further research needs to be as specific as possible in trying to identify possible transmission channels and to take into consideration the possibility that small differences in institutional details can have far-reaching effects.

Vertical Separation of Powers: Federalism

The vertical separation of powers has been analyzed in economics as “fiscal federalism.” Scholars of this approach largely remain within the traditional model, i.e., they assume government to be efficiency maximizing. They then ask on what governmental level public goods will be (optimally) provided, taking externalities explicitly into account. This approach need not concern us here because it does not model politicians as maximizing their own utility (for surveys, see Inman and Rubinfeld 1997; Oates 1999, 2005).

The economic benefits of federalism are thought to arise from the competition between constituent governments (i.e., from noncooperation); its costs are based in the necessity of cooperating on some issues (i.e., from cooperation). Hayek (1939) argued long ago that competition between governments reveals information on efficient ways to provide public goods. Assuming that governments have incentives to make use of that information, government efficiency should be higher in federations, *ceteris paribus*. In Tiebout’s (1956) famous model, the lower

government levels compete for taxpaying citizens, thus giving lower governments an incentive to cater to these citizens’ preferences.

Turning to possible costs of federal constitutions, Tanzi (2000) suspects that government units that provide public goods will be insufficiently specialized because there might be too many of them. Also, federal states need to deal with a moral hazard problem that does not exist in unitary states. The federal government will regularly issue “no-bail-out clauses” but they will not always be credible.

What can we learn from existing empirical studies? For a long time, the evidence concerning the effects of federalism on overall government spending was mixed. Over the last decade, though, this has changed. Rodden (2003) shows that countries in which local and state governments have the competence to set the tax base, total government expenditure is lower. Feld et al. (2003) find that more intense tax competition leads to lower public revenue.

Based on principal component analysis, Voigt and Blume (2012) conclude that institutional detail clearly matters. They find that with regard to a number of dependent variables (budget deficit, government expenditures, budget composition, government effectiveness, and two measures of corruption), frequently used federalism dummies turn out to be insignificant in explaining variation, whereas particular aspects of federalism are significant, some of them very strongly so. One problem for this research strategy is the small number of observations as there are only some 20 federal states countries in the world. One possible way of circumventing this problem is to draw on case studies instead of econometric estimates.

Direct-Democratic Versus Representative Institutions

Representatives of normative CPE ask what rules the members of a society could agree on behind a veil of uncertainty. If they agree on a democratic constitution, they would further have to specify whether and to what extent they want to combine representative with direct-democratic elements. To make an informed decision, the citizens would be interested to know

whether direct democracy institutions display systematic effects.

In most real-world societies, representative and direct democracy are complementary because it is impossible to vote on all issues directly. Among direct democracy institutions, referendums are usually distinguished from initiatives. The constitution can prescribe the use of referendums for passing certain types of legislation, in which case agenda-setting power remains with the parliament but citizen consent is required. Initiatives, in contrast, allow citizens to become agenda setters: The citizens propose legislation that will then be decided upon, given that they manage to secure a certain quorum of votes in favor of the initiative. Initiatives can aim at different levels of legislation (constitutional versus ordinary legislation), and their scope can vary immensely (e.g., some constitutions prohibit initiatives on budget-relevant issues).

These institutions can mitigate the principal-agent problem between citizen voters and politicians. If politicians dislike being corrected by their citizens, direct democracy creates incentives for politicians to implement policies that are closer to the preferences of the median voter. Initiatives further enable citizens to unpack package deals resulting from logrolling between various representatives. If citizens like only half of a deal made by politicians, they can start an initiative trying to bring down the other half.

To date, most empirical studies on the effects of direct democracy institutions have focused on within country studies, in particular with regard to the USA and Switzerland. Most studies have confirmed theoretical priors. Matsusaka (1995, 2004), e.g., finds that US states with the right to an initiative have lower expenditures and lower revenues than states without that institution. Feld and Savioz (1997) find that per capita GDP in Swiss cantons with extended democracy rights is some 5% higher than in cantons without such rights. Frey and coauthors argue that one should investigate not only the outcomes that direct-democratic institutions produce but also the political processes they induce (e.g., Frey and Stutzer 2006). Indeed there is some evidence that citizens in countries with direct democracy institutions have

better knowledge about their political institutions (Benz and Stutzer 2004) and are more interested in politics in general (Blume and Voigt 2014).

Blume et al. (2009b) is the first cross-country study to analyze the economic effects of direct democracy. They find a significant influence of direct-democratic institutions on fiscal policy variables and government efficiency, but no significant correlation between direct-democratic institutions and productivity or happiness. Institutional detail matters a great deal: While mandatory referendums appear to constrain government spending, initiatives seem to increase it. The actual use of direct-democratic institutions often has more significant effects than their potential use, implying that – contrary to what economists would expect – the direct effect of direct-democratic institutions is more relevant than its indirect effect. It is also noteworthy that the effects are usually stronger in countries with weaker democracies.

Constitutional Rules as Explanandum

Procedures for Generating Constitutional Rules

Constitutional rules can be analyzed as the outcome of the procedures that are used to bring them about. Jon Elster's (1991, 1993) research program concerning CPE puts a strong emphasis on hypotheses of this kind. He inquires about the consequences of time limits for constitutional conventions, about how constitutional conventions that simultaneously serve as legislature allocate their time between the two functions, about which effects the regular information of the public concerning the progress of the constitutional negotiations has, and about how certain supermajorities and election rules can determine the outcome of conventions.

Ginsburg et al. (2009) is a survey of both the theoretical conjectures and the available empirical evidence. They expect constitution-making processes centered on the legislature to be associated with greater post-constitutional legislative powers whereas constitution-making processes centered on the executive to be negatively correlated with such powers. Interestingly, the first half of the conjecture is not supported by the data whereas the second half is. In a book-length work on the

life span of constitutions, Elkins et al. (2009) report that public involvement in constitution-making is indeed correlated with a longer constitutional life spans.

The Relevance of Preferences and Restrictions for Generating Constitutional Rules

Procedures are a modus of aggregating inputs and can therefore never produce constitutional rules by themselves. It is thus only a logical step to analyze whether a set of potentially relevant variables can explain the choice of certain constitutional rules. There are good reasons – confirmed by some empirical evidence – to assume that (1) the individual preferences of the members of constitutional conventions will directly enter into the deliberations and that (2) the preferences of all the citizens concerned will be recognized in the final document in quite diverse ways. This would mean that rent-seeking does play a role even on the constitutional level, a conjecture that is often rejected by representatives of normative CPE.

McGuire and Ohsfeldt (1986, 1989a, b) have tried to explain the voting behavior of the Philadelphia delegates and of the delegates to the 13 states ratifying conventions that led to the US constitution. Their statistical results show that, *ceteris paribus*, merchants, western landowners, financiers, and large public-securities holders supported the new constitution, whereas debtors and slave owners opposed it (1989a, p. 175).

Explicit Versus Implicit Constitutional Change

The two approaches toward constitutions as explananda just sketched are rather static. A third approach focuses on explaining modifications of constitutions over time. Constitutional change that results in a modified document will be called explicit constitutional change here whereas constitutional change that does not result in a modified document – i.e., change that is due to a different interpretation of formally unaltered rules – will be called implicit constitutional change.

One approach toward explaining long-run explicit constitutional change focuses on changes of the relative bargaining power of organized groups. Due to a comparative advantage in using

violence (see North 1981), an autocrat is able to establish government and secure a rent from that activity. As soon as an (organized) group is convinced that its own cooperation with the autocrat is crucial for the maintenance of the rent, it will seek negotiations with the autocrat. Since the current constitution is the basis for the autocrat's ability to appropriate a rent, the opposition will strive to change it. In this approach, bargaining power is defined as the capability to inflict costs on your opponent. The prediction of this approach is that a change in (relative) bargaining power will lead to modified constitutional rules (Voigt 1999).

Future Directions

CPE has been developing very fast over the last 20 years. The availability of very large and detailed datasets has definitely contributed to this development. But many questions still need additional research. Here are some of the relevant issues.

Regarding the effects of alternative constitutional rules, we know very little regarding both human rights and constitutional amendment rules. Both can, however, be extremely important. Regarding the effects of alternative constitutional rules in general, establishing causality is a challenge. There is little variation of constitutional rules over time which makes it difficult to establish causality. A possible solution is to change the level of analysis and to leave the nation-state level and move down to the communal level where both the number of observations and the number of changes in the rules are higher.

By now, very detailed datasets concerning the *de jure* content of constitutions are available. But it is well known that the *de facto* situation in many countries does not exactly reflect the *de jure* contents of a constitution. One important area for future research in CPE would, thus, try to identify the determinants for the divergence between *de jure* provisions and *de facto* reality.

There are only a handful of papers trying to identify the determinants of the choice of concrete constitutional provisions such as the choice between a presidential and a parliamentary form

of government (Hayo and Voigt 2013; Robinson and Torvik 2008). It is almost certain that there will be more results on other constitutional rules soon. In a broader context, identifying the determinants that lead to changes from an autocratic to a democratic constitution – and vice versa – remains highly desirable.

References

- Austen-Smith D (2000) Redistributing income under proportional representation. *J Polit Econ* 108(6): 1235–1269
- Benz M, Stutzer A (2004) Are voters better informed when they have a larger say in politics? – evidence for the European Union and Switzerland. *Public Choice* 119:31–59
- Blume L, Voigt S (2014) Direct democracy and political participation. Forthcoming
- Blume L, Müller J, Voigt S, Wolf C (2009a) The economic effects of constitutions: replicating – and extending – Persson and Tabellini. *Public Choice* 139:197–225
- Blume L, Müller J, Voigt S (2009b) The economic effects of direct democracy – a first global assessment. *Public Choice* 140:431–461
- Buchanan JM (1977) Freedom in constitutional contract – perspectives of a political economist. Texas A&M University Press, College Station/London
- Buchanan J (1978) A contractarian perspective on anarchy. In: Pennock JR, Chapman JW (eds). *Anarchism*. Anarchism, New York, pp 29–42
- Buchanan JM (1987) The constitution of economic policy. *Am Econ Rev* 77:243–250
- Buchanan J (1986) Political economy and social philosophy In: ders.; Liberty, market and state – political economy in the 1980s. New York: Wheatsheaf Books, pp 261–74
- Buchanan JM, Tullock G (1962) The calculus of consent – logical foundations of constitutional democracy. University of Michigan Press, Ann Arbor
- Duverger M (1954) Political parties: their organization and activity in the modern state. Wiley, New York
- Elkins Z, Ginsburg T, Melton J (2009) The endurance of national constitutions. Cambridge University Press, Cambridge
- Elster J (1991) Arguing and bargaining in two constituent assemblies, The Storrs Lectures
- Elster J (1993) Constitution-making in Eastern Europe: rebuilding the boat in the Open Sea. *Public Adm* 71(1/2):169–217
- Feld LP, Savioz M (1997) Direct democracy matters for economic performance: an empirical investigation. *Kyklos* 50(4):507–538
- Feld L, Kirchgässner G, Schaltegger C (2003) Decentralized taxation and the size of government: evidence from Swiss state and local governments, CESifo working paper 1087, Dec
- Frey B, Stutzer A (2006) Direct democracy: designing a living constitution. In: Congleton R (ed) *Democratic constitutional design and public policy – analysis and evidence*. MIT Press, Cambridge, pp 39–80
- Ginsburg T, Elkins Z, Blount J (2009) Does the process of constitution-making matter? *Ann Rev Law Sci* 5:5.1–5.23
- Hayek F (1939) Economic conditions of inter-state federalism. *New Commonw Quart* S2:131–149
- Hayo B, Voigt S (2013) Endogenous constitutions: politics and politicians matter, economic outcomes don't. *J Econ Behav Organ* 88:47–61
- Inman R, Rubinfeld D (1997) Rethinking federalism. *J Econ Perspect* 11(4):43–64
- Matusaka J (1995) Fiscal effects of the voter initiative: evidence from the last 30 years. *J Polit Econ* 102(2):587–623
- Matusaka J (2004) For the many or the few. The initiative, public policy, and American Democracy. The University of Chicago Press, Chicago
- McGuire RA, Ohsfeldt RL (1986) An economic model of voting behavior over specific issues at the constitutional convention of 1787. *J Econ Hist* 46(1):79–111
- McGuire RA, Ohsfeldt RL (1989a) Self-Interest, agency theory, and political voting behavior: the ratification of the United States Constitution. *Am Econ Rev* 79(1): 219–234
- McGuire RA, Ohsfeldt RL (1989b) Public choice analysis and the ratification of the constitution. In: Grofman B, Wittman (Hrsg) D (eds) *The Federalist papers and the new institutionalism*. Agathon, New York, pp 175–204
- North DC (1981) Structure and change in economic history. Norton, New York
- Oates W (1999) An essay on fiscal federalism. *J Econ Lit* 37(3):1120–1149
- Oates W (2005) Toward A second-generation theory of fiscal federalism. *Int Tax Public Financ* 12(4):349–373
- Persson T, Tabellini G (2000) Political Economics – Explaining Economic Policy. Cambridge et al.: The MIT Press
- Persson T, Tabellini G (2003) The economic effects of constitutions. The MIT Press, Cambridge, MA
- Persson T, Roland G, Tabellini G (1997) Separation of powers and political accountability. *Q J Econ* 112:310–327
- Persson T, Roland G, Tabellini G (2000) Comparative politics and public finance. *J Polit Econ* 108(6):1121–1161
- Rawls J (1971) A theory of justice. Belknap, Cambridge
- Robinson J, Torvik R (2008) Endogenous Presidentialism. available at: <http://www.nber.org/papers/w14603>
- Rodden J (2003) Reviving leviathan: fiscal federalism and the growth of government. *Int Organ* 57:695–729
- Tanzi V (2000) Some politically incorrect remarks on decentralization and public finance. In: Dethier J-J (ed) *Governance, decentralization and reform in China, India and Russia*. Kluwer, Boston, pp 47–63
- Tiebout C (1956) A pure theory of local expenditures. *J Polit Econ* 64:416–424
- Voigt S (1999) Explaining constitutional change – a positive economics approach. Elgar, Cheltenham

- Voigt S (2011) Positive constitutional economics II – a survey of recent developments. *Public Choice* 146(1–2):205–256
- Voigt S (2013) Veilonomics: on the use and utility of veils in constitutional political economy. Available at: http://papers.ssm.com/sol3/papers.cfm?abstract_id=2227339
- Voigt S, Blume L (2012) The economic effects of federalism and decentralization: a cross-country assessment. *Public Choice* 151:229–254
- Wicksell K (1896) *Finanztheoretische Untersuchungen*. Jena, Fischer

Further Reading

- Voigt S (1997) Positive constitutional economics – a survey. *Public Choice* 90:11–53
- Voigt S (2011) Empirical constitutional economics: onward and upward? *J Econ Behav Organ* 80(2):319–330

Constructivism, Cultural Evolution, and Spontaneous Order

Régis Servant

PHARE, University Paris 1 Panthéon-Sorbonne,
Paris, France

Definition

This essay describes two completely different approaches which have been distinguished by Friedrich A. Hayek, the “constructivist” one and the “evolutionary” one, to the problem of how to develop institutions appropriate for the achievement of a desirable society. It does it by detailing Hayek’s analysis of the two approaches. First, the essay describes the “constructivist” contention that only institutions deliberately adopted by certain competent persons are likely to achieve a desirable society. Then, it presents the “evolutionary” viewpoint defended by Hayek, according to which a lot of beneficial institutions can be discovered only through spontaneous, undesigned growth.

Constructivism

Hayek uses the term “constructivist rationalists” to designate a large and diverse group of scholars which includes Bacon, Descartes, Hobbes,

Leibniz, Spinoza, Voltaire, Rousseau, Bentham, Austin, Hegel, Marx, Comte, Saint-Simon, and the American Institutionalists. These scholars, Hayek contends, are characterized by an arrogant overestimation of man’s intellectual capacity to achieve a desirable society. They believe that people are, or at least can become, intelligent enough to make and remake the institutions of their society as they please so as to render those institutions better suited to their wishes.

For instance, the institutionalist Mitchell (1925) describes the organized group of pecuniary institutions which makes up our economic system as “marvelously flexible” and feels confident that the progress of statistical science will increasingly enable economists to reshape that flexible system, to make it better fitted to our needs. The development of conscious experiments on group behavior and the statistical analysis of their results, may, Mitchell contends, enable humanity to guide the evolution of its institutions more wisely, through a better awareness of the relation between institutions and consequences, “to convert society’s blind fumbling for happiness into an intelligent process of experimentation” (1925, p. 8).

Hayek adds, more specifically, that, according to “constructivist rationalists,” only institutions deliberately adopted by certain competent persons are likely to achieve a desirable society and that spontaneously grown institutions whose benefits are not clearly understood ought to be rejected. Applied to the field of law, for example, such a view means that, to be desirable, a system of law ought to emanate from the deliberate enactments of expert legislators capable of knowing beforehand what laws are better for the community. Thus, under a “constructivist regime,” the development and alteration of the legal system, and of the whole framework of institutions more generally, would depend exclusively on rational foresight and understanding.

Hayek emphasizes one crucial consequence of such a view. From a political and legal standpoint, the logic of constructivism implies that people ought not to be authorized to introduce new institutions outside those deemed desirable by the ruling experts. Indeed, if one believes in the intellectual capacity of certain experts to know beforehand what institutions are better, then one is

naturally led to deny the rest of the people the legal right to try alternative institutions, on the ground that the trials would be useless anyway – a waste of time and energy. Constructivism thus tends to imply *a suppression of freedom*, or, in Hayek's words, "monopolistic power to experiment in a particular field – power which brooks no alternative and is in its essence based on a claim to the possession of superior wisdom" (1958, p. 242). In such a context, the only chance for an outsider to introduce alternative institutions would be to succeed in convincing the chosen experts that her proposal would prove superior. But, in any case, Hayek contends, it would rest exclusively with certain intelligent persons to decide, after rational reflection and argumentation, which institutions are worthy of adoption and which are not. The spontaneous growth of institutions, on the other hand, would not be tolerated, as Hayek describes in *The Constitution of Liberty*.

It is worth our while to consider for a moment what would happen if only what was agreed to be the best available knowledge were to be used in all action. If all attempts that seemed wasteful in the light of generally accepted knowledge were prohibited and only such questions asked, or such experiments tried, as seemed significant in the light of ruling opinion, mankind might well reach a point where its knowledge enabled it to predict the consequences of all conventional actions and to avoid all disappointment or failure. Man would then seem to have subjected his surroundings to his reason, for he would attempt only those things which were totally predictable in their results. (1960, pp. 37–38)

Spontaneous Order

Hayek argues that the history of mankind repudiates the constructivist view. Certain institutions – such as language, writing, the family, money, the price system, and, more generally, the institutions of the market, as well as many other rules of morality and of law – have proven beneficial by assisting people to collaborate effectively, reduce conflicts, achieve a more economical utilization of resources, increase the level of general wealth, etc. *Yet these institutions did not emanate from a deliberate intention of certain experts capable of*

knowing beforehand that they would produce such general benefits. Hayek speaks of "spontaneous growth" or "spontaneous order" to characterize that process by which a lot of beneficial institutions developed without people understanding clearly in what way they benefited their community. Scholars such as Mandeville, Vico, Hume, Burke, Smith, Ferguson, Savigny, Menger, and Maine are praised by Hayek for their "evolutionary" approach which emphasizes the indispensability of such undesigned, blindly growing institutions in the achievement of a desirable society. For instance, Menger (1883) argues that "*institutions which serve the common welfare and which are most important for its advancement can come into being without a common will directed towards establishing them*" (1883, p. 163).

One could retort to Menger, to Hayek, and to "evolutionary rationalists" more generally, that, although certain institutions which have grown spontaneously and whose benefits are not clearly understood may indeed benefit humanity, deliberately adopted institutions ought nevertheless to be exclusively preferred, for they can do even better. Yet, as explained below, Hayek maintains the exact opposite view in his theory of cultural evolution.

Cultural Evolution

Hayek presents a counterfactual history of how societies would have looked if, hitherto, the development and alteration of social institutions had been guided exclusively – or even mainly – by certain people's capacities for understanding and foresight. Under such circumstances, Hayek argues, the institutions available to people would actually have been very primitive. Not only we would be nowadays "much poorer" and "less wise," writes Hayek, "but we would also be less gentle, less moral; in fact we would still have brutally to fight each other for our very lives" (1968, p. 243).

According to Hayek, people discovered the vast majority of their beneficial institutions precisely because they did *not* endorse a constructivist rationalism. That is, they tolerated the growth

of certain institutions such as language, writing, the family, money, the price system, etc., despite the fact that the desirability, for the community, of such institutions was not, at the time, predictable or intelligible.

The chief explanatory example of Hayek's theory of cultural evolution relates to the development of the institutions of the market. People long lived within small food-sharing groups held together by highly collectivist institutions, Hayek explains, until a group of pioneers (period one, t) stumbled upon new rules of conduct, rules of the law of property, tort, contract, etc., which are at the root of the development of a market system, yet without intending thereby to achieve a more economical utilization of resources nor foreseeing the subsequent immense increase of peoples' general wealth. Indeed, those pioneers, says Hayek, "simply started some practices advantageous to them, which then did prove beneficial to the group in which they prevailed" (1979, p. 161).

Next came a second period ($t+1$) when the groups who had adopted the institutions of the market became more prosperous and more prolific than others and gradually displaced (or were imitated by) them. And only long after they grew up ($t+2$) did these superior institutions begin to be understood by a few professional economists. At the time of their development (t), their benefits as regards the utilization of resources and the level of general wealth were not, and could not have been, foreseen by anyone. This is what Hayek explained, for instance, in his Morrell Memorial Lecture on toleration published in 1987.

And to me the proof of this is that even now hardly anybody yet understands what the advantages of private property and the market society are. Man not only did not know what the advantages would be when he introduced it, he still does not understand it fully today although he owes to it (and now I am coming to one of my chief points) the possibility of multiplying the numbers of mankind by roughly 200 times. (1987, p. 40)

Hayek infers that if people had endorsed a constructivist rationalism, they would have rejected the institutions of the market from the very beginning (at t), which would have been a wrong choice. Yet as it was, deviations from constructivism enabled some groups to be selected by

cultural evolution for having adopted a superior sort of institutions. Indeed, under cultural evolution, Hayek insists, institutions are ultimately chosen by an impersonal group selection where the superior practices become known only *afterward* on the basis of what turns out to be more successful, *not beforehand* on the basis of what a body of alleged experts considers worthy of adoption. The following passage presents Hayek's evolutionary viewpoint very clearly.

Man never deliberately created the institutions of private property or the family, or understood why he accepted the moral practices that they entail. The morals of property and the family were spread, and came to dominate a large part of the world, not because those who accepted them were able rationally to convince others that they were correct, and certainly not because they themselves liked them, but because those groups who by accident did accept them prospered and multiplied more than others. Thus we owe our morals not to our intelligence but to the fact that some groups uncomprehendingly, and indeed unwillingly (for mankind has been civilized against its wishes), accepted certain rules of conduct – the rules of private property, of honesty, of the family – and thus enabled the groups practising them to prosper, multiply, and gradually to displace other groups. (1986, p. 689)

Political Consequence

That spontaneous growth may enable people to discover institutions which are better than what deliberate design can ever achieve does not imply, however, that people ought to abstain from employing their capacity of understanding and foresight – as Hayek indeed recognizes. The prescriptive conclusion of the evolutionary approach is, rather, that, whatever their degree of expertise, people ought to acknowledge that the institutions they are capable of deliberately designing may not be the best ones and that alternative institutions whose benefits cannot fully be grasped may prove to be a superior means to success. Therefore, under an "evolutionary regime" à la Hayek, by contrast to a "constructivist" one, each person would be granted the legal right not to be constructivist. That is, the law would, on the one hand, authorize each person to accept certain institutions whose benefits are not clearly understood

and, on the other hand, authorize them to brave the opinion of experts, to try new institutions without being required to rationally demonstrate to any official authority the appropriateness of her proposal to achieve a desirable society.

Future Directions

Many scholars have argued that the “evolutionary rationalism” defended by Hayek implies an apology for freedom of experiment as against government monopolies (see, for instance, Arnold (1980), Gissurarson (1987), Steele (1994), Petroni (1995), Macedo (1999), Ebenstein (2001), Steele (2002), Salle (2003), and Servant (2014, 2018)). Yet, as regards some fields such as the development and alteration of the constitution, the law, a minimum social safety net system, etc., Hayek does not advocate freedom of experiment but, on the contrary, a monopolistic power of control. That a complete abrogation of the State’s exclusive power to experiment is *not* deemed desirable by Hayek signifies that, despite his evolutionary approach, one could recognize *limits* to the appropriate domain of spontaneous order. How and where, then, to locate the boundary between monopolistic and competitively developed institutions? Moreover, an important problem arises from Hayek’s acknowledgment that people who adopt spontaneously grown institutions very often do not understand in what way they benefit their community. In particular as regards institutions which, besides their benefits, also have great costs and may at times involve great disappointments, if people cannot see why they ought to accept those beneficial institutions, the act of accepting them might cause not only an increase but, at the same time, a *decrease* of welfare. In such a context, an area for further research could concern the question of what meaning exactly ought to be attached to the term “better institutions” from an evolutionary perspective.

Cross-References

- [Austrian School of Economics](#)
- [Bounded Rationality](#)

- [Customary Law](#)
- [Hayek, Friedrich August von](#)
- [Institutional Economics](#)
- [Liberty](#)
- [Private Property: Origins](#)
- [Property Rights: Limits and Enhancements](#)
- [Rule of Law](#)

References

- Arnold R (1980) Hayek and institutional evolution. *J Libertarian Stud* 4(4):341–352
- Ebenstein A (2001) Liberty and law. In: Ebenstein A Friedrich Hayek A biography. The University of Chicago Press, Chicago, pp 223–226
- Gissurarson H (1987) Three liberal criticisms of traditionalism. In: Gissurarson H Hayek’s conservative liberalism. Garland, New York, pp 126–132
- Hayek F (1958) Freedom, reason, and tradition. *Ethics* 68(4):229–245
- Hayek F (1960) The constitution of liberty. University of Chicago press, Chicago
- Hayek F (1968) Speech on the 70th birthday of Leonard Reed. In: What’s past is prologue: a commemorative evening to the foundation for economic education on the occasion of Leonard Read’s seventieth birthday. Foundation for Economic Education, New York, pp 37–43
- Hayek F (1979) Law, legislation and liberty, vol. 3: the political order of a free people. Routledge & Kegan Paul, London
- Hayek F (1986) Cultural evolution: the presumption of reason. *World & I* 1(2):683–690
- Hayek F (1987) Individual and collective aims. In: Mendus S, Edwards D (eds) On toleration. Clarendon Press, Oxford, pp 35–47
- Macedo S (1999) Hayek’s liberal legacy. *Cato J* 19(2): 289–300
- Menger C (1883) Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Oekonomie insbesondere. Verlag von Duncker & Humblot, Leipzig
- Mitchell W (1925) Quantitative analysis in economic theory. *Am Econ Rev* 15(1):1–12
- Petroni A (1995) What is right with Hayek’s ethical theory. *Rev Eur Sci Sociales* 33(100):89–126
- Salle C (2003) Fin de l’Histoire et légitimité du droit dans l’œuvre de F. A von Hayek. *Rev Fr Sci Polit* 53(1): 127–166
- Servant R (2014) Libéralisme, socialisme et État providence : la théorie hayékienne de l’évolution culturelle est-elle cohérente ? *Rev économique* 65:373–390
- Servant R (2018) Spontaneous growth, use of reason, and constitutional design: Is F. A. Hayek’s social thought consistent? *J Hist Econ Thought*, forthcoming
- Steele G (1994) On the internal consistency of Hayek’s evolutionary oriented constitutional economics: a comment. *J Économistes Études Hum* 5(1):157–164

Steele G (2002) Hayek's liberalism and its origins: his idea of spontaneous order and the Scottish enlightenment. *Christinia Petsoulas. Q J Austrian Econ* 5(1):93–95

Further Reading

- Boudreaux D (2014) Legislation is distinct from law. In: Boudreaux D *The essential Hayek*. Fraser Institute, Canada, pp 32–38
- Buchanan J (1977) Law and the invisible hand. In: Buchanan J *Freedom in constitutional contract: perspectives of a political economist*. Texas A&M University Press, College Station, pp 25–39
- Diamond A (1980) F. A. Hayek on constructivism and ethics. *J Libertarian Stud* 4(4):353–365
- Hayek F (1946) Individualism: true and false. Hodges, Figgis & Co. Ltd., Dublin
- Hayek F (1965) Kinds of rationalism. *Econ Stud Q* 15(2):1–12
- Hayek F (1966) Lecture on a master mind: Dr. Bernard Mandeville. *Proc Br Acad* 52:125–141
- Hayek F (1967) Notes on the evolution of systems of rules of conduct. In: Hayek F *Studies in philosophy, politics and economics*. University of Chicago press, Chicago, pp 66–81
- Hayek F (1973) Law, legislation and liberty, vol. 1: rules and order. Routledge & Kegan Paul, London
- Hayek F (1988) The fatal conceit: the errors of socialism. Routledge, London
- Kirzner I (1990) Knowledge problems and their solutions: some relevant distinctions. *Cult Dyn* 3(1):32–48
- Moroni S (2014) Two different theories of two distinct spontaneous phenomena: orders of actions and evolution of institutions in Hayek. *Cosmos + Taxis* 1(2):9–23
- Nadeau R (2016) Cultural evolution, group selection and methodological individualism: a plea for Hayek. *Cosmos + Taxis* 3(2):9–22
- O'Driscoll G (2015) Hayek and the scots on liberty. *J Priv Enterp* 30(2):1–19
- Smith V (2008) Rationality in economics: constructivist and ecological forms. Cambridge University Press, Cambridge
- Smith C (2014) Hayek and spontaneous order. In: Garrison R, Barry N (eds) *Elgar companion to Hayekian economics*. Edward Elgar Publishing Limited, Cheltenham, pp 224–245

recognized that agents, namely consumers, are endowed with a bounded rationality. Consumer biases cover a wide range of behaviors, such as quality misperception, status quo bias, projection bias, inertia, and can have various consequences on the market equilibrium. The aftermaths of consumer misperception depend on the type of bias one considers, as well as on the market structure. This entry presents a typology of consumer biases and mentions possible consequences on the market outcome. Policy recommendations to fight against consumer biases, as well as the main counterarguments put forward by libertarians, will finally be discussed.

Definition

Bounded rationality refers to the fact that agents depart in a systematic way from the perfect rationality assumption which prevails in the economic literature (for a more general approach of the topic, see the entry on “► [Bounded Rationality](#)”). As summarized by Herbert Simon, the ambitious aim of behavioral models of rational choice is to “replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist” (Simon 1955, p. 99).

While the issue of bounded rationality is not constrained to the case of consumers (see for instance Rabin 2002 and Kahneman 2011), the latter are particularly prone to various kinds of cognitive biases, because of the context as well as the intrinsic particularities of consumption decisions (Korobkin 2003). Consumers are bound to make numerous and complex decisions, which leads them to use simplifying heuristics. Moreover, they face sophisticated and rational firms, who are likely to enhance or even exploit their weaknesses. Hence, consumer behavior is a particularly prosperous field for behavioral biases. Recent developments in Law and Economics have therefore brought the issue of consumer decision-making back to the center stage. The main

Consumer Bias

Sophie Bienenstock
EconomiX, Université Paris Nanterre, Paris,
France

Abstract

Since the founding work of Simon (1955) and Kahneman and Tversky (1974, 1986), it is

concern is to understand consumer behavior in order to suggest relevant responses in terms of public policy. While the existence of consumer biases is unanimously recognized, the issue of whether and how the regulator should intervene remains debated.

Theoretical Work on Consumer Bias

The main issues tackled in the literature dedicated to consumer biases are twofold:

- First, one needs to describe and understand the consequences of consumers' cognitive limitations. What impact do biases have on consumer choice, and consequently on competition and pricing? How does the market equilibrium change in the presence of consumer biases?
- Once this first step has been addressed, one can tackle the issue of fighting against consumer biases. The natural question that comes to mind is whether inefficiencies linked to consumer biases can be reduced by increasing competition. Will biased consumer behavior be overcome through learning or education?

Such concerns have become central in the economic literature, to the point that some authors assert that "the rational firm-irrational consumer assumption has become the norm, and the question of what firms do to exploit irrationality is often the primary focus" (Ellison 2006). Yet, describing consumer biases, as well as their consequences on the market, is not an easy task. Behind the notion of consumer biases lies a great diversity of behaviors. Consumer biases are numerous and do not refer to a unique cognitive phenomenon. Hence, classifying consumer biases is an essential step towards understanding their economic consequences.

Classifying Consumer Biases Although classifying the numerous biases is an unrelenting and complex task, Huck and Zhou offer a simple typology. The authors identify three dimensions along which consumer choices might be biased (Huck and Zhou 2011).

- First, the willingness to pay bias describes a situation in which agents pay too much for a given quantity of a good. For instance, according to the reference point effect, an agent's willingness to pay depends on a reference point (the status quo, past experience, other products, expectations, etc.). The reference point might lead to an irrational increase in the agent's willingness to pay. Willingness to pay bias can also be due to a misperception of future desired attributes. As in the famous study of DellaVigna and Malmendier (2006), consumers might believe that they will go to the gym more than they actually will. This misperception can be analyzed as over-optimism and willingness to pay bias. Similarly, the common knowledge according to which shopping on an empty stomach leads to overconsumption is another expression of such misperception (Loewenstein et al. 2003).
- Second, a search bias occurs when consumers do not choose the best-suited product because they do not search in a rational way. Consumers might for example be subject to inertia, which is to some extent an expression of the well-known status quo bias. More generally, inertia describes consumers who tend to buy in one of the first shop they see and/or to stick to the same seller in case of repeated purchase, although more advantageous offers might be available. Another form of search bias is price misperception. In the presence of complex price schemes (such as partitioned pricing, drip pricing, baiting, discounts) consumers are likely to misjudge the final price. In response to price misperception, firms might have incentives to adopt artificially complex pricing systems (Spiegler 2006).
- Finally, quality biases refer to any situation in which consumers purchase a quality not fit for their needs. The error can be due to a misperception of the intrinsic quality of the product, or to inaccurate anticipations about one's own needs and capacities to use a product. This observation leads us to the interesting dichotomy proposed by Köszegi (2014): false beliefs either relate to "the contract itself," or to "her own behavior given the contract" (p. 1104).

Modeling Consumer Biases Since consumer biases cover a large scope of behaviors, there is no general model of consumer decision-making in the presence of cognitive bias. Studying consumer biases requires building an ad hoc model, which depends both on the kind of cognitive flaw one wants to describe and on the market structure one considers. Therefore, theoretical articles dedicated to consumer biases generally focus on one specific type of irrationality in a given market structure. One should always keep in mind that the conclusions are necessarily context-dependent and that formulating a general theory of consumer bias is by essence very difficult.

Nonetheless, several papers make a significant contribution in the understanding of consumer biases. For instance, hyperbolic discounting is at the center of several papers by DellaVigna and Malmendier (2004, 2006). Time inconsistent preferences have been studied by Eliaz and Spiegler (2006), while the framing effect is analyzed by Piccione and Spiegler (2012).

The consequences of consumer biases depend on numerous parameters. Yet, authors generally come to the conclusion that consumer biases are detrimental to social welfare because they result in “behavioral market failures” (Bar-Gill 2011). The concept of “behavioral market failures” was coined by Oren Bar-Gill (2011) to describe the deficiencies of market mechanisms in the presence of boundedly rational consumers. Hence, the issue of whether and how one should intervene to constrain the aftermaths of consumer biases naturally comes to mind.

Policy Implications: Should the Regulator Intervene?

In the broad lines, two kinds of responses to consumer biases are conceivable: soft paternalism on the one hand, and debiasing, on the other hand. While they both strive towards a common goal, they differ in the method they use.

Soft Paternalism The notion of soft paternalism, also known as asymmetric paternalism (Camerer

et al. 2003) or libertarian paternalism (Sunstein and Thaler 2003), refers to any legal intervention aimed at protecting agents without encroaching on individual freedom. The concept has been thoroughly studied by Sunstein and Thaler (2008) in their book *Nudge: Improving Decisions About Health, Wealth and Happiness*. Soft paternalism claims to be both ostensibly paternalistic, in that it helps people make decisions that are in their best interest, and libertarian, in that it preserves freedom of choice.

Debiasing Debiasing consists in revealing errors to each agent, so that they can correct their behavior on their own (Jolls and Sunstein 2006). Debiasing differs from paternalism insofar as consumers are fully aware of the process they undergo. The ultimate objective is to give agents the capabilities to act by themselves. Sunstein and Thaler (2008) clearly sum up the concept of debiasing in what they call “RECAP Policies” (Record, Evaluate, Compare, Alternative Prices). While soft paternalism relies on manipulation, debiasing rests on increased transparency. Both raise vivid criticisms from libertarians.

The Libertarian Criticisms The first main criticism, addressed mainly to soft paternalism, concerns the complexity of carrying out a welfare analysis in the presence of consumer bias. According to the libertarian view, the regulator does not have the relevant information to determine the agents’ true preferences in the presence of changing utility functions or inconsistent choices. In this line of thought, Saint-Paul (2011) claims that “it is impossible, in fact, to establish such a result, for one needs a criterion for comparing alternative utility functions; that is, one would have to impose some ‘meta-utility function’ in order to tell us that a given utility function is better than another” (p. 87). Yet, such a meta-utility function does not exist, which renders any assessment on welfare impossible in the presence of changing preferences.

The second main argument can be summed up as the fear of a “slippery slope,” according to which there is a natural tendency to go towards

more paternalism. This slippery slope, which would lead to strong paternalist measures and to the denial of individual freedom, has been put forward by Whitman and Rizzo (2007, 2009).

Beyond the slippery slope argument, also lies the idea that too much regulation would inhibit learning, and ultimately be counterproductive. A systematic intervention to guide citizens towards what is considered to be a good decision would remove all opportunities to make errors. In this line of thought, Klick and Mitchell (2006) allege that regulation leads to a vicious circle of more regulation, which ultimately increases biases. In this approach, cognitive biases are considered to be endogenous, insofar as they depend on the existing regulation.

The last main argument put forward by libertarians is that legal interventions to counter the effects of cognitive biases are useless, since the market remains efficient even if agents are not perfectly rational. For instance, Sugden (2008) contends that the market is an efficient way of allocating resources even if consumers exhibit inconsistent preferences. Sugden's key argument lies in the fact that firms always have incentives to cater to consumer demand, in spite of potentially inconsistent preferences. The idea that the market is the best response to consumer bias has also been suggested by Bebachuk and Posner (2006).

Conclusion

While the ubiquity of consumer biases cannot be denied, the appropriate response remains debated. In various countries, consumer policy is slowly starting to take into consideration the presence of consumer biases, for instance, by using the framing effect in order to steer consumers towards healthy foods. Yet, such measures remain sparse. One of the reasons is that no general policy to fight consumer bias can be enacted, since every situation requires an ad hoc analysis. In spite of the work that remains to be done, we believe that several powerful and simple tools can be used to protect consumers against their own flaws.

Cross-References

- Behavioral Law and Economics
- Bounded Rationality
- Consumption
- Contract, Freedom of
- Endowment Effect
- Harmonization: Consumer Protection
- Naïve Consumers: Contract Economics
- Rationality

References

- Bar-Gill O (2011) Competition and consumer protection: a behavioral economics account. *Law & Economics Research paper series, working-paper no 11–42*
- Bebchuk L, Posner R (2006) One-sided contracts in competitive consumer markets. *Mich Law Rev* 104: 827–836
- Camerer C, Issacharoff S, Loewenstein G, O'Donoghue T, Rabin M (2003) Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". *Uni Pennsylvania Law Rev* 3(151):1211–1254
- DellaVigna S, Malmendier U (2004) Contract design and self-control: theory and evidence. *Q J Econ* 119(2): 353–402
- DellaVigna S, Malmendier U (2006) Paying not to go to the gym. *Am Econ Rev* 96(3):694–719
- Eliasz K, Spiegel R (2006) Contracting with diversely naive agents. *Rev Econ Stud* 73(3):689–714
- Ellison G (2006) Bounded rationality in industrial organization. In: Blundell R, Newey WK & Persson T (eds), 'Advances in economics and econometrics: theory and applications', Vol. 2 of Ninth World Congress, Cambridge university Press, pp. 142–219. <http://economics.mit.edu/files/904>
- Huck S, Zhou J (2011) Consumer behavioral biases in competition: a survey, Final Report OFT1324, Office of Fair Trading
- Jolls C, Sunstein C (2006) Debiasing through law. *J Leg Stud* 35(1):199–242
- Kahneman D (2011) *Thinking fast and slow*. Farrar, Staus and Giroux, New York
- Kahneman D, Tversky A (1974) Judgment under uncertainty: heuristics and biases. *Science* 185(4157): 1124–1131
- Kahneman D, Tversky A (1986) Rational choice and the framing of decisions. *J Bus* 59(4):S251–S278
- Klick J, Mitchell G (2006) Government regulation of irrationality: moral and cognitive hazards. *Minnesota Law Rev* 90:1620–1663
- Korobkin R (2003) Bounded rationality, standard form contracts, and unconscionability. *Uni Chicago Law Rev* 70(4):1203–1295

- Köszegi B (2014) Behavioral contract theory. *J Econ Lit* 52(4):1075–1118
- Loewenstein G, O'Donoghue T, Rabin M (2003) Projection bias in predicting future utility. *Q J Econ* 118(4):1209–1248
- Piccione M, Spiegler R (2012) Price competition under limited comparability. *Q J Econ* 127:97–135
- Rabin M (2002) A perspective in psychology and economics. *Eur Econ Rev* 46:657–685
- Saint-Paul G (2011) The tyranny of utility: behavioral social science and the rise of paternalism. Princeton University Press, Princeton
- Simon H (1955) A behavioral model of rational choice. *Q J Econ* 69(1):99–118
- Spiegler R (2006) Competing over agents with boundedly rational expectations. *Theor Econ* 1(2): 207–231
- Sugden R (2008) Why incoherent preferences do not justify paternalism. *Constit Polit Econ* 19: 226–248
- Sunstein C, Thaler R (2003) Libertarian paternalism. *Am Econ Rev* 2(93):175–179
- Sunstein C, Thaler R (2008) *Nudge: improving decisions about health, wealth and happiness*. Yale University Press, New Haven, CT, USA
- Whitman DG, Rizzo M (2007) Paternalist slopes', *NYU. J Law Econ* 2:411–443
- Whitman DG, Rizzo M (2009) The knowledge problem of new paternalism. *Brigham Young Uni Law Rev* 4:904–968. <http://digitalcommons.law.byu.edu/lawreview/vol2009/iss4/4/>

Further Reading

- Bernheim D, Rangel A (2009) Beyond revealed preferences: choice-theoretic foundations for behavioral welfare economics. *Q J Econ* 124(1):51–104
- Rabin M (2002) A perspective in psychology and economics. *Eur Econ Rev* 46:657–685
- Rachlinski J (2003) The uncertain psychological case for paternalism. *Northwest Univ Law Rev* 97(3):1165–1225
- Spiegler R (2011) *Bounded rationality in industrial organization*. Oxford University Press, Oxford

Consumer Law

- [Harmonization: Consumer Protection](#)

Consumer Legislation

- [Harmonization: Consumer Protection](#)

Consumer Policy

- [Harmonization: Consumer Protection](#)

Consumption

Fav Tsoin Lai

National Chi Nan University, Puli, Taiwan

Abstract

Consumption involves consumer behavior of how people make choice in selecting to enjoy bundle with specific quantities of one or more goods and services within a period. Economists use an often-used word “demand” to term peoples’ choices when consuming a good, and model the choices beginning with an account of consumer preferences over commodity bundles. Goods are categorized as normal or inferior by the responses of consumers’ demands to the income changes, and economists break the effect of price changes on demands into two parts, namely, substitution and income effects. People have their own time preference over the consumptions in different periods. Two cases are illustrated: one is that current consumptions are the substitutes for future consumptions, and the other is that the pleasure of current consumption is enhanced by the past consumptions. Decisions people make regarding consumptions not only occur in the economic situations with certainty but also with uncertainty. Most of economists nail uncertainty down as risk, and it refers to situations, in which consumers can list all possible outcomes and subjectively know the likelihood of each occurring. The ways that people rank plans of consumption under uncertainty are similar to the ones that people have preferences over consumption bundles under certainty. An individual may be not only concerned with his/her own self, but also be connected to the selves of the others, and hence his/her demand for some goods could depend on the

demands on the part of other consumers. In the case of positive network externalities, the interdependence preferences boost the demand for a good or service, but in the case of negative ones they reduce the consumption.

Definition

The using up of goods or services by individuals' choice.

Consumption can be understood as the using up of goods or services by individuals' choice. It involves consumer behavior of how people make choices in selecting to enjoy bundle with specific quantities of one or more goods and services within a period. Economists use an often-used word "demand" to term peoples' choices when consuming a good. The individual's demand for a good can be independent of another person's and just associated with his/her own tastes and income and the prices of goods. However, for some goods, one person's demands are related to the demand of other people. While discussing consumers' demands for goods, most of microeconomics textbooks start from and focus on the choices that are not associated with social interactions.

The scholars of economics develop a practical way to describe the reasons why an individual makes a specific good or service not others; they clarify what is affordable, describe the assumptions on which preferences are based, present how consumers make consumption decisions, and analyze the characteristics of demands. All the combinations of goods that the consumer may choose are referred to as consumption bundles or market bundles. Each bundle is a list with specific quantities of goods and services. Consumers can afford the bundles that do not cost more than his/her income, and the collection of such affordable bundles is referred to as budget set. The relative price (price ratio) can be a measure that fathoms the opportunity cost of consuming one good in terms of the other good. In a well-organized market, it is a kind of objective exchange rate in which most of the people will trade one good for another.

Consumer Behavior

Some economists believe that their impositions on the ordering of consumer preferences hold for most people in most situations. The first one is completeness. Consumers are able to compare any two bundles, rank them, and hence make a choice between them. The second one is that consumer preferences are transitive. It says that if bundle A is preferred to bundle B and bundle B is preferred to C, then it must be the case that A is preferred to C. Transitivity is a hypothesis about people's choice behavior and is normally regarded as necessary for consumer consistency. People, who violate this condition, have circular preference that they cannot have the same attitude toward a thing and are often contradictory when dealing with the logically connected events. Usually, economists believe that these two characteristics, completeness and transitivity, are the most important parts of consumers' rationality. (For more details about the assumptions about consumers' preferences, refer to Varian (1992), Mas-Colell et al. (1995), and Pindyck and Rubinfeld (2012a)).

As long as the goods do not satiate the consumers, they always prefer more of any goods as opposed to less. In this case, consumer preferences are referred as monotonic by economists. There could be further assumptions regarding consumer preferences, which make economists' practices proceed more smoothly. For example, well-behaved preference has the traits: averages are preferred to extremes, and the ordering of consumer preferences is continuous. The former trait seems to be a usual observation of ordinary peoples' preference, and the latter indicates that a small change in the combination of goods will not cause a significant disturbance in their ordering of preferences. All these assumptions provide the classic, preference-based approach to illustrate how a person's actual consumption is determined.

After recognizing that consumer's choice over bundles is only related to the rank of the satisfactions between two bundles and not the metrics of satisfactions, economists construct a utility function in such a way that the number assigned to the more-preferred bundle is larger than the one

assigned to the less-preferred bundle. The utility function presents the order of preferences over the bundles, not how much one bundle is preferred to the other, so that the magnitude of the difference between the utilities of the two bundles does not matter. What matters is the order of preferences, not the strength of people's preference. When the quantity of good changes by a very small amount, the number assigned will be changed; the ratio of the change in utility to the quantity change of the good is referred as marginal utility with respect to it. Since the magnitude of utility is for ordinal function, the marginal utility does not have behavior content. All the bundles that have the same utility constitute a set in the commodity space that is referred as indifference curve by economists. Any bundle in the indifference curve has the same utility and hence can be used to establish a measure that is called marginal rate of substitution – the maximum amount of one good that a person is willing to give up to obtain an additional unit of the other good. It is a consumer's subjective exchange rate at which a consumer is willing to trade one good for the other.

The preference of a consumer is subjective; everyone has his/her own taste and ranks his/her preference over consumption bundles. The diversity of personal tastes gives rise to a range of preference orderings that are very different from each other. Economists insist that no metric exists enabling one to make interpersonal comparison of well-being. The difference between the utility a person assigns to a bundle and the utility the other person assigns to it tells nothing about what the difference between their satisfactions is. An individual with subjective preference can compare different groups of items available for purchase and rank all of them, but there are many restrictions on the quantities of goods they can buy. Limited real incomes are the foremost constraint that forces consumers to choose among alternatives. Consumers consider not only nominal income but also the prices of goods because the amount of money they spend on the goods should be no more than the total amount they can spend. Higher income and lower prices make consumers' budget set larger and allow them to afford more. Given their rationality and limited real income,

consumers make optimal choice over affordable set; that is, they choose the best to consume and maximize their utility. Consumers' consumption or demand is a function of prices and their incomes. When a consumer makes a choice involving all kinds of goods, his/her subjective exchange rate will be equal to the objective exchange rate. This is a marginal condition for consumers' consumption choice – the marginal rate of substitution is equal to the ratio of prices.

Consumers are concerned with the changes in their incomes and the prices of goods, because these changes in economic environment affect what they can afford. When their budget sets are changed, consumers will adjust their choice to make themselves as good as possible. Economists examine the impacts of these changes on consumers' demand independently. In the case where the prices remain unchanged, the goods are described as normal: consumers want to buy more of the goods as their incomes increase. Some economists define normal goods further according to consumers' demand responses to the income changes. If the demand for a good goes up by a greater proportion than income (the demand elasticity of income is greater than one), it is called luxury good, and if it goes up by a smaller proportion than income, it is called necessary good (the demand elasticity of income is less than one). Alternatively, such a good is called inferior good: an increase of income results in a reduction in the consumption of the good. As a person's purchasing power increases, his/her satisfaction from the consumption should not decrease. Thus, for a well-behaved preference, it is impossible for all goods to be inferior.

Price adjustments for a good not only change the market exchange rates between it and the other goods but also affect consumers' real purchasing power. Economists break up the effect of price changes on consumers' demands for goods into two parts, namely, substitution and income effects. A good becomes cheaper as its price falls, and consumers will tend to buy more of it, but the other goods become relatively more expensive now, and consumers will be inclined to buy less of them. The change in price allows consumers to substitute cheaper goods for the

relatively expensive goods; economists describe this kind of response as the substitution effect, and it always moves in the opposite directions to the price changes. A fall in the price of a good not only allows consumers to buy the original choice but also enables them to afford extra amount of goods in their money income. This is equivalent to the movement that occurs when purchasing power goes up while the relative prices remain constant. Consumers would adjust their demand following the changes in their real income; this response is called the income effect. If the good is normal, the increases of purchasing power that is due to the fall in its price will lead to an increase in its demand, and then the direction of income effect will be opposite to the movement of its price. However, if the good is inferior, the direction of income effect will be the same as the movement in price. There are no definite signs for income effect. For a good, the impact of its price changes on its demand is related to how consumers' choices in regard to consumption respond to income change. If it is a normal good, income effect should reinforce the substitution effect, and its demand for the good must increase when its price decreases. However if it is an inferior good, the income effect has the movement that is opposite to the one for substitution effect, and the demand for the good does not always increase when its price falls. For an inferior good, if income effect is strong enough to dominate substitution effect, a fall in its price trumps consumers' demands for it. In considering this event, economists call such a good a Giffen good. The income effect is usually small compared to the substitution effect because most of the expenditures on individual good make up a small part of consumers' budgets. In addition large income effects are often related to normal good, seldom to inferior good. Thus, in the real world, Giffen goods are rarely to be encountered, and instead it is ubiquitous for people to consume more of a good as its price falls (Pindyck and Rubinfeld (2012a) *Microeconomics*, Pearson, Taipei).

For many goods, the demand for one good is often associated with the consumptions and prices of other goods. For the goods like rice and wheat

that have some similar characteristics, when one good gets more expensive, the consumer switches to consume the other good; the consumer substitutes away from the more expensive good to the cheaper one. Two goods are substitutes if an increase in the price of one leads to an increase in the quantity demanded for the other. To some consumers, the pleasure of consuming a cup of coffee can be enhanced when they can consume with one or two spoons of sugar. Sugar and coffee are consumed together and complement each other. When sugar gets more expensive, the consumer not only consumes less sugar but also consumes less coffee; two goods are complements if an increase in the price of one leads to a decrease in the quantity demanded for the other.

Intertemporal Choice

In their life span, people have income streams that do not remain constant and have their own time preference over the consumptions in different periods, and hence their choices of consumptions are involved in saving and consuming over time. The returns of hoarding are the relative prices of consumption between periods and are equal to the principal of any amounts saved plus interest rate, and with income streams they form consumers' budget constraints. How much a consumer is willing to substitute consumption today for consumption tomorrow depends on his/her time preferences that are constituted by his/her particular patterns of consumptions over time. Time preference is a main factor that determines the intertemporal marginal rate of substitution at which a consumer is just on the margin of being willing to substitute current consumption for future consumption. Most economists take rates of time preference as given or exogenous for convenience in dealing with the subjects they focus on, but some economists think they could be changed by consumers' choices over time. (Uzawa (1968), Lucas and Stokey (1984), and Epstein (1987) consider the possibility that the rate of time preference ultimately depends on consumption flows. Becker and Mulligan (1997) postulate a model, in which a consumer can make

an effort to reduce the discount on future utilities.) A consumer who doesn't care whether he/she consumed today or tomorrow is such a patient person that there is no time discount between his/her consumptions in different periods; future consumptions can be perfectly substituted for current consumptions. The people, who value current consumption more than future consumption, have time impatience with delaying consuming a certain bundle of goods or services, there is a time discount against the values of their future consumptions. As what the consumers' choices in a static setup are, there exist substitutions between the consumptions at different periods, and it is also possible that consumptions can be complemented for each other.

If there is a complementary between the consumptions of a particular good at different times, the pleasure of current consumption is enhanced by the past consumption. Goods like cigarette and heroin are potentially addictive, because for a large number of consumers who consume these goods, their past consumption complements current consumptions (Becker and Murphy 1988). People that become addicted into smoking cigarettes will increase their current consumption if there is an increase in their past consumption. Because of the complementary between the consumptions today and that of tomorrow, a price increase in the price of tomorrow consumption will lead to a decrease in the consumption today. For example, an addicted smoker that anticipates there will be a tax on smoking may reduce his/her current consumption of cigarettes. The choices in regard to consumptions over time are very similar to the choices of consumptions over several goods within one period; there exists a complementary between the current and past consumptions. This can also be seen by the observation that environmental cues affect consumer behavior. The smell of cookies being baked, sound of ice cube falling into a whisky tumbler, and sight of a pack of cigarettes could be the cues for consuming the goods associated with them (Laibson 2001). Cigarette addicts feel a keen desire for nicotine when they see smoking cues, like an open box of cigarettes. Addict formation effects are triggered and halted by the occurrence and nonappearance of

cues that have been related with the past consumption of addict forming goods.

Uncertainty

People's decisions makings regarding consumptions discussed in the previous paragraphs have been implicitly assumed to occur in the economic situations with certainty. However, there is considerable uncertainty involved in everyday life of people. Uncertainty can be interpreted broadly; however, most of the economists nail it down as risk, which refers to situations in which consumers can list all possible outcomes and subjectively know the likelihood of each occurring. Then, uncertainty is associated with the probability distribution that consists of a list of different outcomes and the subjective probability associated with each outcome. According to his/her experience and judgment, a person evaluates the possibility that an outcome will occur to form his/her subjective probability, and hence it is not necessary that this probability is the same as the rate of recurrence with which this outcome has actually occurred in the past. Since people subjectively gauge the likelihood that an outcome occurs, they may have different probability distributions over outcomes and different decision-making.

People can rank probability distributions as they can do the bundles of goods under certainty; that is, the way that consumers have preferences for various goods can be applied to the one in which consumers have preferences for different probability distributions. When an economic environment is characterized by uncertainty, a person will attach different values on a consumption bundle in different circumstances under which the consumption turns out to be reachable. Before they are sure which state of the world will occur, people have a plan that they will follow to have their consumption while facing different outcomes. An outcome of a random event can be interpreted as a state of nature, and what people actually consume is contingent on it. A plan for consumption under uncertainty constitutes of probabilities and outcomes, and it is a risky

alternative. Since each outcome gives an individual a level of satisfaction, he/she can evaluate every risky alternative according to his/her expectations and rank all risky alternatives. The ways that people rank plans of consumption under uncertainty are similar to the ones that people have preferences over consumption bundles under certainty. The difference between the consumer's choice under certainty and the one with uncertainty is the independent assumption for the outcomes under uncertainty. What an individual intends to consume in one state is independent of the choices he/she makes in other states, since the consumption choices in the different states of nature cannot coexist. Except this, the theory of consumers' choice under certainty can be applicable for analyzing people's choices among different risky alternatives. Not only can an individual's rationality, completeness, and transitivity be applied to his/her preference over risky alternatives, but also the other assumptions regarding the preferences over actual bundles are also applicable to them under a particular structure. Each outcome corresponds with consumption and can be assigned a number to represent the pleasure it brings up to a consumer. Then weighting the number by the probability of its occurrence and summing all the weighted values establish a consumer's expected utility of a risky alternative.

In the presence of uncertainty, there are two important measures that are used to describe a risky alternative, expected value and variability. Each outcome is associated with a payoff and a probability; the expected value of a risky alternative is the sum of payoffs being weighted by probabilities. The expected value is the payoff that an individual would expect on average, and hence it reveals the main propensity for the distribution of outcomes. The variability of a risky alternative is the spread of possible outcomes and tells the association with the riskiness of a random event. Most people love the risky alternative with high expected value and low variability. In addition, an individual's expected satisfaction about a risky alternative is related to but not necessarily proportional to its expected monetary values. People's attitudes toward risk determine

their choices over risky alternatives and therefore their actual consumption. Economists categorize people's willingness to bear risk by identifying whether an individual ranks a certain income over an uncertain income with the same expected value. (This definition of consumers' preferences toward risk can be seen in Pindyck and Rubinfeld (2012b)). An individual who is risk averse prefers a certain income to a risky income with the same expected value; that is, he/she would rather have the expected value of his/her wealth rather own the lottery. A person who is risk neutral is indifferent between a certain income and an uncertain income with the same expected value. An individual who is risk loving prefers an uncertain income to a certain one, even if the expected value of uncertain income is less than that of certain income. However, some experiments to see the extent to which people's attitude toward risk fit those three propensities reveal that most people have the propensity to prefer avoiding losses over acquiring gains. This kind of multifaceted feelings about losses and gains is depicted by Kahneman and Tversky (Kahneman and Tversky 1979) as "loss aversion," which is often seen in inexperienced investors but not so often in experienced investors (List 2003).

Interdependence and Consumer Behavior

In the case where consumers have preferences independent of one another, their demands are only related to the prices of goods and their own incomes and tastes. However, an individual may be not only concerned with his/her own self but also be connected to the selves of the others, and hence his/her demand for some goods could depend on the demands on the part of other consumers. If the interdependence preferences boost the individual demand for a good or service, a positive network externality exists – the quantity of a good demanded by a usual consumer increases as the demands of other consumers increase (Leibenstein 1950). When people watch a game or play, they jointly consume the service it supplies in the presence of others, and this activity

becomes more worth having when they can be shared with a group of peers. Similarly, some social activities, like restaurant eating, attending a concert, talking about books, and chatting about the booming of stock markets, let a consumer experience more enjoyment the more that these activities are participated in by others (Becker 1991). Those have the common feature of a bandwagon effect – the longing for being fashionable, for owning a good because everyone has it, or to derive unconstrained pleasure from a fashion. The pleasure from a good is greater when many people want to consume it, perhaps because a person wishes to be in the same step with what is in fashion or because he/she is more certain that the food, writing, or performance has its quality when a restaurant, book, or theater is more popular. Moreover, because people take part in many activities with their families, coworkers, neighbors, and friends, interdependent preferences can be spread across and through several networks, which are channeled by those persons close to them.

Network externalities are sometimes negative. A specially designed sports car not only supplies a service of transportation but also brings prestige to its owner; it is a conspicuous good that is characterized by its exclusivity and provides its consumers with a signal of status, which is enhanced by material displays of wealth. Consumers purchase conspicuous good to satisfy their material needs and social needs as provided by the product. Conspicuous consumption is such an activity that consumers' indulgence in it is recognized by their peers and differentiates their consumption from that of the other groups. The desire to own the uniqueness and exclusivity of a conspicuous good motivates consumers to purchase it and produces a negative network externality, which is called snob effect by some economists (Leibenstein 1950). A branded watch has its instrumental function for its consumers, but what is more valuable is the prestige, status, and exclusivity resulting from the fact that not many people own one like it. When more of the others consume the conspicuous good, its status value declines. Thus an individual demand for a snob good decreases as the more people own it, and

there is a negative feedback between individual consumption and aggregate consumption. An increase in the price of the good will reduce consumption but will enhance the status value of the good.

Since a conspicuous good signals the status of wealth, its price is much higher than the value of its instrumental function. Most of the conspicuous goods have branded names to show their exclusivity and to assure quality. The counterfeits of conspicuous goods do not have the same quality as genuine ones but have the appearance of being genuine. Thus, the producers of counterfeits separate the status and quality aspects of the product and thereby allow some consumers to purchase the former aspect even though they would not be willing to pay the high price of purchasing the two aspects together (Grossman and Shapiro 1988). People, who cannot afford the consumption of a genuine good, could consume its counterfeits that are in much lower price. Counterfeits become the available substitutes for the genuine good due to their price advantage, and hence the consumptions of counterfeits are getting widespread proportions in developing countries. The illegitimate producers may impose an externality on the owners of branded genuine goods by reducing the snob appeal of their possessions (Higgins and Rubin 1986). However, it has been documented that the consumption of some counterfeits can promote the sale of genuine goods. In the event that genuine goods can be differentiated from their counterfeits, the sale of counterfeits can enhance the exclusivities of genuine goods and make the conspicuous goods more valuable, and because of the increase of snob appeal, more people consume the genuine goods (Lai and Chang 2012).

An individual enjoys his/her own consumption but also cares for the welfare of the people that are associated with him. Altruism can be defined as individual's behavior that intended to benefit another to gain himself a higher satisfaction, even when doing so may require the payment of an implicit price. The way parents care for their children is just an altruistic example. In addition to this biological relation, altruism also can influence when making consumption decisions. Via their

consumption choice, consumers can express their concern about whether a company deals business ethically. A significant subset of consumers is the ethical consumer who feels responsible toward society and expresses his/her or her altruism by means of his/her or her purchasing behavior. The purchase of a product is associated with consumers' altruism ethic if they concern a certain ethical issue (human rights, labor conditions, animal well-being, environment, etc.) as they make consumption decisions. Ethical consumer behavior is associated with the consumptions that intend to either benefit the natural environment (e.g., environmentally friendly products, legally logged wood, animal well-being) or help people (e.g., products free from child labor, Fair trade products). Some altruistic consumers will buy the products that have specific positive qualities (e.g., green products) to show their ethical concerns (Pelsmacker et al. 2005). For example, a considerable number of consumers have shown a willingness to pay a premium for products labeled as "Fair Trade" and a favorite to retailers that are seen to be more generous to their suppliers and employees, domestically and internationally. Marketing appeal to the consumer's altruism is the most important feature of Fair Trade products. The more an individual consumer's altruism, the higher the marginal altruistic value from giving more profit to the farmers. Raising awareness of Fair trade may promote individual's altruism and result in ethical consumption. Because of Fair trade, farmers can receive greater profits, which in turn induce farmers to invest more in producing their crops. Consumers' altruism may make it happen that a firm offers a Fair trade product in a competitive environment (Reinstein and Song 2012).

Not only can an individual's own tastes and habits be the determinants of his/her decisions on consumptions, but also social activities have effects on his/her demands for some goods. Consumers make their consumption choices under social influences, when they take up the attitude of others around them without being aware of they being doing so. Derived from a cause-consequence of social influence that is not technically measurable so far, the argument is much weaker in

the power of predicting consumers' choice than the one based on the factors that are supported by observables, they lead to some hypotheses and hardly to some theories. However, as the technology of data mining progresses and more resources have been invested in testing the unobservables, the theories of consumers' choices will be enriched and their predictions will be more powerful in the future.

References

- Becker GS (1991) A note on restaurant pricing and other examples of social influences on price. *J Polit Econ* 95:1109–1116
- Becker GS, Mulligan CB (1997) An endogenous determination of time preference. *Q J Econ* 112:729–758
- Becker GS, Murphy KM (1988) A theory of rational addiction. *J Polit Econ* 96:675–700
- Epstein GL (1987) The global stability of efficient intertemporal allocation. *Econometrica* LV:329–355
- Grossman GM, Shapiro C (1988) Counterfeit-product trade. *Am Econ Rev* 78:59–75
- Higgins RS, Rubin PH (1986) Counterfeit goods. *J Law Econ* 29:211–230
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47:263–292
- Lai FT, Chang SC (2012) Consumers' choices, infringements and market competition. *Eur J Law Econ* 34:77–103
- Laibson D (2001) A cue theory of consumption. *Q J Econ* 116:81–119
- Leibenstein H (1950) Bandwagon, snob, and veblen effects in the theory of consumers' demand. *Q J Econ* 64:183–207
- List JA (2003) Does market experience eliminate market anomalies? *Q J Econ* 118:41–71
- Lucas R, Stokey N (1984) Economic growth with many consumers. *J Econ Theory* XXXII:139–171
- Mas-Colell A, Whinston MD, Green (1995) *Microeconomic theory*. Oxford University Press, New York
- Pelsmacker PD, Driesen L, Rayp G (2005) Do consumers care about ethics? Willingness to pay for Fair trade coffee. *J Consum Aff* 39:363–383
- Pindyck RS, Rubinfeld DL (2012a) *Microeconomics*. Pearson, Taipei
- Pindyck RS, Rubinfeld DL (2012b) Difference Preferences toward Risk, 5.2 preferences Toward risk, Chapter 5, pp 159–161
- Reinstein D, Song J (2012) Efficient consumer altruism and Fair trade products. *J Econ Manag Strateg* 21:213–241
- Uzawa H (1968) Time preference, consumption function, and optimum asset holding. In: Wolfe JN (ed) *Value, capital, and growth: papers in honor of Sir John Hicks*. University of Edinburgh press, Edinburgh
- Varian H (1992) *Microeconomics*. Norton, New York

Contest Theory

► Tournament Theory

Contract, Freedom of

Péter Cserne
University of Hull, Hull, UK

Abstract

Freedom of contract is a principle of law, expressing three related ideas: parties should be free to choose their contracting partners (“party freedom”), to agree freely on the terms of their agreement (“term freedom”), and where agreements have been freely made, parties should be held to their bargains (“sanctity of contract”). This entry provides an overview of the economic justifications and limitations of this principle.

Freedom of Contract: Meaning and Significance

Freedom of contract is a fundamental principle of most modern contract laws, expressing three related ideas: parties should be free to choose their contracting partners (“party or partner freedom”), to agree freely on the terms of their agreement (“term freedom”) and where agreements have been freely made, parties should be held to their bargains and contracts should be enforceable by state institutions (“sanctity of contract”) (Brownsword 2006, p. 50). Freedom of contract prevails to “the extent to which the law sanctions the use of contracts as a commitment device,” leaving the terms of the contract agreement to the parties (Hermalin et al. 2007, p. 18).

Freedom of contract is an ideologically charged notion which attracts strongly held political views among both defenders and critics (Craswell 2000, p. 82). “Outside the legal academy, ‘freedom of contract’ largely serves as a

slogan for laissez-faire capitalism. Even within contract theory, the term retains a particular libertarian flavor” (Dagan and Heller 2013, p. 1).

Historical research has established that the idea of a general enforceability of agreements comes from late medieval and early modern theological and philosophical debates on the moral foundations of contract law. Ancient and medieval laws did not recognize the general enforceability of consensual agreements; only certain types of agreements based on consent were enforceable. Later, both the general principle of freedom of contract and its limits have been systematically discussed in late scholastic natural law theories, thus providing moral underpinning for the rise of economic freedoms (Gordley 1991; Decock 2013).

The philosophical discussion is still ongoing as to which exchanges are morally permissible, which are problematic, and why. Yet there seems to be a reasonably broad consensus that at least a limited version of freedom of contract may be supported by both autonomy-based and welfarist theories, and, perhaps less prominently but also importantly, by aretaic (virtue-based) arguments (Cserne 2012, pp. 82–89).

Legal doctrines of most modern legal system reflect this overlapping consensus to a considerable extent. Modern Western legal systems attach a high value to freedom of contract as a basic legal principle but also set several limits to this freedom, going well beyond punctual exceptions. In these countries, “most contracts that support legitimate economic exchange are at least presumptively enforceable. Still, the limits of freedom of contract vary among Western countries and are an important element of regulatory policy” (Hermalin et al. 2007, p. 19). In fact, today’s contract law regimes can be seen as long lists of exceptions to the principle of contractual freedom.

These exceptions have all been subject to analysis in mainstream law and economics scholarship: some make more (economic) sense than others. From an economic perspective, the presumption for freedom of contract is supported by its conduciveness to welfare and can be typically rebutted by identifying a bargaining failure or a market failure (Cooter and Ulen 2012, p. 341).

This, in turn, suggests either imposing mandatory terms on contracting parties or refusing the enforcement of their agreement. This functional linking of particular rules and doctrines to their incentive effects is a microlevel economic analysis of the limits of contractual freedom which provides valuable contribution to both economic and legal scholarship.

The Economic Case for Freedom of Contract

The operation of a modern market economy relies on freely negotiated enforceable contracts. Overall, in mainstream economic theory freedom of contract, sometimes under the label of consumer sovereignty (Persky 1993), has been traditionally supported by its likely benefits in terms of social welfare.

This case for freedom of contract is based on a contingent empirical generalization: “Most people look after their own interests better than anyone else would do for them” (Cooter and Ulen 2012, p. 342). In neoclassical economics, the “predilection for private ordering over collective decision-making is based on a simple (perhaps simple-minded) premise: if two parties are to be observed entering into a voluntary private exchange, the presumption must be that both feel the exchange is likely to make them better off, otherwise they would not have entered into it” (Trebilcock 1993, p. 7).

Freedom of contract and a competitive market economy seem to simultaneously promote individual autonomy and social welfare, converging toward what could be called the private ordering paradigm. According to this “convergence claim,” freedom of contract is supported by a combination of autonomy-based and welfare-based arguments (Pincione 2008). Mainstream economics claims that promises should be enforced when they provide an *ex ante* Pareto improvement, i.e., if and only if promisor and promisee both benefit from the agreement. This is often assumed to be equivalent to the assumption that both parties wanted the agreement to be enforceable when it was made. As we shall see below, however, this

convergence is not universal. In case of divergence, economists tend to give priority to social welfare considerations. “An economic case for or against freedom of contract is based on the consequent welfare implications” (Hermalin et al. 2007, p. 21).

In a perfectly competitive market, there should never be inefficient contract terms. Therefore, there is no way to improve efficiency by forbidding certain terms. This case is a useful benchmark in the sense that when one or more of the conditions of market perfection is not fulfilled, there is potential for improving efficiency by restricting freedom. In other words, circumstances when these conditions do not prevail provide an economic justification for rules and doctrines of contract regulation. The two main cases for regulating contract are then third-party effects and bargaining (contracting) failures.

Welfare economics provides theoretical justification for freedom of contract by reference to a general equilibrium economy (the First Theorem of Welfare Economics) or at least the efficiency of singular competitive markets (Hermalin et al. 2007, pp. 21–30). The Coase theorem suggests that freedom of contract is desirable even more generally. Costless contracting and the parties’ rationality guarantee that they will exhaust all possibilities for mutually beneficial exchanges, thus bringing about maximum welfare.

When we relax the zero transaction costs assumption of the Coase theorem, there might be situations with positive transaction costs that justify either default or mandatory rules. Yet the transaction costs of contract regulation need to be taken into account as well. “In other words, while it is true that restrictions on private contracts can possibly enhance efficiency when the private parties incur transactions costs, one must assess that observation in light of real-life limitations on what the legal system can do and the cost at which it can do it” (Hermalin et al. 2007, pp. 27–28).

In fact, contract law (and more generally, third-party enforcement) is one among many governance mechanisms for private transactions; its use varies historically and cross-culturally. The importance of law (courts and other state

institutions) in contract enforcement depends on its merits and costs relative to other governance mechanisms (Dixit 2004). On the one hand, contract law does not seem necessary for an exchange economy to operate. As Piccione and Rubinstein (2007) showed, many equilibrium features of competitive markets (an exchange economy) can be achieved even “in the jungle,” i.e., in an economy without property rights and freedom of contract where resource allocation is regulated by physical strength. On the other hand, there is ample evidence that legal enforcement of contracts is not sufficient for the welfare benefits of competitive markets to be realized. Well-functioning markets rely on social norms and informal institutions in various ways.

Economic analysis thus provides a *prima facie* justification for freedom of contract and also “suggests two potential grounds on which to argue against (complete) freedom of contract: (i) actors who are not party to a contract (third parties) are affected by externalities resulting from the contract; and (ii) problems in negotiating a contract prevent the parties from writing the optimal contract” (Hermalin et al. 2007, p. 30).

Constitutive, Procedural, Informational, and Substantive Limits to Freedom of Contract

Contract law is understood here as a body of legal rules that pertains to the enforcement and regulation of voluntary private agreements. Contractual freedom is not only a matter for contract law, however. It may be limited by rules outside the domain of contract law, for instance, when anti-discrimination laws constrain parties’ freedom to choose their contracting partners.

The distinction between state regulation and state enforcement (negative and affirmative government sanction) needs to be noted as well: “there are many agreements that cannot be enforced in the courts but that can still be useful as commitment devices if the parties can manage to implement them privately” (Hermalin et al. 2007, p. 19).

In what follows our main focus will be on freedom of contract and its limits as they appear in rules, principles, and doctrines of contract law. The rules and doctrines of contract law have been analyzed extensively in terms of their impact on social welfare. These limits are sometimes classified into constitutive, procedural, informational, and substantive limits to freedom of contract (Cserne 2012, pp. 93–135). With respect to each, the question for economic analysis is whether and how the legal instrument in question can be illuminated, explained, justified, or criticized in light of the empirical findings and normative criteria of economics.

Contracting practices and the private ordering paradigm implicitly assume some “constitutive limits” on freedom of contract (Kennedy 1982). This term refers to those minimal conditions of individual rationality and voluntariness which are necessary for the working of even a libertarian (unregulated) contract regime. Virtually all legal systems impose threshold conditions for the making of enforceable contracts, requiring capacity, and prohibiting duress and fraud. The key idea is that these “limits” constitute our idea of what contracts are, rather than constraining freedom of contract. These constitutive limits of freedom of contract not only guarantee contracting as a domain of individual autonomy but are also instrumental for increasing social welfare (Cserne 2012, pp. 93–106). While the importance of constitutive limits is quite intuitive, incorporating them into economic models is not straightforward. “Economists frequently extol the virtues of voluntary exchange, but economics does not have a detailed account of what it means for exchange to be voluntary” (Cooter and Ulen 2008, p. 12).

Procedural limits do not constrain the parties’ agreement on terms they choose (accept or bargain for), but merely require certain actions to be taken (or not taken) before contracting or during the contractual relationship. They relate to the process of agreeing on a contract and the manner of recording or authenticating the agreement. For instance, they prescribe written form, waiting period, mandatory advice, or mandatory withdrawal rights (cooling off periods).

Informational limits regulate the information flow between the parties before or during the contract. These include rules that mandate the precontractual furnishing of information and prohibit the provision of fraudulent, misleading, or irrelevant information.

Substantive limits set mandatory terms for the contracts either directly, e.g., by regulating interest rates and other terms of consumer credit contracts by statutory rules, or indirectly, e.g., by nonenforcement of terms that courts find unconscionable, unreasonable, or unfair.

Bargaining Failures and Market Failures

Cooter and Ulen's textbook treatment of regulatory doctrines of contract law (Cooter and Ulen 2012, Tables 9.3 and 9.5) can be seen as a systematic translation or linking exercise between various shortcomings of a perfectly functioning market on the one hand and the respective contract law doctrines (formation defenses and performance excuses) triggered or justified by a market failure on the other.

They focus on two kinds of shortcomings. The first include failures of the bargaining process that prevent welfare maximization (or make it unlikely) and are further classified as cases of *bounded rationality* (lack of stable and well-ordered preferences), addressed by the rules on (in)capacity, or cases of *constrained choice sets*, addressed by the doctrines of duress, necessity, or impossibility. When a contracting party faces dire (as opposed to moderate) scarcity and the other party either generates or takes advantage of this situation, the resulting contract is often not enforced. Constitutive limits to freedom of contract are always triggered in this context. Whether empirical evidence on bounded rationality justifies further limitations to freedom of contract is discussed below.

The second kind of shortcomings includes market failures that can be categorized according to three types of transaction costs which arise, respectively, from spillovers (externalities), information imperfections, and market power. These market failures are in turn addressed by various

contract law doctrines as well as noncontractual regulations (Cooter and Ulen 2012, pp. 341–372).

Externalities justify the unenforceability of contracts which derogate public policy or violate a statutory duty. This category will be analyzed further below.

Symmetrical or asymmetrical *information imperfections* also generate market failures (Trebilcock 1993, Chaps. 5 and 6). These are addressed by contract doctrines such as frustration of purpose or mutual mistake (in case of symmetric imperfections) and fraud, unilateral mistake, or failure to disclose (in case of information asymmetry). Information asymmetry has been the subject of economic analysis at least since Akerlof's (1970) seminal work, and the findings generally suggest that "such distortions must imply a loss of welfare vis-à-vis the symmetric-information benchmark. [...] Whenever the parties negotiate imperfect contracts, the question arises whether there is scope for the legal system to improve matters, either by restricting the set of possible contracts ex ante or through appropriate court action ex post" (Hermalin et al 2007, p. 34).

The third type of market failure includes structural or situational *monopoly* or at least significant market power to restrict competition. "Competitive markets can be expected to maximize welfare in the absence of externalities. When, however, one or more entities have market power, the market can no longer be expected to yield the social welfare-maximizing allocation" (Hermalin et al. 2007, p. 39). Market power is addressed primarily by competition law, but situational monopolies also trigger contract law doctrines such as necessity and unconscionability.

Externalities, Simple and Subtle, Justifying Intervention

Externalities impose costs or benefits from a particular exchange transaction on third parties who are not involved in the transaction. Positive externalities pose incentive problems, leading to a suboptimal quantity of the good or transaction in question. Negative externalities are arguably more important with respect to contract regulation.

If such an effect can be detected, this provides reason for interfering with contractual freedom.

The efficiency of markets and private contracting is contingent on there being no third-party externalities. For instance, the market equilibrium with a competitive, but heavily polluting, industry does not maximize welfare—the supply of the good in question is determined by the private costs incurred by the manufacturers rather than the social costs that account for both those private costs and the harm the pollution imposes on society. Because social costs are greater than private costs, more than the welfare-maximizing quantity gets sold. [...] More generally, in a market, bilateral contracts may generate externalities and reduce the welfare on an aggregate level. [...] The inefficiency of the market when externalities are present can justify restrictions on private contracts. (Hermalin et al. 2007, p. 30)

Although in principle externalities could be solved by bargaining toward a grand contract including all third parties, the number of these parties may be too big and some of them could be unknown or not yet exist. A bargaining solution would often generate insurmountable transaction costs (Hermalin et al. 2007, pp. 30–31).

Some limitations on freedom of contract can be easily and plausibly justified by externalities: “antitrust authorities may frown upon contracts that have potentially harmful effects on competition (most favored nation clauses, contracts that induce predatory or collusive behavior, etc.). Contracts between a firm and a creditor may exert externalities on other creditors, either directly through priority rules in the case of bankruptcy or indirectly through the induced change in managerial incentives. The Internal Revenue Service warily investigates employment contracts that might dissimulate real income” (Tirole 1992, p. 109).

Some other limits to freedom of contract can be seen as responding to specific forms of harmful externalities. For instance, in his *Principles of Political Economy*, John Stuart Mill referred to the statutory limitation of working hours as an example of what we would call now governmental solutions to a collective action problem: “classes of persons may need the assistance of law, to give effect to their deliberate collective opinion of their own interest, by affording to every individual a guarantee that his competitors will pursue the

same course, without which he cannot safely adopt it himself” (Mill 1848, Book V chapter XI § 12). He argued that even if workers as a class would prefer to work for shorter periods, they cannot achieve it without mandatory rules limiting working hours, because each would have an individual interest working longer. While a full-fledged economic analysis of such problems is rather complex, even in a competitive market, there may be cases when such interference is Pareto improving (Basu 2007). Kaushik Basu argued that in this category of cases, overriding the principle of freedom of contract can be justified within a welfarist framework, without reference to paternalism, moralism, or even autonomy.

Similarly, in a thoughtful article, Eric Posner suggested that many protective laws of modern welfare states serve to redress imbalances created by social security and welfare laws (Posner 1995). By providing a social safety net, welfare states effectively truncate the downside of financial and other risks to citizens. This regulatory environment of a welfare state has the unintended effect of encouraging socially harmful behavior, such as irresponsible spending, risky borrowing, and overindebtedness. This suggests that many seemingly paternalistic limitations on freedom of contract may be justified by harmful externalities. When individuals take on too much risk in reliance on the welfare state, they impose external costs on society. Thus, what at first looks like a rule protecting vulnerable groups may in fact be protecting the public budget.

Virtues and Vices of Reductionist Accounts

While the above cases may be plausibly analyzed in terms of externalities, the regulatory relevance of externalities is bound with problems. The issue is both theoretical and practical: what kind of externalities should contract regulation take into account? Similar to autonomy-based theories which face difficulties delimiting relevant harms that would justify coercion, welfare-based theories have difficulties in delimiting the kinds of third-party effects that would justify welfare-enhancing intervention.

First, third-party effects are pervasive: “virtually any contract may cause some external harm, [at least by] denying other potential contracting parties the opportunity to contract with the parties to the contract in question” (Shavell 2004, p. 320). If all external effects are taken into account, the private ordering paradigm is largely at an end (Trebilcock 1993, p. 58).

To be sure, from a welfarist perspective, the mere presence of an externality is not sufficient to justify limitations. “The harm to third parties must tend to exceed the benefits of a contract to the parties themselves for it to be socially desirable not to enforce a contract” (Shavell 2004, p. 320). But as externalities are ubiquitous, regulators need additional criteria to determine what kind of externalities matter and how much weight should be attached to them. Welfare-based analyses alone do not provide tools for selecting between relevant and irrelevant externalities.

For instance, so-called moral externalities refer to the fact that some people find certain conducts morally offensive or simply disgusting. Should these effects matter for contract regulation? Mainstream economics is unlikely to provide help in this matter because without further analytical tools, economics cannot properly distinguish other-regarding preferences and tangible externalities (Hatzis 2006).

Second, even if externalities are clearly tangible or measurable in monetary terms (such as costs on dependents, the social welfare system, or the public health care system), it is contestable whether such externalities provide sufficient reason for regulatory intervention. For instance, unhealthy lifestyles or risky leisure activities may have an impact on the public budget, but it is not clear whether the freedom to purchase unhealthy food should be limited for this reason alone (Trebilcock 1993, p. 75).

More generally, one may question whether all reasons for contract regulation can be fruitfully analyzed in terms of individual and social welfare.

In its simplest versions, economic arguments classify limitations to freedom of contract according to whose welfare is increased (whose losses are prevented) by nonenforcement. Steven Shavell distinguishes two rationales for legislative or judicial overriding of contracts: the existence of

harmful externalities and welfare losses to the contracting parties themselves. More interestingly, he claims that this exhausts the set of valid reasons. Other justifications for nonenforcement are merely stands-in for the previous ones. Inalienability and paternalism ultimately protect, maybe in subtle or complex ways, the welfare of either the contracting parties or third parties. Consequently these rationales are reducible to the two previous ones (Shavell 2004, p. 322).

This argument is fully compatible with the methodological and substantive assumptions of mainstream economics: welfare maximization and consumer sovereignty. Having identified negatively affected third parties, specific transaction costs, and/or informational imperfections, instances of contract regulation are considered economically justified by their social welfare benefits, i.e., to the extent that they remedy such market failures. The plausibility and success of this reductionist account, however, deserves further analysis.

Some economists argue that concerns such as commodification or inalienability cannot be easily translated into welfare terms; at most, they can be modeled as specific preferences (Hermalin et al. 2007, pp. 47–48). Paternalism is also not easily translated into economic terms (Buckley 2005; Cserne 2012, Chap. 3).

Arguably, it is both an advantage and a problem for the reductionist approach that it surpasses the conflict between welfare and autonomy inherent to certain kinds of contract regulation. The attraction of such a reductionist approach comes from its simplicity or even elegance. Not only the case for freedom of contract but all of its justified limitations can be explained in relatively narrow terms by neither relaxing the rationality assumptions nor resorting to fairness arguments.

The danger is, however, that if we consider this issue from a purely welfarist economic perspective, there is no principled limit to paternalism. Indeed, as far as their normative views are concerned, economists can be anywhere on the range between hard paternalism and hard anti-paternalism. Relying on an ad hoc mixture of welfarist and autonomy-based principles, economists are ill-equipped to handle problems of

freedom of contract which arise precisely from the conflict of these two principles. If economics acknowledges some exceptions or limits to the private ordering paradigm, as it usually does in practice, then in order to justify these exceptions, “some theory of paternalism is required, the contours of which are not readily suggested by the private ordering paradigm itself” (Trebilcock 1993, p. 21).

Pragmatic Arguments and Institutional Design

Within a reductionist economic framework, limits to allegedly welfare-maximizing interventions can be added in a contingent way, with reference to empirical facts about the functioning of the institutional mechanisms that are supposed to be used for carrying out the intervention. These contingent empirical circumstances provide pragmatic arguments for freedom of contract (Cserne 2012, pp. 31–33).

Pragmatic anti-interventionist arguments draw attention to the side effects and non-intended, often counterintentional, consequences of contract regulation. These arguments are not specific to freedom of contract but need to be taken into account in designing regulatory policies. Nor do they categorically support freedom of contract. In other words, freedom of contract may be justified *faute de mieux*, by the costs of possible interventions. More specifically, pragmatic arguments refer to (1) the overinclusiveness of rules, (2) ensuing redistributive effects, (3) the lack of information, or (4) inadequate motivations of the regulators.

The pragmatic question asked here is whether the state is more or less able to prevent undesirable contracts than other mechanisms. It is worth noting that sometimes there are very few resources needed for effective state intervention: “as long as the courts are needed to enforce contracts, the contracts will not be made, and the state does not need to police the actual making of contracts and root out the undesirable ones” (Shavell 2004, p. 322).

This leads us further into questions of institutional design. Some market failures cannot be appropriately addressed judicially, i.e., through

private law constraints on freedom of contract but there may be other regulatory tools available. Ultimately, economics is likely to suggest a mix of policy instruments for contract regulation. For instance, Trebilcock argues for a “relative institutional division of labor” in which “the common law of contracts will be principally concerned with autonomy issues in evaluating claims of coercion, antitrust and regulatory law [with] issues of consumer welfare, and the social welfare system [with] issues of distributive justice” (Trebilcock 1993, p. 101).

Behavioral Economics and Freedom of Contract

At first sight, behavioral economics seems the right way to go, not only with regard to contractual freedom but policymaking more generally. By testing the assumptions of economics empirically, making economic theories descriptively more precise, one can expect to make economics more useful for policy design. Behavioral economics has identified and/or provided evidence for the functioning of various techniques of choice architecture (sticky default rules, options, menus, information provision) which take into account empirical findings on human decision making and thus can be put to socially beneficial use in contract regulation.

What else has behavioral economics to tell about freedom of contract? It is sometimes argued that psychological research provides new arguments for limiting freedom of contract. This idea seems misconceived (Cserne 2012, pp. 43–54, 137–139). There are distinct economic arguments for limiting freedom of contract in certain circumstances, and empirical research is indispensable for identifying whether and to what extent these circumstances prevail. Also, empirical findings give more detail and in this respect further support to the existing argument about the imperfection of the correlation between individual choice and welfare maximization. The increased attention to these findings is expected to lead to more precise and better-founded knowledge about the circumstances when contracting parties make

suboptimal choices. Yet, the evidence on various cognitive biases does not, in itself, call for limiting freedom of contract any more or less than common sense observations about human frailties. Indeed, it is an open question whether behavioral findings justify more or less paternalistic regulation than we currently can observe. (Note that the relevant comparison is this, not the one between hard paternalism and a hypothetical libertarian “regulatory” regime.) Some empirical research draws attention to psychological advantages of giving people the freedom to choose (Feldman 2002). More importantly, the normative standards for contract regulation need to come from elsewhere than empirical research. In this regard, behavioral economics does not fare any better or worse than mainstream economics.

Cross-References

- [Liberty](#)
- [Limits of Contracts](#)
- [Market Failure: Analysis](#)
- [Market Failure: History](#)
- [Public Goods](#)
- [Rationality](#)

References

- Akerlof GA (1970) The market for lemons: quality uncertainty and the market mechanism. *Quarterly J of Economics* 84:488–500
- Basu K (2007) Coercion, contract and the limits of the market. *Soc Choice Welfare* 29:559–579
- Brownsword R (2006) Freedom of contract. In: Brownsword R (ed) *Contract law. Themes for the twenty-first century*, 2nd edn. Oxford University Press, Oxford, pp 46–70
- Buckley FH (2005) *Just exchange: a theory of contract*. Routledge, London
- Cooter R, Ulen T (2008) *Law and economics*, 5th edn. Pearson Education, Boston
- Cooter R, Ulen T (2012) *Law and economics*, 6th edn. Pearson Education, Boston
- Craswell R (2000) Freedom of contract. In: Posner E - (ed) *Chicago lectures in law and economics*. Foundation Press, New York, pp 81–103
- Cserne P (2012) *Freedom of contract and paternalism. Prospects and limits of an economic approach*. Palgrave, New York
- Dagan H, Heller MA (2013) *Freedom of contracts*. Columbia law and economics working paper no. 458. Available at SSRN. <http://ssrn.com/abstract=2325254>
- Decock W (2013) *Theologians and contract law. The moral transformation of the Ius Commune (ca. 1500–1650)*. Martinus Nijhoff, Leiden
- Dixit A (2004) *Lawlessness and economics. Alternative modes of governance*. Princeton University Press, Princeton
- Feldman Y (2002) Control or security: a therapeutic approach to the freedom of contract. *Touro L Rev* 18:503–562
- Gordley J (1991) *The philosophical origins of modern contract doctrine*. Clarendon, Oxford
- Hatzis AN (2006) The negative externalities of immorality: the case of same-sex marriage. *Skepsis* 17:52–65
- Hermalin BE, Katz AW, Craswell R (2007) *Contract Law*. In: Polinsky AM, Shavell S (eds) *The handbook of law & economics*, vol 1. Elsevier, Amsterdam, pp 3–136
- Kennedy D (1982) Distributive and paternalist motives in contract and tort law, with special reference to compulsory terms and unequal bargaining power. *Maryland Law Rev* 41:563–658
- Mill JS (1848) *Principles of political economy with some of their applications to social philosophy*. <http://www.econlib.org/library/Mill/mlP.html>
- Persky J (1993) Consumer sovereignty. *J Econom Perspect* 7:183–191
- Piccione M, Rubinstein A (2007) Equilibrium in the jungle. *Econom J* 117:883–896
- Pincione G (2008) Welfare, autonomy, and contractual freedom. In: White MD (ed) *Theoretical foundations of law and economics*. Cambridge University Press, Cambridge, pp 214–233
- Posner EA (1995) Contract law in the welfare state: a defense of the unconscionability doctrine, usury laws, and related limitations on freedom of contract. *J Legal Stud* 24:283–319
- Shavell S (2004) *The foundations of economic analysis of law*. Harvard University Press, Cambridge
- Tirole J (1992) Comments. In: Werin L, Wijkander H (eds) *Contract economics*. Basil Blackwell, Oxford, pp 109–113
- Trebilcock MJ (1993) *The limits of freedom of contract*. Harvard University Press, Cambridge

Further Reading

- Atiyah PS (1979) *The rise and fall of freedom of contract*. Clarendon, Oxford
- Ben-Shahar O (ed) (2004) *Symposium on freedom from contract*. *Wisconsin Law Rev* 2004:261–836
- Buckley FH (ed) (1999) *The fall and rise of freedom of contract*. Duke University Press, Durham
- Craswell R (2001) Two economic theories of enforcing promises. In: Benson P (ed) *The theory of contract law: new essays*. Cambridge University Press, Cambridge, pp 19–44
- Kerber W, Vanberg V (2001) *Constitutional aspects of party autonomy and its limits: the perspective of*

constitutional economics. In: Grundmann S, Kerber W, Weatherhill S (eds) *Party autonomy and the role of information in the internal market*. De Gruyter, Berlin, pp 49–79

Kronman AT, Posner RA (1979) *The economics of contract law*. Little & Brown, Boston

Schwartz A, Scott RE (2003) *Contract theory and the limit of contract law*. Yale Law J 113:541–619

Contracts of Adhesion

Elena D'Agostino

Department of Economics, University of Messina, Messina, Italy

Abstract

In the globalized mass production economy with a large number of individual consumers, transactions very often take place between parties who are not physically present, such that communication between them turns out impossible or, at least, highly expensive. For that reason contracts are usually proposed by sellers to consumers on a take-it-or-leave-it basis without negotiation, referred to as contracts of adhesion. Consumers usually do not read the whole contract, and sellers can include inefficient one-sided clauses in fine print. This work reviews the main legal and economic literature on this topic presenting the traditional reasons to justify regulation in favor of consumers and highlighting its risks.

Introduction

According to a well-known definition, a contract is “a meeting of minds” between two or more parties who bargain on its terms and conditions. Moving from the legal definition, contracts are also viewed as a fundamental and irreplaceable tool in economics to realize an efficient allocation of limited resources among agents.

It is unanimously recognized that what should characterize an effective contract is the balance between parties' power in order to avoid that any

of them may exploit some bargaining power against the other. On the contrary, a contract should be the natural conclusion of a bargaining process in which parties do not fight each other but have to reach a compromise between their opposite interests after a (more or less long and, Coasian speaking, expensive) discussion about what terms to be bound to.

If it is true that contracts are enforceable when all parties knowingly consent, nevertheless a knowledgeable consent is sufficient but not necessary to make the contract enforceable. Indeed in the globalized mass production economy with a large number of individual consumers, transactions very often take place between parties who are not physically present, such that communication between them turns out impossible or, at least, highly expensive. As a consequence, it is surprising that most of the contracts we sign (or simply accept in words) every day do not come out from a bargaining process, but every term is proposed by one party to the other, and the latter simply limits her will to adhere to the preprinted content: for this reason lawyers refer to this category of contracts as *contracts of adhesion*. Furthermore, when the same content is reproduced in every contract for the same good and proposed to any potential customer on a take-it-or-leave-it basis, the contract is not simply adhesive but also standard.

Standard contracts of adhesion characterize several markets (think of transportation, bank contracts, insurance policies, and the huge and endless list of online contracts) and are not necessarily bad contracts but rather make transactions quicker in so far as they are able to economize on some costs, like writing down the contract, finding legal references for each clause, and so on. The downside, however, is that the drafter party could exploit his bargaining power to insert one-sided clauses, that is, clauses whose content turns out very onerous for the other party and very generous for himself. To put these clauses away from the counterparty's eyes, the drafter usually relegates them at the bottom of the contract, in some annexes or footnotes, and writes them in fine print using very unfriendly legal terms to make the content

intentionally unavailable or at least obscure for nonprofessional readers.

It turns out that whether the nondrafter party reads such terms is not an open question, especially when she is a final consumer not professionally involved into the transaction: in respect to consumers, the answer to the question is quite simple, and consumers usually do not read standard contracts and, in particular, do not read fine print, so their signatures do not necessarily imply that their consent has been knowledgeable (see Bakos et al. 2014 on infrequent reading).

The reason why consumers do not read is not necessarily found in their incapacity to realize the presence of these clauses and the risk involved into signing without reading, but it could be the rational decision of not investing time and resources into an activity that may turn out useless either because it is too costly or because the consumer needs the good and knows that, even if she does not totally agree with every clause, they are unalterable and the only alternative is to reject the whole contract.

Suppose, per contra, the consumer decides to read every term. Given the obscure language used to write some or all these clauses, we have already noted that it is very unlikely that she will be able to understand their content, especially if she is not an expert in legal terms. For this reason the literature stresses on the fact that reading implies a cost on the side of consumers. It does not mean that consumers sign without having any idea of contract terms; rather it is likely that they limit their attention to some clauses only (above all those fixing price), for this reason defined as salient, and skip some other (like those regulating insurance, restoration in case of damages, place of jurisdiction, etc.), accordingly defined as non-salient.

Contracts of Adhesion in the Law Literature

The law literature contains at least two different approaches, both aiming to justify regulation of complex contracts with fine print to protect potentially unaware buyers.

The first approach is based on market structure. In this sense, Kessler (1943) argues that monopolists exploit their market power by offering contracts containing onerous terms with buyers not able to understand their content and/or to renegotiate some or all terms. While Kessler does not discuss regulation, his argument suggests that courts should be more prone to strike down standard-form contracts drafted by a monopolist. More recent versions of this approach include Kornhauser (1976), who claims that oligopolists would agree to draft onerous terms to facilitate price-fixing, and Shapiro (1995), who argues that competition would protect buyers from exploitation. Some courts have used market structure as a criterion for treating terms as procedurally unconscionable (and therefore unenforceable), notably in *Henningsen v Bloomfield Motors* (NJ 1960) and more recently in *Pack v. Damon Corp* (E.D. Mich 2004) and *Flores v. Transamerica HomeFirst Inc.* (Cal. Ct. App. 2001). This view corresponds to the traditional economic thought based on the supposition that a free competitive market provides efficient clauses in equilibrium: an argument that collapses if consumers have to pay a cost to read and to understand contract terms as efficiency in a competitive market requires that they are fully informed and rational.

Kessler's argument has been discredited on both empirical and theoretical grounds. Theoretically speaking, competitive firms also offer non-negotiable, complex contracts, whose terms are usually not less onerous. Empirically speaking, there are studies focusing on specific markets which demonstrate that the severity of terms included in standard-form contracts does not depend on market structure (see Marotta-Wurgler 2008 for the market for software licenses, and Priest 1981 specifically for warranty terms included in standard-form contracts). Moreover, according to a conventional argument, monopolists are better served by raising price than by including onerous terms (for a detailed analysis, see Rakoff 1983).

On an alternative account, regulation is considered a necessary step when contracts are pre-printed by one of the two parties because the other party, usually consumers, should be protected as

they lack the expertise to understand and/or could be too naive to regard unread terms skeptically. A recent behavioral literature treats the infrequency of reading as indicative of consumer naivety and, in contrast to the Kessler tradition, argues that competitive sellers lack an incentive to educate such buyers: cf. Gabaix and Laibson (2006) and Gilo and Porat (2011). Regulations may therefore benefit such naive buyers, irrespective of market structure (see ► [Naïve Consumers: Contract Economics](#)).

Contracts of Adhesion in the Economic Literature

Various papers incorporate a cost of reading terms that are drafted by one of the parties to the contract. Katz (1990) analyzes a monopoly in which the only seller can choose non-price terms from an interval that depends on the legal regime under consideration. A monopolist chooses price that consumers observe at no cost and also whether to disclose its chosen non-price terms at some cost. If the seller does not disclose, then each buyer can either accept or read, also at some cost. In the unique equilibrium outcome, the monopolist discloses whenever this is cheap enough; otherwise, she offers the most onerous terms that the regime allows (conditional on not disclosing): the monopolist optimally sets terms such that a reading buyer never strictly prefers to accept. Katz considers a family of regulations which includes mandating favorable terms and shows that mandating the same terms as a disclosing seller would choose induces an efficient outcome. This regulation solves a commitment problem in Katz because a monopolist who did not disclose would offer the most onerous terms that the regime allows.

D'Agostino and Seidmann (2016) analyze a monopoly in which the only seller offers a menu of either complex or simple nonnegotiable contracts: the latter contains a (transparent) price and default non-price terms alone, while the former also contains a shrouded non-price term which is either favorable, default, or onerous (for all buyers). Consumers incur a (sunk) cost if they

read the shrouded terms in a complex contract; and trade on favorable terms is socially efficient. If some buyers are sophisticated, then in equilibrium trade must be inefficient because of a commitment problem which is intrinsic to fine print. On the one hand, sellers cannot credibly promise that the terms in unread complex contracts are favorable and can only be disciplined to (sometimes) draft such terms if buyers read. Without such discipline, the monopolist would offer onerous contracts, which (sophisticated) consumers would then reject. On the other hand, sophisticated consumers will only incur the costs of reading if the seller randomizes over terms, and the utility difference between the best and the worst terms justifies the cost. Equilibrium play is always inefficient because trade sometimes occurs in contracts that do not include favorable terms, and sophisticated consumers then engage in socially wasteful reading.

The authors consider the effects of two sorts of regulation of non-price terms. They first compare the effects of a regulation that mandates favorable terms (that also correspond by assumption to efficient terms). Given the model they present, such regulation resolves the commitment problem, but the identification of the beneficiaries differs from Kessler's. Precisely the authors prove that resolution of the commitment problem helps the stronger side of the market in a monopoly. Thus, a regulated monopolist gains because she can now extract consumers' maximal surplus, while her sophisticated consumers are unaffected. In sum, the distributional effects of this regulation depend on market structure but in the opposite direction to Kessler's suggestion. The authors also consider the effects of a regulation that prohibits onerous terms alone. This mitigates, but does not eliminate the commitment problem; and, as a result, regulated trade is also inefficient. Indeed, such regulation may lower welfare by reducing the utility difference between the best and the worst terms in complex contracts, which drives any complex contracts out of the market by deterring consumers from reading, and thereby disciplining the seller.

Turning to a competitive market, in Che and Choi (2009) competitive sellers can decide

whether to disclose (at some cost) or not to disclose their terms that could be either favorable or unfavorable. Consumers hold heterogeneous preferences over non-price terms and can either accept or read (at some cost) the contract, possibly resampling another seller if they reject on a first instance. If sellers do not disclose, whenever consumers read in equilibrium, they are also indifferent between accepting and reading, and use price as a signal for the quality of contract terms. If sellers disclose, then consumers can immediately have access to contract terms without paying any cost. Che and Choi (2009) compare play in legal regimes with a duty to read and with a duty to speak: the former regime corresponds to an unregulated market, and sellers separate into those offering favorable terms with high probability and not disclosing and those who are sure to offer onerous terms and disclose; sellers must disclose in the latter regime but can still offer onerous contracts. Che and Choi find that the two regimes cannot be welfare ranked. However, the duty to read regime welfare dominates when, *ceteris paribus*, reading is cheap enough, whereas buyers are better off in the duty to speak regime if enough of them care about non-price terms. The latter regime does not result in efficient trade because sellers charge a single price to consumers with heterogeneous valuations and because of the cost of speaking.

Regulations

As pointed out in previous sections, a large consensus has grown up in the last decades among lawyers and economists that consumers must be protected against evidently unfair, non-negotiated terms. The key question in examining alternative legal systems is how protective these measures should be. Comparing different legal systems, such as the US and the EU systems, it turns out that the former opted for a regulation based on clause disclosure, whereas the latter preferred a regulation of contract terms.

What clearly emerges from the US system is, however, the lack of an organic regulation of contracts of adhesion and rather a proliferation

of acts or pieces of laws applying to specific markets and/or to specific transactions. Examples can be found in the Truth in Lending Act of 1968 for the credit market, in the Magnuson-Moss Warranty Act of 1975 applying to the market for warranties, in the Nutritional Labeling and Education Act of 1994 about the food market, and, more recently, in the Principles of the Law of Software Contracts of 2009 approved by the American Law Institute. All these laws have limited application to specific markets, but looking at their rules, there is a common feature: all of them emphasize the importance of consumer's awareness of terms and conditions, and regulation mainly consists of mandating disclosure of clauses written in fine print but leaving the drafter free to decide the content of these clauses.

This approach aims to protect the main principle in contract law, well known as freedom of contract. However, as argued by Korobkin (2003), such an intervention will sort out a positive effect if consumers are not sophisticated, that is, they are not able to understand the risk involved in signing the contract without reading every clause (see again entry on ► [Naïve Consumers: Contract Economics](#)). There is empirical evidence against this argument: Marotta-Wurgler (2012) shows that making it available for the consumer to access terms and conditions in online contracts for software licenses (as measured by the number of clicks required to access the corresponding window) has a negligible impact on the rate at which consumers actually read them. On the same line, Ben-Shahar and Schneider (2011) argue that to be effective, disclosure should be brief, easy, and simple, but how brief, easy, and simple it should be is still an open question.

Moving to the EU system, the 93/13/EEC Directive is the main piece of law regulating the usage of contracts of adhesion with standard clauses. On a first look, we can find some rules following the same spirit of US regulation to the extent that the aim seems to disclose at least the existence, if not the content, of some risky clauses: e.g., it is required that the consumer declares to have read terms and conditions by putting an additional signature (or by ticking a box if the transaction is online). On a second and

deeper reading, however, the approach of the EU Directive looks like pretty different from the US system. First of all, the Directive has a general application to every contract of adhesion and not just to specific markets. Secondly its main concern is evidently to look directly at the content of fine print in order to avoid their use when it turns out so one-sided to become vexatious. The Directive also contains a black list of clauses that are presumed to be vexatious and, for this reason, never enforceable even if included in the contract.

Despite these differences between the two systems, it must be reminded that courts in both systems (and also national courts for European countries) have played an important role in identifying high-risk cases and have also used the market structure criterion in order to regulate those markets where sellers exploit some market power.

Cross-References

► [Naïve Consumers: Contract Economics](#)

References

- Bakos Y, Marotta-Wurgler F, Trossen D (2014) Does anyone read the fine print? *J Leg Stud* 43:1–36
- Ben-Shahar O, Schneider CE (2011) The failure of mandated disclosure. *Univ Pennsylvania Law Rev* 159:647–749
- Che Y-K, Choi A (2009) Shrink-wraps: who should bear the cost of communicating mass-market contract terms? mimeo
- D'Agostino E (2005) Contracts of adhesion between Law and Economics. Rethinking the unconscionability doctrine. Springer
- D'Agostino E, Seidmann DJ (2016) Protecting buyers from fine print. *Eur Econ Rev* 89:42–54
- Gabaix X, Laibson D (2006) Shrouded attributes, consumer myopia and information suppression in competitive markets. *Q J Econ* 121:505–540
- Gilo D, Porat A (2011) Viewing unconscionability through a market lens. *William Mary Law Rev* 52:133–195
- Katz A (1990) Your terms or mine? The duty to read the fine print in contracts. *RAND J Econ* 21:518–537
- Kessler F (1943) Contracts of adhesion – some thoughts about freedom of contract. *Columbia Law Rev* 43:629–642
- Kornhauser L (1976) Unconscionability in standard forms. *Calif Law Rev* 64:1151–1183
- Korobkin R (2003) Bounded rationality, standard form contracts, and unconscionability. *Univ Chicago Law Rev* 70:1203–1295
- Marotta-Wurgler F (2008) Competition and the quality of standard form contracts. *J Empir Leg Stud* 5:447–475
- Marotta-Wurgler F (2012) Does contract disclosure matter? *J Inst Theor Econ* 168:94–119
- Priest G (1981) A theory of the consumer product warranty. *Yale Law J* 90:1297–1352
- Rakoff T (1983) Contracts of adhesion: an essay in reconstruction. *Harv Law Rev* 96:1173–1284
- Shapiro C (1995) Aftermarkets and consumer welfare: making sense of Kodak. *Antitrust Law J* 63:483–511

Contracts, Forward

Haksoo Ko

School of Law, Seoul National University, Seoul, Republic of Korea

Abstract

A forward contract is an agreement to buy or sell an asset at a specified future time at a pre-specified price. While financial underpinnings of forward contracts are well-known, law and economics research in forward contracts is underdeveloped.

Definition

A forward contract is an agreement to buy or sell an asset at a specified future time at a pre-specified price. A forward contract can be contrasted to a spot contract, which is an agreement to buy or sell an asset almost immediately. A forward contract is similar to a future contract. A main difference is that, whereas future contracts are typically standardized and trade on exchanges, forward contracts are nonstandardized and trade on the over-the-counter market.

Contract, Forward

A forward contract is an agreement to buy or sell an asset at a specified future time for

a pre-specified price (Hull 2014; Bodie et al. 2010). A party to a forward contract assumes a “long position,” and the other party assumes a “short position.” The party with a long position agrees to buy the underlying asset on the specified future date for a pre-specified price, while the party with a short position agrees to sell the asset on the same date for the same price.

A popular type of forward contracts is for foreign exchanges. As an illustration, imagine a European company which must purchase materials from Chinese suppliers and pay in Chinese Yuan on a regular basis. The European company would then be subject to cost uncertainty due to the unpredictable nature of the future exchange rate between Euro and Yuan. This company can reduce the uncertainty by entering into a forward contract with a bank, fixing the exchange rate at maturity in advance, regardless of the then prevailing exchange rate. Forward contracts can be entered into for all sorts of assets, whether financial or physical.

Parties enter into a forward contract for a variety of reasons. A common reason would be parties’ different expectations about the future, e.g., as to whether a foreign exchange rate would rise or fall. In such a case, a forward contract would simply reflect the parties’ different views or different evaluations on future circumstances in the foreign exchange market. Since the value of a forward contract at maturity will change depending on the spot price which will then be prevailing, the parties’ agreement in this context can be viewed as their bet on the spot price in the future. Another reason why parties enter into a forward contract would be to hedge against fluctuations of the value of the asset that is subject to the contract. For a party, entering into a forward contract would be like buying insurance, while it would be like selling insurance for the other party. That way, the parties could reallocate and share risks.

A forward contract would show a distributional effect, largely reflecting the difference between what the parties expected at the time of entering into the contract and the actual outcome at maturity. A forward contract could also have an effect on allocation if, e.g., a party is better

positioned to pool a similar type of risks together. Thus, for instance, between a bank and a manufacturing company exposed to risks arising from foreign exchange rate fluctuations, it would typically be the bank which provides the function of pooling risks.

The structure of a forward contract is often simple and straightforward. As such, so long as there are no disputes as to the validity of the forward contract and also as to the applicable contract terms, contract obligations would become clear at maturity. Thus, if there are disputes between the parties, it may well be regarding the inherent validity of the contract, rather than regarding the interpretation of specific contract terms.

In that context, a forward contract could prove to be problematic, if a party has sophisticated expert knowledge about the relevant market, while the other party lacks such knowledge. In such a case, depending on the market situation at maturity, the party claiming the lack of knowledge may refuse to carry out contract obligations and may challenge the validity of the contract. In challenging the validity of the contract, this party may claim that the forward contract is a fundamentally unfair contract and allege violation of the general legal principle requiring good faith when entering into a contract. Fraud, mistake, failure to explain, and various other legal doctrines could also be cited in challenging the validity of the forward contract.

However, without the detailed factual information surrounding the parties’ dealings at the time of entering into the forward contract, it would be difficult to assess whether the party’s claim challenging the validity of the forward contract is itself made bona fides or whether the party is engaging in ex post opportunistic behavior.

In assessing the parties’ judgment when entering into a forward contract, a behavioral law and economics perspective could be helpful. Behavioral law and economics could shed light on the parties’ motivations and behavior at the time of entering into contract and could help in learning if psychological biases and limitations such as investor myopia, optimistic bias, and herd mentality had impact on one or both parties (Ko and

Moon 2012). In particular, if a contracting party lacks sophistication, enhancing understanding as to how individual contract provisions were inserted into a specific forward contract would help in assessing whether the forward contract could be viewed as a result of arm's-length dealings. In entering into a forward contract, the parties often use a standard form contract, and as such, the law and economics literature on standard form contracts could be helpful as well.

Overall, law and economics research in forward contracts is underdeveloped. As we learn more about what precisely transpires when the parties enter into a forward contract, academic research in this area will become richer and more interesting.

References

- Bodie Z, Kane A, Marcus A (2010) *Investments*, 9th edn. McGraw-Hill, New York
- Hull JC (2014) *Options, futures, and other derivatives*, 9th edn. Prentice Hall, New Jersey
- Ko H, Moon W (2012) Contracting foreign exchange rate risks: a behavioral law and economics perspective on KIKO forward contracts. *Eur J Law Econ* 34:391–412

Cooperative Game and the Law

Samuel Ferey
CNRS, BETA, University of Lorraine,
Nancy, France

Definition

While noncooperative game theory applied to the law is now a subfield of law and economics literature, cooperative game theory has a more strange history. Many of the founding fathers of the cooperative game theory (Shapley, Shubik, Owen, Aumann) were interested in legal examples to illustrate their games; however, law and economics literature has not systematically investigated the meaning of cooperative game theory for the

law and is still mostly noncooperative oriented. The aim of the entry is to draw a general picture of what cooperative game theory may add to the law and economics literature. We focus on the positive and normative aspects of cooperative game theory, and we provide illustrative examples in different fields (private law, public law, regulation, theory of the law).

Cooperative Game Theory and the Law: A Missed *Rendezvous*?

In their famous book *Game Theory and the Law* published in 1994, Baird, Gertner, and Picker said nothing about cooperative games and the law (Baird et al. 1994). The authors mainly focused on noncooperative games and gave to the Nash equilibrium the most preeminent role to understand strategies of legal players (plaintiffs, defendants, and judges). Most of the subfields of law and economics have been deeply changed by the use of game theory in place of the Chicago-style price theory. Compared with noncooperative game theory, cooperative game theory is still underestimated. From a historical perspective, here lies a paradox. While most of the classical games studied by cooperative branch of game theory have obvious consequences for the law (the ownership game by Shapley and Shubik 1967; the bankruptcy game by Aumann and Maschler 1985; the airport games by Littlechild and Thomson 1977), no structured paradigm on law/economics/cooperative games has emerged compared to the noncooperative game approach. The first reason lies in the fact that the legal examples modeled with cooperative game theory were not explicitly oriented in a law and economics perspective. The second reason is that, at the very beginning of the law and economics movement, leaders of the field, like Posner or Calabresi, were almost exclusively interested in the efficiency of the common law (maximization of wealth) which is not an issue addressed by cooperative game theorists.

Much has changed in recent times. A lot of new and original models focus on unprecedented applications that can be envisaged through cooperative game theory as applied to private law, public law, regulation, and legal theory. These

legal-oriented models may overlap applications in public economics or industrial organization (imperfect competition, public goods, matching, and networks). We have chosen to focus on some of the most suggestive applications of cooperative game theory for the law.

As we would like to avoid non-useful technical complexities, we deal with the most basic and simple games to show how they renew our ideas on what economics may add to the law. More importantly, we insist on the twofold features of cooperative game theory which can be considered from a positive point of view (how people cooperate and share the surplus created by cooperation) and from a normative view (how a judge or an arbitrator should settle a case when several people are in conflict about how to share a joint surplus).

First, definition and notation are introduced in order to better understand what cooperative game theory is. Second, we deal with more contemporary works using cooperative game to highlight regulation and public law issues. Third, we show that cooperative game approach is useful to renew private law and mainly torts. Last, we give some intuitions at a more abstract level to see how the legal theory could be influenced by the cooperative game approach.

Definition, Notation, and Meaning

A transferable utility game (TU game) is a couple (N, v) with N the set of players and v the characteristic function. The two basic blocks of cooperative game theory are the coalitions and the characteristic function. The Grand coalition $\{1, 2, \dots, n\}$ is the coalition of all the players. The singletons $\{i\}$ are coalitions restricted to single players. There are also all the subsets of players between singletons and N . With n players, there are $2^n - 1$ coalitions. The worth of the Grand coalition is $v(N)$ and is equal to the worth to be shared among players. The worth $v(i)$ is what player i could get if he decided to behave on his own. More generally, the worth $v(S)$, with S a coalition of players, is what the coalition S is able to get for its members. The characteristic function associates to each coalition its worth but says nothing about how this worth is then

shared among the members of S . Mostly, a hypothesis of transferability holds. Transferability is a serious assumption and states that utility is measurable in terms of money: side payments are possible among the players.

Once coalitions and characteristic function are defined, then the most important challenge for cooperative game theory is to solve the game that is to say to find a vector of payment $(x_1, x_2, \dots, x_n)$ for the players. At first sight, infinity of vectors could work. But, two rationality requirements will be added. First is collective rationality condition: the sum of the payments should be necessarily equal to the worth of the Grand coalition (no more, no less). The vector of payment should be feasible (it is impossible to give players more than what is created by the Grand coalition) and should ensure that resources are not wasted (it would be irrational to give them less than the surplus jointly created). Second is the individual rationality condition: the payment of a player i should be more important than its worth $v(i)$: without this condition, a single player would have no incentive to cooperate with others. The vectors which satisfy these two conditions are called imputations.

Then, a first set of solutions is the core. The core of a game $C(N, v)$ is defined as follows:

$$C(N, v) = \{x \in \mathbb{R}^n \mid x(N) = v(N) \text{ and } x(S) \geq v(S) \text{ for all } S \subset N\}$$

Behind the core is the idea of stability. In case of non-empty core, no individual nor any coalition has an interest to leave the Grand coalition for their own. On the contrary, an empty core means no guarantee on the stability of cooperation among players: some of them have an incentive to leave the Grand coalition and get more, out from the Grand coalition. Cooperative game theorists have then studied the conditions under which the core is non-empty (e.g., when games are convex – that is to say $v(S) + v(T) \leq v(S \cup T) + v(S \cap T)$ for all S and T – the core is non-empty). Still, the set of the core may be large enough.

It is possible to define other concept solutions with an axiomatic perspective. A trade-off arises between “not enough” and “too much” axioms.

If few axioms are required, it will be easy to find allocations which solve the game, but the set of solutions is likely to be too large. If too much axioms are required, the set of solution is likely to be empty. Each axiom is debatable on rational and normative grounds.

A particularly interesting rule to solve a game is the Shapley value. The Shapley value may be explained in two alternative ways. First, the Shapley value depends on the marginal contribution of players to the coalitions. For a player i , its marginal contribution to a coalition S is the difference between the worth of the coalition $v(S)$ and the worth of the coalition $v(S \setminus i)$. The Shapley value allocates to player i his average marginal contribution. Second, the Shapley value – defined on any cooperative game – follows four axioms which characterize uniquely this rule. The first is efficiency (the worth of the Grand coalition should be shared); the second is the null player axiom (players who contribute zero to any coalition get nothing back); the third is the additivity axiom (for a game that is the sum of two games, the Shapley value for the former is the sum of the Shapley values for the latters); and the fourth is symmetry (two substitutable players should receive the same payoff).

Beyond the core and the Shapley value, other sharing rules are discussed in the literature (e.g., the nucleolus). More recently, new paths in cooperative game theory have been developed: games with a priori unions (Owen 1977) or graph-restricted games (Myerson 1977). In these games, a structure of cooperation is defined through a network of preexisting links, and the value is calculated on this structure.

Cooperative Games, Antitrust, and Regulation: From Stability to Fairness

A contemporary perspective on cooperative game theory and the law would consider that public regulation is one of the main fields where cooperative game theory is useful. In part I, we have shown that cooperative game theory may be considered from a twofold perspective: first, it models “situations in which the players may conclude binding agreements that impose a particular action or a series of actions on each player” (Maschler

et al. 2013); second, axiomatization of solution concepts indicates how a judge or an arbitrator should allocate the value among the players. The first perspective could be said positive and the second normative. As soon as public agencies aim at regulating the behaviors of individual or firms implied in a common activity, cooperative game theory has something to say about the following: (1) Is cooperation among players stable? (2) How will be (or should be) the surplus due to cooperation shared?

A first subfield is antitrust. Collusion and anti-competitive practices may be considered as cooperation among players: colluders coordinate their actions in order to get monopoly profits. Some of antitrust scholars assert that cooperative game theory is irrelevant insofar as collusion being illegal, there is no way to conclude binding agreements; on the contrary others consider that the non-emptiness of the core is useful to better understand the stability of cartels and the efficiency of anticompetitive practices (Telser 1985; criticized in Wiley 1987). Analyzing the famous *Addyston Pipe* case, one of the most famous cases in antitrust according to Judge Bork, Bittlingmayer and Telser argue that cartelization may be useful for the firms to share fixed costs (Bittlingmayer 1982). This cooperation avoids the inefficiencies in sectors where marginal cost is under the average cost. This idea is followed by contemporary literature on oligopoly games. Both Cournot and Bertrand oligopolies are concerned. The main issue addressed is the stability of collusion. Some of anticompetitive behaviors (as prohibitive restrictive agreements or concerted practices) imply that players coordinate their strategies to get their best outcome and may organize side payments. In some models, a hypothesis of sharing the best technology is done, but it is not a necessary condition (Lardon 2017). The stability of the agreement among members of the oligopoly is one of the key elements of this literature. The core is consequently the most studied solution: if the core is non-empty, stability of the cartel is expected; on the contrary, the emptiness of the core implies that the cartel is unstable insofar as some players or groups of players have an incentive to leave the Grand coalition. Zhao applies this

reasoning to the sugar cartel in the USA at the end of the nineteenth century. For this author, the increasing of the cartel organizational costs due to the Sherman Act made the core empty and lead to instability (Zhao 2014). What is interesting in this literature of oligopoly games is to work out different scenario implying different blocking rules. Indeed, when a player or a group of players decide to leave the Grand coalition, the interaction between them and the remaining players becomes strategic, and several strategies are conceivable. For example, they may behave to maximize their own utility or to minimize the utility of the other players. The conclusions drawn are particularly interesting for antitrust authorities regarding the best way to enforce antitrust law and to enhance instability of collusion.

A second subfield is concerned with public regulation, public utilities, and facilities at large. In many contexts covered by public regulation, agents (firms, individuals, or groups like municipalities) cooperate and are looking for the best way to share their costs: fisheries conferences, commons, pools, water or telecom networks, facilities that will be jointly used, and water from a river at a national or international level are some of the numerous examples of such situations. Take the example of the airport fees analyzed by Littlechild and Thomson (1977) which deals with how to share the cost of a common facility (a runway). Assume three aircraft companies which cooperate to build a new runway. Due to the size of the planes they own, the first company needs a small airstrip, the second a medium one, and the third a large one. The cost of each airstrip is $c1 < c2 < c3$. If firms do not cooperate, they bear a total cost of $c1 + c2 + c3$, while cooperation leads to a total cost equals to $c3$ (we suppose that there is no congestion and the largest airstrip is enough to land all the aircrafts whatever their size is). The core is non-empty but large enough and let unsolved the precise cost paid by each firms. A regulator could use more specific sharing rule. A natural solution would say that the total cost $c3$ should be divided as follows: the smallest part of the airstrip is used by all the companies: $c1$ should be equally divided among the three companies. The additional cost ($c2 - c1$)

should be paid only by the companies 2 and 3. The additional cost ($c3 - c2$) should be paid by the third firm (the only firm which needs a large airstrip). This intuitive solution is the Shapley value of the airport game. As such, the Shapley value is a fair compromise that a regulator could implement to sharing the costs among firms.

Sometimes, legislators and regulators play a major role in cooperation. This is the case when regulated firms are required by the law to cooperate. The REACH legislation in the European Union (Registration, Evaluation, Authorization and Restriction of Chemicals) is very illustrative. The European Union has decided in 2006 to make a systematic evaluation of the toxicity of chemicals. The number of chemicals is so high that it is impossible for public authorities to evaluate by themselves their danger. At the same time, firms privately own information regarding the toxicity of chemicals (scientific reports, experiences, etc.). REACH legislation creates SIEF which are forum to organize the exchange of information among firms. The difficulty lies in the side payments: some firms may provide very valuable data, some add information already get by others, some have no information at all, etc. Based on the costs of replication of data, Dehez and Tellone (2013) provide a cooperative game model and advocate the Shapley value as a fair compromise to calculate payments/rewards of each participant to a SIEF. Beal and Deschamps follow this path and discuss different rules in order to compare their axiomatic properties (Beal and Deschamps 2016).

Last, and more recently, a third important subfield using cooperative game theory has grown: the implementation of public matching algorithms. Regulators may face complex issues of matching the two sides of the market. In a famous paper, Roth studied the National Resident Matching Program in the USA that aims at matching hospitals with resident doctors (Roth 1984). The NRMP is a centralized “clearing-house” introduced in the 1950s to allocate resident doctors to hospitals. Roth discovered that the algorithm used in the NRMP is very close to a Gale-Shapley algorithm (Gale and Shapley 1962) that aims at finding stability allocation (stable

matches means that no matched couple would like to break up and forms new matches to be better off). The key aspect from the regulator perspective is that there are several stable sets with different normative properties regarding the side of the market which is favored. Such public algorithms are now implemented by the law in many fields including health care, education (universities and applicants), labor markets, etc.

Cooperative Game Theory and Private Law

Private law – property, torts, and contracts – is also a field studied by cooperative game theory. First, property rights are concerned. In a famous paper on land ownership, Shapley and Shubik (1967) show how different types of property (feudal world, private property, village commune, corporate and joint ownership, etc.) may be modeled in terms of cooperative games. Shapley and Shubik show how the characteristic function has to be changed to well describe different types of property rights.

Second, tort law has been studied through cooperative game theory. Here too, literature is divided in a positive branch and a normative one. From the positive point of view, cooperative game theory has added to a better understanding of the Coase theorem. The Coase theorem is well known in law and economics: in case of zero transaction costs, competition, and perfect delineation of property rights, agents may reach a mutually advantageous agreement. The level of externality is independent of the initial distribution of property rights and maximizes the value of the production (the result is invariant and efficient). Aivazian and Callen (1981) reconsidered the issue and show that the Coasean result is not robust (Coase 1981) when there are more than two players (e.g., several polluters and one victim): the emptiness of the core leads to the impossibility to reach a stable agreement (players have incentive to block any agreement with another coalition). More recently, Aivazian and Callen (2003); Gonzalez et al. (2016); Gonzalez and Marciano (2017), and others provided more complex examples with several victims. The results converge: the Coase theorem does not hold insofar as the delineation of property right influences the

emptiness of the core. More importantly, positive transaction costs lead unexpected results: some authors consider that positive transaction costs make the stability of agreement more likely (precisely because renegotiation is costly), and others think not.

In a different perspective, Dehez and Ferey use cooperative game theory from a purely normative point of view as a description of legal adjudication (Ferey and Dehez 2016a, b). The cases studied are tort cases implying multiple tortfeasors who jointly cause a unique harm to a victim or a group of victim. They first consider the Shapley value to estimate the part of the damage to be paid by each tortfeasor, and then they show that the principles of the American Restatement are consistent with the Shapley value principles. Concept solutions are here considered from a normative perspective to better solve conflicts among defendants about their respective shares of responsibility and to describe the behaviors of judges.

In the same vein, a third interesting example regarding private law is about debt and insolvability. In a famous model, Aumann and Maschler (1985) address the issue of the burden of insolvability of a firm. In that case, suppose that players are creditors and get claims against a firm (the debtor) which is unable to pay all its debts back. Suppose E be the total amount of value available (the value of the assets left) and suppose d_1 , d_2 , and d_3 the claims of the debtors 1, 2, and 3 with $d_1 + d_2 + d_3 > E$. In that case, the worth of each coalition S is defined as the maximum that S can get once all the others creditors ($N \setminus S$) get their claims back, $d(N \setminus S)$. Formally, $v(S) = \text{Max} (0, d(N \setminus S))$. The authors compare different solution concepts and show that the rules advocated in the Talmud is close to the nucleolus. Their paper is not explicitly oriented in a law and economics perspective, but they suggest that cooperative game theory is very useful to better understand the deep reasons of legal and/or moral reasoning.

Many other examples from private law could be added: patents (several firms cooperate to get a patent), condo (several people have to share the costs of a condo, Crettez and Deloche 2014), or

even the contracts used by bitcoin miners when they pool together to find the relevant hash and have to share the bitcoins get in common. These examples show how fruitful cooperative game theory is for the law.

Conclusion: Cooperative Games and Legal Theory

Law seems to be a quite natural field of application of cooperative game theory. We have provided many examples of legal topics studied with cooperative game theory. To conclude, we would like to add some more speculative views. Is cooperative game theory descriptive, predictive, or normative (Aumann 1985)? Sure, in some cases, cooperative game theory is useful to predict how agents will behave; in other cases, the normative aspect of cooperative game theory is more interesting and provides some axiomatic solutions that help judges or arbitrators to settle conflicts. As Aumann states, “Normative aspects of game theory may be subclassified using various dimensions. One is whether we are advising a single player (or group of players) on how to act best in order to maximize payoff to himself, if necessary at the expense of the other players; and the other is advising society as a whole (or a group of players) of reasonable ways of dividing payoff among themselves. The axis I’m talking about has the strategist (or the lawyer) at one extreme, the arbitrator (or judge) at the other” (Aumann 1985).

Cross-References

- Coase Theorem
- Coase Theorem and the Theory of the Core, The
- REACH Legislation

References

- Aivazian VA, Callen JL (1981) The Coase theorem and the empty core. *J Law Econ* 24:175–181
- Aivazian VA, Callen JL (2003) The core, transaction costs, and the Coase theorem. *Constit Polit Econ* 14:287–299

- Aumann RJ (1985) What is game theory trying to accomplish ? In: Arrow K, Honkaphoja S (eds) *Frontiers of economics*. Basil Blackwell, Oxford, pp 5–46
- Aumann RJ, Maschler M (1985) Game theoretic analysis of a bankruptcy problem from the Talmud. *J Econ Theory* 36:195–213
- Baird DG, Gertner RH, Picker RC (1994) *Game theory and the law*. Harvard University Press, Harvard
- Beal S, Deschamps M (2016) On compensation schemes for data sharing within the European REACH legislation. *Eur J Law Econ* 41:157–181
- Bittlingmayer (1982) Decreasing average cost and competition: a new look at the *Addyston Pipe* case. *J Law Econ* 25:201–229
- Coase RH (1981) The Coase theorem and the empty core: a comment. *J Law Econ* 24:183–187
- Crettez B, Deloche R (2014) Cost sharing in a condo under law’s umbrella. Working paper
- Dehez P, Tellone D (2013) Data games: sharing public goods with exclusion. *J Public Econ Theory* 15:654–673
- Ferey S, Dehez P (2016a) Multiple causation, apportionment and the Shapley value. *J Leg Stud* 45:143–171
- Ferey S, Dehez P (2016b) Overdetermined causation, contribution and the Shapley value. *Chicago-Kent Law Rev* 91:637–658
- Gale D, Shapley LS (1962) College admissions and the stability of marriage. *Am Math Mon* 69:9–15
- Gonzalez S, Marciano A (2017) New insights on the Coase theorem and the emptiness of the core. *Revue d’Économie Politique* 127:579–600
- Gonzalez S, Marciano A, Solal P (2016) The Social cost problem, rights and the (non)empty core. Working paper. Available on https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2863952
- Lardon A (2017) Coalition games and oligopolies. *Revue d’Économie Politique* 127:601–636
- Littlechild SC, Thomson GF (1977) Aircraft landing fees: A game theory approach. *Bell J Econ* 8:186–204
- Maschler M, Solan E, Zamir S (2013) *Game theory*. Cambridge University Press, Cambridge
- Myerson R (1977) Graphs and cooperation in games. *Math Oper Res* 2:225–229
- Owen G (1977) Values of games with a priori unions. In: Moeschlin O, Hein R (eds) *Mathematical economics and game theory: essays in honor of Oskar Morgenstern*. Springer, New-York
- Roth AE (1984) The evolution of the labor market for medical interns and residents: A case study in game theory. *J Political Econ* 92:991–1016
- Shapley L, Shubik M (1967) Ownership and the production function. *Q J Econ* 81:88–111
- Telser L (1985) Cooperation, competition, and efficiency. *J Law Econ* 28:271–295
- Wiley JS (1987) Antitrust and core theory. *J Law Econ* 54:556–589
- Zhao J (2014) Estimating the merging costs and organizational costs: methodology and the case of 1887–1914 sugar monopoly. Working paper, University of Saskatchewan

Copyright

Antonio Rodriguez Andres¹ and Nora El-Bialy²

¹Departamento de ADE y Economía, Universidad Camilo José Cela, Villanueva de la Cañada, Madrid, Spain

²Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Abstract

IPRs indices in general developed to measure the quality or strength of IPR institutions across countries usually do not differentiate between *de jure* and *de facto* IPR institutions. In addition, they neglect the different impact these individual variables comprising these indices might have by simply averaging all variables or assigning arbitrary weights. The main contribution of this essay is to shed light on the relative importance of individual institutional variables in forming the *de facto* institutional framework of IPR protection as opposed to their *de jure* counterparts mainly focusing on Copyright. For that purpose we discuss the drawbacks of the common indices and offering some suggestions for building more reliable ones. Our main recommendation is to look at the formal copyright institutions in a more careful way to be able to code the different provisions that we think are probably most influential for Institutional Quality and that would enable better enforcement. Second: Countries must make data available about the enforcement process of Copyright laws and number of piracy cases filed in courts and the imposed sanctions in order to be able to develop a *de facto* measure for the quality of copyright institutions.

Definition

Copyright means individuals are not allowed to copy someone else's works.

Introduction

In legal terms, intellectual property rights (henceforth, IPRs) refer to legal rights to creations of mind like inventions, literary works, artistic works, etc. IPRs are divided into two main categories: industrial property that includes inventions (patents), trademarks, industrial designs, geographic indications of source, copyright, and rights related to copyright. Although the interest in IPRs has been increasing, most empirical research works are focused on patents leaving aside other types of IPR protection. As regards IPR protection, one serious concern for copyright holders is piracy: that is, the unauthorized use of copyrighted goods. Even though piracy occurs for all types of intellectual property and can take several forms depending on the access type and intellectual property mechanism, one of the most worrying areas is the piracy of business software applications. This entry addresses the study of copyright institutions and intends to measure the strength of the IP systems for copyright institutions for a cross-national sample as other authors have done for copyright laws in Europe (Andres 2006).

The divergence between institutional reforms (legal reforms) and factual implementation of copyright laws has recently become more critical and apparent (Table 1 in the appendix shows the number of IPR laws issued or amended and related agreements signed by each country as opposed to the prevailing software violations in order to show that the existence of copyright institutions is necessary but cannot be considered a sufficient condition for enforcement). The correlation between the number of copyright laws issued and the perceived level of IPR protection is -0.34 . In the absence of enforcement and adequate sanctions, IPR reforms and signed agreements tend to fall short of their proclaimed goals, leading to inefficient IPR institutions. Hence, in order to design an efficient IPR reform strategy, one must first identify whether the failure of existing IPR laws is due to the lack of formal written aspects of IPR legislations (*de jure* IPR institutions) or it is caused by the inefficiency of enforcement authorities responsible for implementing the law (*de facto* IPR institutions).

Copyright, Table 1 The number of IPR-related laws issued and agreements signed by each country

	Agreements	Laws	Software piracy rates
Argentina	14	7	69.88
Armenia	9	3	92.33
Australia	15	14	30.52
Austria	15	14	32.52
Azerbaijan	10	1	91
Belgium	12	8	32.70
Bolivia	11	12	83.05
Bosnia	15	3	67.71
Brazil	14	5	61.47
Bulgaria	9	1	77.64
Canada	8	15	36.88
Chile	11	14	60.65
China	9	19	89.29
Colombia	9	11	58.52
Costa Rica	15	5	69
Croatia	13	7	64.1
Cyprus	12	2	59.76
Czech Republic	13	10	44.47
Denmark	15	2	29.70
Ecuador	15	13	70.76
Egypt	5	2	68.59
El Salvador	8	5	82.82
Estonia	7	11	52.2
Finland	12	9	31.70
France	16	34	44
Germany	17	15	31.23
Greece	12	16	67.29
Honduras	10	5	75.58
Hong Kong	1	16	55.17
Hungary	13	3	50.88
Iceland	6	23	50.66
India	12	5	70
Indonesia	5	2	89.23
Ireland	9	30	47.2
Israel	9	12	43.9
Italy	16	38	50.2
Japan	14	21	32.9
Jordan	6	4	69.3
Kazakhstan	7	0	80.9
Kuwait	2	1	74.5
Latvia	9	2	56.9
Lithuania	9	10	55.5
Luxembourg	11	4	32.8
Malaysia	4	10	66.6
Malta	6	4	54.6

(continued)

Copyright, Table 1 (continued)

	Agreements	Laws	Software piracy rates
Mauritius	8	2	67.7
Mexico	13	11	62.1
Morocco	13	9	68.9
Netherlands	14	9	39.2
New Zealand	9	17	27.7
Nicaragua	14	9	81.6
Norway	12	14	36.6
Pakistan	7	3	85.7
Panama	16	7	70.2
Paraguay	13	2	83.2
Peru	14	14	69.6
Philippines	11	15	74.2
Poland	11	4	59.9
Portugal	15	31	45.8
Romania	7	3	76.2
Russia	12	21	83.8
Saudi Arabia	7	4	60.1
Serbia	15	2	76.6
Singapore	7	11	46.3
Slovakia	14	4	49.2
Slovenia	17	9	63
South Africa	5	12	42
Spain	13	16	51.8
Sweden	16	10	33.6
Switzerland	30	6	32
Thailand	4	3	79.4
Tunisia	11	3	77.4
Turkey	7	7	70.8
Ukraine	11	13	88.9
United Kingdom	15	42	28.9
United States	15	7	23.4
Uruguay	13	16	71
Venezuela	10	11	72.6
Vietnam	7	14	92.8
Zambia	4	1	82.2

Sources: Software piracy rates. Business Software Alliance (BSA) (2009) Sixth annual BSA and IDC global software piracy studies (online). BSA. <http://global.bsa.org/globalpiracy2008/index.html>. Accessed Mar 2010

The first option requires a reform strategy targeting improvement of the existing IPR laws and regulation by issuing, for example, further enforcement or sanctions. The latter requires

launching a judicial reform or improving the quality of the enforcement organizations in general.

It is worth noting that most pirating or counterfeiting countries have already issued national IPR laws and may even have special provisions that serve as a copyright law with a relevant sanction mechanism. However, these laws have not translated into significantly lower software piracy rates. Hence, it appears that the existing IPR institutions might turn out to be inefficient, especially in most developing countries. In such cases, the use of IPR rules and legal provisions might be an unreliable measure for reflecting the level of IPR enforcement in a country. Thus, there is a wide gap in our understanding of how to measure the quality of copyright institutions. Therefore, measuring the strengths of IP systems for copyright at country level may enable a better policy design by giving information about the determinants of IP protection, as well as their economic and social effects. This is the problem this entry addresses.

Measuring the Quality of Copyright Institutions

The level of economic development is considered a main determinant of all forms of IPR piracy including copyright infringement across countries (Maskus and Penubarti 1995; Park and Ginarte 1997; Gopal and Sanders 1998, 2000; Maskus 2000). With the recent rise of the new institutional economics (NIE) and the growing role of institutions in economic studies, scholars started to highlight the impact of institutions on IPR piracy.

Institutions in general comprise a rule that is accompanied by a sanction. Accordingly, institutions that lack a sufficient sanctioning mechanism cannot fulfill their role and end up written down *only* in formal decrees. Hence, setting institutions that are in charge of administering the system constitutes just one of the pieces that form a national system of IPR protection. These institutions must interact with some organizations in order to realize factual implementation and protect these rights and enforce the institutions. North (1981) described the

process of property right enforcement as making it obligatory or put it into force. This process protects right holders by preventing the infringement of property rights by free riders. Thus, organizations in charge of enforcing rights become crucial elements of the system's overall effectiveness, as they play a main role in setting the legal infrastructure.

Institutions are just the rules of the game, while organizations are the players. Hence, it can be said that the efficiency of institutions and the performance of organizations are highly interdependent.

Mansfield (1994) determined three areas of concern in assessing the strength of IPR protection in a country. These are (a) existing laws, (b) the prevailing legal infrastructure, and (c) the willingness of governments to actively enforce property rights. Whereas development economists have typically treated the state as an exogenous actor in the development process, North (1990), however, pointed out that the state can never be treated as an exogenous actor in development policy. He explained that third-party enforcement represents the development of the state as a coercive force able to monitor property rights and enforce contracts effectively. In other words, this third party is responsible for enforcing the contracts carried out by different parties and punishing the party that violates the agreement (Harris et al. 1995) (for further information, see Harris et al. 1995, p. 22; Sened 1997, p. 49). The enforcement of agreements in economic markets is ultimately a function of the political markets of economies. The political market imposes the polity that specifies property rights and provides the instruments and resources to enforce contracts. Hence, an explanation of property right enforcement in general requires an understanding of the behavior of the government and assumption that government is exogenous because government normally defines and enforces property rights (North 1990). Even so, the government is not the sole actor of the enforcement process, as its main role practically ends after formally signing the law. Afterward, the implementation process will be handed to the law enforcers which are responsible for deterring and sanctioning those who break the law. It starts with the police who raid the suspected sites and catch infringers to file a case. After filing it, the case will be handed over

to the prosecutors to gather all the needed information and evidence to finally submit it to the courts. Hence, it can be considered that the overall efficiency of the formal enforcement system in a country to handle copyright infringement among other cases of commercial or civil trials plays a huge role in determining the adequacy of copyright law enforcement. This process, where it is inadequate, requires special reforms within the enforcement organizations themselves in order for the participating bodies (police, prosecutors, and judiciary) to enjoy a high degree of accountability. Williamson (1975) analyzes the overall impact of poor performance of courts on contractual obligations. The study shows that court litigations are time-consuming and might result in errors or in the absence of choice. This can happen when judges dismiss cases due to wrong litigation procedure or the lack of sufficient evidence, which is typical for most copyright suits.

Aspects of the “De Jure” Copyright Institutions

De jure IPR institutions include formal institutions (laws, formal amendments of rules, agreements, etc.) designed to protect IPRs.

For research purposes, many indicators are used to measure the quality of institutions within the IPR research context. For instance, using a measure of institutional strength developed by Knack and Keefer (1995) as an explanatory variable for IPR infringement, Marron and Steel (2000) argue that institutional factors in general have greater influence on piracy than economic conditions of a country. This indicator uses data published in the firm’s International Country Risk Guide (ICGR) and identifies five variables related to property rights protection, which are tradition of law and order, the government’s propensity to repudiate contracts, the quality of bureaucracy, the extent of corruption, and risk of expropriation. By contrast, most studies use either a dummy variable to represent the strength of IPR institutions in a country or an index composed of multiple variables reflecting the quality of institutions at the national level. Both methods are problematic to some extent as the brief discussion below illustrates.

A Dummy Variable Approach

Using a dummy variable representing a country’s membership in international agreements or the possession of a national IPR code is often used to represent copyright institutions (e.g., Papadopoulos 2003; Van Kranenburg and Hogenbirk 2005; Andres 2006). However, as mentioned before, most countries are already members in one or more international IPR agreements and already have issued their own IPR codes. Hence, the use of a dummy variable to represent membership in agreements or IPR laws would be inappropriate. This is a necessary condition but not sufficient as argued by Maskus (2000). As argued in the theoretical part of the study and supported by the data observed from reviewing the IPR laws and regulations of different countries, it is noticed that countries keep changing their IPR laws by adding extra provisions, extending the scope of coverage, or maybe even issuing a new law to ensure a better enforcement. For this purpose, counting the number of laws issued or agreements signed by a country over time might be considered as a proxy for measuring the efforts made toward improving IPR institutions rather than reflecting the qualities of these institutions. The existence of the copyright institution per se cannot be considered a determinant of its quality, but rather we should focus on the details and the provisions included in the law and the efforts of the legislators to improve the code by doing amendments and modifying the framework of the law to suit the countries’ conditions. Especially, in copyright software piracy matters and with the fast growing technology and innovative piracy devices, codes and provisions have to be updated through time to ensure clear enforcement provisions.

Another appropriate measure would be constructing variables through revising existing copyright laws and international copyright agreements in order to search for differences among the single provisions of the different laws. The index for patent rights developed by Ginarte and Park (1997) was the first attempt toward coding the content and scope of the laws but with the fast developments of the IP

codes in most countries. They have composed a patent protection index for a panel of 110 countries from 1960 to 1990. It uses an unweighted sum of five variables to build up the index. The variables are the scope of coverage, membership in international agreements, provisions for loss protection, enforcement mechanisms, and the duration of IPR protection (for more details about the composition of the index, see Ginarte and Park (1997), pp. 286–287). Each variable is assigned a value of one if present and zero otherwise, and then all of them are aggregated to obtain a maximum score of 5.0 in case all 5 factors were taken care of in each countries' law. However, there have been a variety of new laws issued and agreements signed, such that the differences in the variables used in their index became invariant among countries to a large extent. One important remark is that the indexes are largely measuring the laws on the books rather the actual practice. A common concern is that these IP indexes measure only the perceived protection not actual protection. As argued by Park (2001), nevertheless, there is a high correlation between statutory laws and the current level of enforcement. Countries that have strong laws on the books tend to be the ones that carry out the laws. Hence, developing measures that capture the significantly different aspects of IPR codes across countries might be a more proper way to measure IPR institutions (see the UNESCO portal: http://portal.unesco.org/culture/en/ev.php-URL_ID=39069&URL_DO=DO_TOPIC&URL_SECTION=201.html and WIPO: <http://www.wipo.int/clea/en/> and <http://www.wipo.int/about-ip/en/ipworldwide/country.htm>). For example, one could code the severity of sanction provisions if mentioned at all in a law. In addition, one could also count the number of clauses within each country law regarding the implementation and enforcement procedure of a piracy crime. It might be expected that countries having specified implementation clauses in their law will observe better law enforcement and hence experience lower piracy rates. Finally, a variable measuring the number of copyright laws and regulations issued by executives rather than the legislative body might also be

significant in determining the quality of copyright institutions.

Aspects of the “De Facto” Copyright Institutions

In an attempt to measure the factual IPR law implementation or determine the quality of enforcement, formal empirical studies used the level of corruption in a country as a proxy for institutional quality (e.g., Ronkainen and Guerrero-Cusumano 2001; Papadopoulos 2003; Bagchi et al. 2006). The indicators of Kaufmann et al. (2011) “graft/corruption index” (Kaufmann et al. 2011) in addition to the Transparency International's 1999 Corruption Perception Index (CPI) (Lambsdorf 2000) are mainly used in the previous studies. These indices reflect the compilation of perceptions of the quality of governance and are used to represent the overall corruption level among custom agents, police, prosecutors, and judicial in the country.

Whereas Shadlen et al. (2005) employ the Kaufmann et al. “government effectiveness indicator” to examine the effect of institutional factors on piracy, Holm (2003) considers the “rule of law” as most appropriate when attempting to proxy for the institutional aspects of a country, as it aims to measure the efficiency of the judicial system (available in Kaufmann et al. (2004)). Fischer and Andrés (2005) explain that the degree of efficient law enforcement which is determined by the rule of law variable may increase the probability of imposing a deterrent punishment; hence, it may best capture the de facto aspects of copyright institutions. The rule of law measures the government's administrative capacity in enforcing the law. More recently, Ping (2010) uses the rule of law to measure laws and institutions and finds it to be an important determinant of the rate of software piracy.

However, using the Worldwide Governance Indicators (henceforth, WGI) of Kaufmann et al. (2006, 2011) as measures of the institutional quality of a country might be however somehow problematic (the WGI comprise six dimensions of governance: voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law, and

control of corruption). In addition, most of these bundled indicators may be criticized by the fact that they do not focus on a single variable or institution but relate to dozens or even hundreds of variables that are not equally weighted and are not necessarily related to each other. Also note further that the weights of individual variables appear to be different when comparing the old with new WGI versions. Voigt (2013), e.g., shows that most of the partial correlations between a selected number of the rule of law variables are quite low, which implies that the rule of law is indeed made up of various dimensions that are not necessarily highly correlated with each other. Moreover, each variable or group of variables synthesized in each indicator is weighted differently from another indicator. Hence, it can be said that the indicator does not represent the variables with equal strength, which may hinder one's ability to identify the relative significance, if any, of each institution or variable individually.

We argue that it might be useful to unbundle the indices and to consider the performance of the different enforcement organizations (police, judiciary, executives) individually to measure de facto IPR institutions in a more accurate way. For that purpose, we raise the following two questions: (i) Is the IPR enforcement process treated in the way a regular penal act is treated (pirates are caught by the police), or is it handled by "specialized agencies? (ii) Are copyright infringement cases filed in specialized courts, or are they counted as civil or criminal acts and hence treated in regular first instance courts? All these details matter; if copyright infringement cases are treated in a special manner and in a way that deviates from the regular penal cases, then accounting for the quality or accountability of the police and the regular court becomes irrelevant and hence would not be a proper measure for de facto copyright institutions and hence enforcement.

Cross-References

► [Economic Analysis of Law](#)

Appendix

See Table 1.

References

- Andrés A.R. 2006. The relationship between copyright software protection and piracy: evidence from europe. *European Journal of Law and Economics*, 21, 29–51
- Bagchi K, Kirs P, Cervený R (2006) Global software piracy: can economic factors alone explain the trend? *Commun ACM* 49(6):70–75
- Business Software Alliance (BSA) (2009) Sixth annual BSA and IDC global software piracy studies (online). BSA. <http://global.bsa.org/globalpiracy2008/index.html>. Accessed Mar 2010
- Fischer J, Andrés A (2005) Is software piracy a middle class crime? Investigating the inequality-piracy channel. University of St. Gallen working paper series, Department of Economics, University of St. Gallen
- Ginarte JA, Park W (1997) Determinants of patent rights: a cross-national study. *Res Policy* 26:283–301
- Gopal R, Sanders G (1998) International software piracy: an analysis of key issues and impacts. *Inf Syst Res* 9(4):380–397
- Gopal RA, Sanders G (2000) Global software piracy: you can't get blood out of a turnip. *Commun ACM* 43(9):82–89
- Harris J, Hunter J, Lewis C (1995) *The new institutional economics and third world development*. Routledge, London
- Holm H (2003) Can economic theory explain piracy behaviour? *Top Econ Anal Pol* 3(5):1–15
- Kaufmann D, Kraay A, Mastruzzi M (2006) *Governance matters V: governance indicators for 1996–2005*. World Bank policy research working paper
- Kaufmann D, Kraay A, Mastruzzi M (2011) *Worldwide governance indicators*. Accessed at: <http://info.worldbank.org/governance/wgi/index.asp>, Accessed on 20 Feb 2014
- Knack S, Keefer P (1995) Institutions and economic performance: cross-country tests using alternative institutional measures. *Econ Polit* 7:207–227
- Mansfield E (1994) Intellectual property protection, foreign direct investment and technology transfer, International finance corporation discussion paper, 19. The World Bank. <http://www.bvindicopi.gob.pe/colec/emansfield2.pdf>. Accessed Oct 2008
- Marron D, Steel D (2000) Which countries protect intellectual property? The case of software piracy. *Econ Inq* 38:159–174
- Maskus K (2000) *Intellectual property rights in the global economy*. Institute for International Economics, Washington, DC
- Maskus K, Penubarti M (1995) How trade related are intellectual property rights? *J Int Econ* 39: 227–248

- North D (1981) Structure and change in economic history. Norton, New York
- North D (1990) Institutions. Institutional change and economic performance. Cambridge University Press, New York
- Papadopoulos T (2003) Determinants of international sound recording piracy. *Econ Bull* 6(10):1–9
- Png IP (2010) On the reliability of software piracy statistics. *Electronic Commerce Research and Applications* 9:366–373.
- Park W (2001) Intellectual property and economic freedom. In J. Gwartney and R. Lawson (eds.), *Economic Freedom of the World*, Vancouver: Fraser Institute
- Park W, Ginarte J (1997) Intellectual property rights and economic growth. *Contemp Econ Policy* 15(3):51–61
- Ronkainen I, Guerrero-Cusumano J (2001) Correlates of intellectual property violation. *Multinatl Bus Rev* 9(1):59–65
- Sened, I (1997) The political institution of private property. Cambridge: Cambridge University Press
- Shadlen K, Schrank A, Kurtz M (2005) The political economy of intellectual property protection: the case of software. *Int Stud Q* 49:45–71
- Van Kranenburg H, Hogenbirk A (2005) Multimedia, entertainment, and business software copyright piracy: a cross-national study. *The Journal of Media Economics*, 18:109–129
- Voigt S (2013) How (not) to measure institutions? *J Institut Econ* 9(1):1–26
- Williamson O (1975) Markets and hierarchies: analysis and antitrust implications: a Study in the economics of internal organization. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship (online). <http://ssrn.com/abstract=1496220>. Accessed Mar 2010
- WIPO (2001) WIPO intellectual property handbook: policy, law, and use, vol 489(E), WIPO Publication. WIPO, Geneva
- Lambsdorf J (2000) Background paper to the 2000 corruption perceptions index, Transparency International, Sept. www.transparency.org. Accessed on Feb 2014

the optimal structure of corporate sanctions to deter crimes. The distinction between civil and criminal corporate liability is addressed, and a brief discussion of the corporate criminal enforcement in the United States and Europe is presented.

Definition and Structures of Corporate Liability

Corporate liability is the liability of one party (the firm) for the misconduct of another party (the employee). Corporate liability can assume different forms ranging from *vicarious liability* to some forms of *duty-based liability*. Vicarious liability is a strict (absolute) form of secondary liability that arises under the legal doctrine of *respondeat superior*, that is the common law doctrine of agency for which a party (the principal) is responsible for the acts committed (within the scope of employment) by its agents that has the legal right or duty to control. Duty-based liability regimes are forms of secondary liability where the principal is held liable for the misconduct of the agent only if he contravenes a legal duty – that is, if he does not observe due care in preventing or reporting a violation. There may also be mixed regimes under which firms are liable, but the level of sanctions depends on whether the corporation fulfilled its policy responsibilities.

Besides being vicarious or duty-based, corporate liability may be civil or criminal. This entry addresses the theoretical reasons for corporate criminal liability by discussing the following three issues:

- (i) Under what conditions it is optimal to impose sanctions also on the firm (principal) rather than only on the employee (agents) who misbehaved, namely, why there should be *joint individual and corporate liability*?
- (ii) Which form of corporate liability, vicarious versus duty-based, is socially optimal?
- (iii) Whether the socially optimal corporate liability should be criminal or not. We then briefly discuss the corporate liability in the USA and Europe.

Corporate Criminal Liability

Paolo Polidori and Désirée Teobaldelli
Department of Law, University of Urbino,
Urbino, Italy

Abstract

This entry reviews the literature on corporate criminal liability. It first describes the different forms of corporate liability and then discusses

The Need of Joint Individual and Corporate Liability

To understand the optimal deterrence of corporate crimes, it is useful to keep in mind that such crimes are crimes committed by individuals working for firms, i.e., they are crimes committed in presence of an agency relationship between the firm and the employee. However, the starting point of the analysis is represented by the results of the classic model of individual criminal liability in absence of corporations when individuals are risk neutral and are not wealth constrained and sanctioning costs are zero (Becker 1968; Polinsky and Shavell 2000). As individuals will commit a crime only when the expected benefit exceeds the expected cost, the socially optimal level of deterrence requires that the state imposes a criminal fine equal to the ratio between the social cost of crime to society and the probability of crime being detected. This result also holds when the probability of detection depends on the enforcement expenditure with the caveat that the optimal deterrence scheme involves minimizing enforcement costs, which means that the probability of crime detection is at its lower bound. The analysis for unintentional crimes leads basically to same results as the state will induce the optimal individual's level of effort to avoid the crime by choosing a sanction scheme that equalizes the expected sanction to the social cost of crime.

Once we introduce corporations into the analysis, the first issue to be addressed is whether corporate liability is necessary to optimally prevent corporate crimes. Let us begin considering a framework without imperfections, i.e., a world where the following conditions are satisfied:

1. Firms and employees have no wealth constraints.
2. There is a positive probability that the state sanctions a crime even without spending resources on enforcement.
3. There are no costs on imposing sanctions.
4. All parties are rational and there is perfect information.
5. Firms and employees can contract at zero cost.

In this perfect world, there is no justification for joint individual and corporate liability because crime can be optimally deterred by imposing the liability on either the individual or the firm at the same social and private costs. As the state is indifferent between individual and corporate liability (i.e., the two forms of liability are complete substitute in this framework), this result is known in the literature as the neutrality principle (Kornhauser 1982; Sykes 1984; Polinsky and Shavell 1993).

The assumptions on which the neutrality principle is based are very strong and are generally not satisfied in reality. This opens the way for an important role played by corporate liability. In particular, it is immediate that the state cannot optimally deter corporate crime using individual liability and monetary sanctions because employees do not generally have sufficient assets to pay the sanctions; corporate crimes often generate large social costs (see, e.g., the case of financial and consumer frauds), and the probability of sanction for such kind of crimes tends to be low, especially in absence of significant expenditures on enforcement. Again, while imprisonment can increase the level of sanctions imposed on the wealth constrained individuals, this is unlikely to make individual liability alone sufficient to deter corporate crime for the following reasons:

- (a) The maximal punishment, i.e., life imprisonment, is likely to imply a limited duration of this penalty given that corporate wrongdoers tend to be relatively old.
- (b) Imposing long prison sentences for nonviolent crimes, such as corporate crimes, may not be possible because this reduces the room of appropriate sanctions for violent criminal offences.
- (c) Imprisonment entails very large costs to society both because of the standard reasons associated to the detention of individuals (e.g., removal of productive individuals from the labor market, expenditures for guards, protective services, and equipment) and the losses suffered by risk-averse agents who face the risk of being liable also when they have not committed the crime.

Another important reason behind the failure of the neutrality principle and the need of corporate liability is related to the complexity of corporate crime as such crimes may involve many individuals and determining those responsible for the wrongdoings is often a difficult task. This means that optimal deterrence requires that both the state and the firm spend resources on prevention and enforcement. In particular, the firm can play an important role in deterring corporate crimes by:

- (i) Adopting measures of crime prevention that reduce the incentives to commit a crime; this can be made by adopting policies, such as those related to compensation and promotion that influence the extent to which the employee benefits from the crime (e.g., firms may limit the adoption of high-powered short-run compensation policies who generally provide an incentive to managers to commit crimes; Arlen and Carney 1992); corporations can also make committing crime more costly by promoting a culture of legal compliance within the firm so to increase the likelihood that employees report suspect wrongdoings or that wrongdoers bear higher direct psychological costs (Conley and O'Barr 1997; Tyler and Blader 2005).
- (ii) Implementing policing measures that increase the probability of crime detection and conviction; this can be done *ex ante* through the adoption of compliance (monitoring) programs that allows the firm to detect crime and to collect evidence on wrongdoers more easily and *ex post* (i.e., after the crime is committed) by reporting detected crimes and by cooperating with the authorities to investigate the crime (Arlen 1994; Arlen and Kraakman 1997).

In sum, corporate crimes are committed in presence of an agency relationship between the firm and the employee and optimal deterrence needs to take into account that the firm has a comparative advantage in the collection of information relative to the state. Therefore, optimal

deterrence of corporate crimes requires that the state not only has to invest optimally in enforcement and to impose the optimal sanctions on individuals but also has to provide the optimal incentives to firms to undertake policies of prevention and policies. In other words, corporate liability is essential (Arlen and Kraakman 1997; Arlen 2012). Of course, there are many factors affecting the decision of the firm to prevent crimes for a given structure of corporate liability; for example, variables such as the firm's size, its productivity, and its market power as well as the profitability of illegal activities end up being important determinants of whether (and how) the firm prevents manager wrongdoings (Polidori and Teobaldelli 2016).

Individual liability in presence of corporate liability, and therefore the joint liability, is necessary because under various circumstances, the firm may be unable (or find it not optimal) to impose the optimal level of sanctions on employees. One of these situations, especially relevant for closely held firms and smaller publicly held firms, is when the firm has insufficient asset to pay the optimal corporate sanction. In this case the firm may find it optimal to impose a suboptimal level of sanctions to the employees (Kornhauser 1982; Kraakman 1984; Sykes 1984); moreover, pure corporate liability is likely to create other kind of distortions as managers have the incentive to keep the firm undercapitalized. In larger publicly held firms, individual liability is necessary because of the agency problems characterizing these firms where there is a separation of ownership and control. Large corporations may be unable to impose the optimal level on sanctions to their employee because managers can often influence the decisions of the board of directors, directly or by controlling the information available to the board (Arlen and Carney 1992). In all these cases, individual liability ensures that the state can impose larger (monetary and nonmonetary) sanctions than the firm to the employee responsible of the wrongdoer and improve social welfare (Segerson and Tietenberg 1992; Polinsky and Shavell 1993).

Vicarious Versus Duty-Based Liability

The optimal structure of corporate liability requires that the firm deters crime by adopting measures of crime prevention and by undertaking policing activities at the optimal level. This result cannot be obtained under vicarious liability because the activities of prevention and policing generate a liability enhancement effect, namely, an increase of the firm's expected liability for the crimes that the firm does not deter (or that cannot be deterred), which reduces the firm's incentive to pursue such policies. Indeed, under strict liability, the activities of crime prevention (such as the adoption of monitoring programs) increase the probability that crimes will be detected and the firm sanctioned (Arlen 1994); similarly, the ex post policing (self-reporting and cooperation with the authorities) may reduce crime ex ante, but it increases the probability that committed crimes are sanctioned so that the net effect on the firm's expected sanctions may well be positive (Arlen and Kraakman 1997).

While strict corporate liability may provide to the firm too little incentive for measures of prevention and policing, a multitiered duty-based sanction regime may overcome this problem and induce the firm to monitor and police optimally (Arlen 2012). More precisely, the state should induce the firm to:

- (i) Employ crime-preventing measures by imposing a penalty if it has not monitored optimally (e.g., it has not adopted appropriate monitoring programs); the penalty can be avoided by the firm choosing the optimal monitoring.
- (ii) Engage in ex post policing by imposing additional sanctions that can be avoided if the firm self-report the crime and fully cooperate with the authorities to convict the wrongdoers.

Although some authors (e.g., Weissmann 2007) argue that the firm should not be liable if it prevents and police optimally, others argue that the optimal structure of corporate liability also requires that the state imposes a residual strict corporate liability. This residual liability (that

should be civil, not criminal) is required to eventually impose additional sanctions equalizing the firm's expected sanction, net of the costs beared by the firm through market sanction, to the total social cost of crime (Polinsky and Shavell 1993; Shavell 1997; Arlen and Kraakman 1997). The main market sanction is generally represented by the reputational penalties which reflect the greater difficulty of *criminal* firms in contracting with other parties on favorable terms; the market's anticipation that the firm will produce lower profits – owing to either higher costs or lower revenues – may lead to a reduction in the firm's value (Klein and Leffler 1981; Karpoff and Lott 1993).

The works on reputational penalties suggest there are large variations in the size of these penalties depending on the type of crime. In particular, parties contracting with the firm will react negatively to news that the firm committed crimes if those crimes (such as fraud) harm them or other contracting parties; however, theory also suggests that firms should not be punished by the market for crimes committed by noncontracting third parties as may occur in cases of regulatory or environmental violation (Karpoff and Lott 1993; Alexander 1999). The available empirical evidence provides support to these theoretical arguments (Karpoff et al. 2005, 2017). Higher reputational penalties clearly provide an incentive to the firm to prevent crime, and this effect is likely to be larger when the firm operates in more competitive markets (Polidori and Teobaldelli 2016).

Civil Versus Criminal Corporate Liability

The doctrine of corporate criminal liability is controversial for various reasons. One of these reasons is the conceptual difficulty of imposing criminal liability on a corporation, an artificial and collective entity, given that criminal liability requires a culpable mental state or *mens rea* (Alexander 1999; Hamdani and Klement 2008). At the same time, one of the main advantages of the criminal law system, that is, the sanction of incarceration as a punishment, is unavailable

against enterprises as these are not physical entities and, therefore, cannot be imprisoned (Khanna 1996; Fischel and Sykes 1996). This consideration leads to the second critical issue concerning the rationale for imposing criminal liability on a corporation given that the criminal and civil corporate sanctions are quite similar in nature (as they are mainly monetary) and can have the same magnitude.

Despite the similarity between corporate criminal and civil liability, they differ along some dimensions. First, there are important procedural differences; criminal cases require a higher burden of proof for the conviction of the wrongdoers, but the authorities have more powerful tools of investigation. Second, whereas corporate criminal and civil liability share the form of monetary sanctions, those levied for criminal behavior are much stronger and can lead to enormous collateral sanctions (e.g., debarment and de-licensing); furthermore, criminal sanctions are not only higher than civil penalties but can also be imposed in addition to them. Third, criminal sanctions are generally associated with much larger reputational losses than civil penalties.

The key distinctive feature of criminal liability of imposing larger sanctions on convicted firms may allow the state to structure a multitiered duty-based liability that induces the firm to prevent and police optimally more efficiently than a pure civil liability regime. However, criminal penalties may also create the risk of overdeterrence of crimes which in case of corporate crimes may have important negative effects on social welfare. Therefore, a well-structured duty-based liability is essential to produce socially efficient outcomes.

Corporate Criminal Liability in the United States and in Europe

Corporate criminal enforcement in the United States is characterized by joint individual and corporate criminal liability for business crimes and firms are formally subject to a strict *de jure* liability regime for employees' crimes. However, criminal liability is *de facto* duty-based as firms can avoid indictment if they self-report

wrongdoing and cooperate with government authorities to convict individual wrongdoers. This Department of Justice policy was initiated in 1999 by Eric Holder (then the US Deputy Attorney General) who issued guidelines to federal prosecutors on when firms should be indicted for employees' crimes committed during the scope of their employment. The Holder memo encouraged prosecutors not to indict firms for employee crimes if they adopted compliance programs, self-reported the wrongdoings, and fully cooperated with the federal authorities. In the existing regime firms that report and cooperate are still subject to some form of expected monetary sanction; prosecutors impose both monetary and nonmonetary sanctions by means of deferred prosecution and non-prosecution agreements (DPAs and NPAs, respectively). These agreements often impose also nonmonetary performance mandates on the firm; for example, they may require the firm to adopt prosecutor-approved compliance programs or to appoint an outside corporate monitor who reports to federal authorities (Garrett 2007; Arlen and Kahan 2017). Arlen (2012, p. 152) concludes that US enforcement practice is consistent with the optimal level of corporate liability.

Even though formal conditions governing leniency are quite broad and could apply to all firms, large and publicly held firms typically avoid formal conviction (by way of DPAs and NPAs), whereas small owner-managed, closely held firms are usually convicted (Arlen 2012). The most likely reason for this difference is that cooperation with prosecutors is the key element determining whether prosecution is avoided or not, and closely held firms tend to cooperate less because they are inclined to protect their owner/managers. In contrast, in large publicly held firms, the owners are rarely directly involved in day-to-day management, and decisions to cooperate with the authorities are often delegated to outside directors who have strong incentives to cooperate in return for leniency.

At the European Union level, after a number of harmonization efforts, there is a general trend in almost all the Member States toward the adoption of corporate criminal liability regimes. With the

exception of the UK and the Netherlands, where there is a long history of recognizing corporate criminal liability, the concept of corporate criminal liability is relatively recent in Europe. The first European country that introduced corporate criminal liability was France in 1994, followed by Finland in 1995, Denmark in 1996, Belgium in 1999, Italy in 2001, Poland in 2003, Austria in 2005, Romania in 2006, Portugal in 2007, Luxembourg and Spain in 2010, Czech Republic in 2012, and Slovakia in 2016. Germany represents a remarkable exception, since corporate criminal liability is considered incompatible with the essence of German criminal law based on the notion of individual culpability; however, in late 2013, the Department of Justice of North Rhine-Westphalia presented a first draft of a Corporate Penal Code, although this remained under discussion (Clifford Chance 2016).

Despite harmonization efforts, relevant differences still exist between the Member States, due to different legal tradition and, also, to different approaches followed in the implementation of the proposed EU legislation. While most countries applied criminal procedure rules, others have imposed quasi-criminal administrative liability regimes. In some jurisdictions, the category of employees which can activate corporate liability is restricted to those with management responsibilities, and the employee's misconduct must be done in the interests of or for the benefit of the company. Similarly to the USA, also at the EU level, a common feature characterizing the legal framework is the possibility for the corporation to have a reduction of the potential penalties in case of adoption of adequate compliance systems to prevent the crime and of cooperation with the authorities. While penalties vary across countries, many of them complement the standard monetary sanctions with nonmonetary sanctions such as judicial supervision of a corporation's affairs, exclusion from public procurement tenders, limitations on the use of checks and credit, confiscation of assets gained from the offence, posting of notices throughout the media, publication of the judgment in a registry, and even the dissolution of the corporation, in the most severe cases (Sun Beale and Safwat 2004).

References

- Alexander CR (1999) On the nature of reputational penalty for corporate crime: evidence. *J Law Econ* 42:489–526
- Arlen JH (1994) The potentially perverse effects of corporate criminal liability. *J Leg Stud* 23(2):833–867
- Arlen JH (2012) Corporate criminal liability. Theory and evidence. In: Harel A, Hylton KN (eds) *Research handbook on the economics of criminal law*. Edward Elgar, Northampton, Massachusetts, USA
- Arlen JH, Carney WJ (1992) Vicarious liability for fraud on securities markets: theory and evidence. *Univ Ill Law Rev* 1992:691–740
- Arlen JH, Kahan M (2017) Corporate governance regulation through nonprosecution. *Univ Chic Law Rev* 84 (1):323–387
- Arlen JH, Kraakman R (1997) Controlling corporate misconduct: an analysis of corporate liability regimes. *N Y Univ Law Rev* 72(4):687–779
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Clifford Chance (2016) Corporate criminal liability. Available at www.cliffordchance.com
- Conley JM, O'Barr WM (1997) Crimes and custom in corporate society: a cultural perspective on corporate misconduct. *Law Contemp Probl* 60:5–22
- Fischel DR, Sykes AO (1996) Corporate crime. *J Leg Stud* 25:319–349
- Garrett BL (2007) Structural reform prosecution. *Va Law Rev* 93(4):853–957
- Hamdani A, Klement A (2008) Corporate crime and deterrence. *Stanford Law Rev* 61:271–310
- Karpoff JM, Lott JR Jr (1993) The reputational penalty firms bear from committing fraud. *J Law Econ* 36(2):757–802
- Karpoff JM, Lott JR Jr, Wehrly EW (2005) The reputational penalties for environmental violations: empirical evidence. *J Law Econ* 48(2):653–675
- Karpoff JM, Lee DS, Martin GS (2017) Foreign Bribery: incentives and enforcement. Available at SSRN: <https://ssrn.com/abstract=1573222>
- Khanna VS (1996) Corporate criminal liability: what purpose does it serve? *Harv Law Rev* 109:1477–1534
- Klein B, Leffler KB (1981) The role of market forces in assuring contractual performance. *J Polit Econ* 89(4):615–641
- Kornhauser LA (1982) An Economic analysis of the choice between enterprise and personal liability for accidents. *Calif Law Rev* 70(6):1345–1390
- Kraakman RH (1984) Corporate liability strategies and the costs of legal controls. *Yale Law J* 93(5):857–898
- Polidori P, Teobaldelli D (2016) Corporate criminal liability and optimal firm behavior: internal monitoring versus managerial incentives. *Eur J Law Econ* 1–34. <https://doi.org/10.1007/s10657-016-9527-2>
- Polinsky AM, Shavell S (1993) Should employee be subject to fines and imprisonment given the existence of corporate liability. *Int Rev Law Econ* 13:239–257
- Polinsky AM, Shavell S (2000) The economic theory of public enforcement law. *J Econ Lit* 38(1):45–76

- Segerson K, Tietenberg T (1992) The structure of penalties in environmental enforcement: an economic analysis. *J Environ Econ Manag* 23:179–200
- Shavell S (1997) The optimal level of corporate liability given the limited ability of corporations to penalize their employees. *Int Rev Law Econ* 17:203–213
- Sun Beale S, Safwat A (2004) What developments in Western Europe tell us about American critiques of corporate criminal liability. *Buffalo Crim Law Rev* 8:89–163
- Sykes AO (1984) The economics of vicarious liability. *Yale Law J* 93(7):1231–1280
- Tyler TR, Blader SL (2005) Can businesses effectively regulate employee conduct? The antecedents of rule following in work settings. *Acad Manage J* 48(6):1143–1158
- Weissmann A (2007) A new approach to corporate criminal liability. *Am Crim Law Rev* 44:1319–1342

Corporate Liability

► Organizational Liability

Corporate Social Responsibility

► Codes of Conduct

Corruption

Maurizio Lisciandra
Department of Economics, University of
Messina, Messina, Italy

Definition

Corruption can be generally intended as an (illicit) exchange between a member of an organization and another subject at the expense of either the organization itself or the rights of others, in which acts of power in contrast with official duty are exchanged for personal advantage. This general definition is usually narrowed to consider corruption as an act of misuse of public power for private

profit against the common good. *Sensu stricto*, bribery is considered the only actual form of corruption. It consists of promising, offering, or giving, as well as soliciting or accepting, a corrupt exchange between some utility (e.g., kickbacks, gratuities, sweeteners) and the actions of individuals (e.g., bureaucrats, politicians, employees in private enterprises) in charge of a legal or public duty. Some habits, such as tipping, gift-giving, and patronage, may not be considered corruption *per se*, but they are corruption-like situations when they are intentionally conceived to induce someone to act dishonestly. Finally, according to many countries' specific legal standards, certain transfers that are similar to bribes are not considered illicit payments, as in the case with lobbying, campaign contributions, and postretirement offers to politicians or public officials. In these circumstances, the elements of a corrupt exchange are all in place apart from the difficulty of identifying the abuse of the official duty and the distortion of market competition.

Introduction

The phenomenon of corruption was originally a matter of investigation in law and criminology studies. However, corruption is not just a legal or ethical issue but also has important economic implications. Starting with the seminal work by Rose-Ackerman (1978) and becoming a well-established field in the economic literature after the 1980s, the advances in corruption studies have profited from the set of tools and results of the theoretical and empirical economic analysis. The general outcome is that corruption affects both distribution and efficiency and, by its nature, restricts market competition. General discussions and surveys on the advances of corruption studies from an economic perspective can be found in Jain (2001), Tanzi (2002), Svensson (2005), Lambsdorff (2007), Lisciandra (2014), and Aidt (2003, 2016), among many.

Corruption is highly context-dependent, as social norms, culture, institutional setting, and the stage of development affect corruption; however, there are also common patterns in human

behavior that explain what generally causes corruption. Corruption also feeds back on the institutional and economic dynamics to such an extent that consequences and causes are confounded by the complexity of all interactions. These issues are presented in this brief essay, which also provides a conceptual framework of corrupt behavior according to the widely used principal-agent model and a summary of the empirical evidence on the determinants and effects of corruption. However, before moving on to the contributions of the existing literature, it is important to address one important topic which regards the measurement of corruption and is closely intertwined with its definition and the empirical analysis.

Measurement of Corruption

Corruption is a multifaceted phenomenon, hard to delimit and measure. However, many institutions have tried to come up with some measurement that could capture its diverse dimensions within a country, a region, or some economic activity. Currently, data on corruption are abundant and are mainly of two types: objective and subjective.

Objective measures are typically reported crimes, court cases, or convictions of bribery and other corruption-related crimes. Unfortunately, objective data have several shortcomings. They do not necessarily measure the actual level of corruption, but may capture instead the quality of investigative, prosecution, and judiciary departments. Sometimes, it may even be the case that low figures of corruption offenses hide a high pervasiveness of the phenomenon. In addition, this type of data is reliable for comparison only if law enforcement and institutional setup are the same in time and space.

Subjective measures are constructed upon the experience or perception of representative samples of the population or field's experts. In particular, some measures are the results of polls of the general public, while other measures use the information available from expert assessment or a combination of factual data such as laws, regulations, institutional data, and on-site visits. Regarding the weaknesses of subjective measures,

especially perception indices, see an extensive analysis in Lambsdorff (2005).

There currently exist numerous cross-national perception measures of corruption. The most widely used measure is the Corruption Perception Index (CPI). This composite index measures perception of corruption in the public sector (i.e., administrative and political corruption) and combines surveys and assessments of corruption from independent institutions devoted to governance and business climate analysis. It is an "index of indices" that was originally launched in 1995 and currently covers 176 countries. Notice, however, that the scores of each country are not comparable over time due to the frequent updates in the methodologies and the inclusion of new data sources.

The International Country Risk Guide Index for corruption is another measure that assesses corruption within the political system in terms of bribery but also secret party funding, excessive patronage, nepotism, and suspicious ties between politics and business. It was originally created in 1980 and derives from the collection of political information that is eventually converted into risk scores. It is issued on a monthly basis and now covers 140 countries. This index is also used to construct the abovementioned CPI.

Another measure of cross-country corruption is the Control of Corruption Index from the Worldwide Governance Indicators that captures the perceptions about petty and grand forms of corruption through a large set of sources such as surveys of firms and households and subjective assessments of several commercial business information providers, as well as many other non-governmental and governmental organizations. It covers over 200 countries starting from 1996.

There are further subjective cross-country indicators that capture the multidimensional phenomenon of corruption in many of its forms. For a more comprehensive overview of measurements and indicators, see Kaufmann et al. (2007) and Malito (2014). In general, all perception measures appear highly correlated, while, as observed by Mocan (2008), experience-based measures may not be correlated with perception measures due to common perception biases or interviewer biases.

The distinction between objective and subjective measures has important implications in the empirical investigation. Due to the extreme difficulty of finding objective measures that can be comparable across countries (see Escresa and Picci 2015, for a first attempt), empirical analyses adopt subjective measures for cross-country investigations. By contrast, within-country investigations can use both subjective and objective measures. In addition, compared to cross-country investigations, within-country analyses can considerably reduce the institutional differences existing across countries (e.g., administrative controls, investigative departments, criminal laws) that can eventually explain most of the variability in the subjective corruption measures across countries.

Another approach to measurement is through indirect measures of corruption. This is especially the case of corruption in public procurements. Many signals can hide the presence of corrupt transactions, such as delays of work completion, overrun costs, and bad quality of goods and services. For a large overview on this type of measure, see OECD (2007).

Finally, more recently, corruption has also been measured through laboratory and field experiments or so-called natural experiments. These are controlled situations that severely reduce the problem of the identification of causation which exists with both subjective and objective data. Experimental data are typically more suitable to investigate the psychological determinants of corrupt behavior. For a recent survey on the pros and cons of experimental data on corruption and the relevant results, see Lambsdorff and Schulze (2015).

Microeconomic Theory of Corruption

Starting from the 1960s with the analysis of criminal behavior, and exploring from the 1970s onward corrupt behavior in particular, economic theory has provided an important conceptual framework that is useful to analyze many issues involving corruption and its implications.

According to the Beckerian seminal contribution on crime modeling, criminals behave as rational economic agents with a specific risk attitude and decide to breach the law if the utility from breaching is higher than the utility from compliance. Corruption is a two-sided criminal endeavor that follows a similar paradigm. In particular, corruption deterrence would depend on (i) the probability of detection and punishment and (ii) the severity of punishment. These components are rather complementary, to such an extent that a very low level of one of the two makes deterrence extremely weak. The expected cost of committing corruption also depends on other components such as (iii) the psychological cost to infringe some moral or social norms and (iv) the opportunity cost to perform legal activities.

This analytical contribution can be associated with the theoretical analysis of another strand of literature that is widely used in contract theory, the so-called principal-agent model (see Klitgaard 1988). In fact, modern societies are characterized by structured and hierarchical organizations with several delegatory relationships. This type of relationship arises with the advent of division of labor and market exchange. Opportunistic behavior that gives rise to corruption is embedded in many of these delegatory relationships, in which a delegate (agent) exploits the informative advantage with respect to the delegator (principal). In other words, once delegated, the agent is endowed with discretionary power that can be used at the expense of the principal. A typical incident of corruption is when the principal delegates the agent to assign a reward (e.g., a public procurement) or a punishment (e.g., tax collection) to a third subject in the principal's interest, but the agent exploits the informative advantage to his/her own interest, for instance, by pocketing a bribe.

There are two types of informative advantage: adverse selection and moral hazard. The former describes a situation in which the principal does not know some characteristics of the agents (e.g., level of morality), while the latter refers to a situation in which the principal may not perceive correct information on the actions the agent should perform in the principal's interest. Thus,

the principal faces two problems: how to select the agents who are best endowed with morality (i.e., solution to adverse selection) and how to design contract schemes inducing agents to behave honestly rather than dishonestly (i.e., solution to moral hazard).

A solution to the first problem would require that (i) moral virtues should be distributed non-uniformly and that (ii) hiring the best agents should not be very expensive. Unfortunately, moral virtues are not easily observable, and they may not remain constant over time. Nonetheless, there are some indirect signals, such as personal wealth, willingness to work for free, and personal history. All these signals are rather imperfect, especially the first one, and should be used carefully. Consider the problem of adverse selection of political representatives, which is very common in any democracy and gives rise to serious corrupt episodes. Many signals coming from electoral campaigns turn out to be flawed to such an extent that the electoral body (principal) oftentimes takes the bad signals on the morality levels of the candidates (agents) as good ones. In the long run, the general decadence of moral values in politics implies that only the most dishonest individuals would run for elections, while honest individuals would tend to decline any candidacy because they are not willing to reach compromises or to be confused with the dishonest ones. In sum, the worst ones crowd out the good ones. Of course, this is a context of multiple principals in which principals' objectives can be the most diverse, and sometimes a good percentage of the electoral body is not even interested in the moral virtues of candidates. This would eventually exacerbate the problem of adverse selection.

The principal can use the "carrot and stick" method to face moral hazards. The carrot works by aligning the agent's interests to the principal's ones. This can mainly be done, where possible, by contract schemes such as pay-for-performance. The stick is more flexible and aims to increase the expected disutility from sanctions. This can be achieved by increasing the probability of being detected (e.g., monitoring, contrast of interests between briber and public official), the probability of being sanctioned (e.g., investing in the

judiciary, extending the statute of limitation), and the severity of sanctions (e.g., easing firing, confiscation, non-electability).

Increasing the community's moral virtues would be another policy in helping to reduce the occurrence of both types of informative advantage. This would mainly increase the psychological cost to breach the law. This is achieved by reducing deprivation, improving law enforcement, introducing rehabilitative punishment, and education around civic-mindedness. In other words, policies that incentivize producers of moral virtues (e.g., the criminal justice system, the education system, volunteering) and discourage producers of dishonesty (e.g., organized crime) would be beneficial against corruption.

Therefore, policies should aim to increase the expected disutility from corruption. For instance, an efficiency wage-like policy would increase the expected disutility from punishment to public officials. This may also increase the potential bribes solicited by public officials, thereby discouraging bribers. However, any policy comes at a cost, which sometimes can be prohibitive. Other policies modify the environment in which corruption takes place, such as (i) reducing public officials' discretionary power by narrowing terms and conditions of their activities, (ii) introducing forms of competition among officials such that more agencies can provide substitutable benefits, (iii) when possible deregulating public purchasing of goods and services, and (iv) simplifying taxation.

Another strand of economic literature that explored corrupt exchange is game theory (e.g., Macrae 1982; Dabla-Norris 2002). Corruption requires an agreement, at least between two parties, and involves strategic interaction between them since they can each cooperate or defect. Cooperation is more likely if defection can be punished. For instance, the briber can punish the uncooperative public officials by excluding them from subsequent corrupt exchanges, hindering their careers, or even encouraging their firing. Evolutionary game theory finds that if corruption is originally confined to small groups that cooperatively sustain the corrupt exchange over time, then corruption is likely to spread to the rest of that

society. In other words, corruption is a sort of disease that is fed by cooperative behavior.

Causes and Effects of Corruption

Corruption is a phenomenon in which causes and effects are closely interdependent. This makes the causation analysis extremely difficult, because the effects of corruption very often feed back into what we expect to be the sources of corruption to such an extent that it is extremely difficult to ascertain which is the cause and which is the effect. In econometrics, this situation is called simultaneity, in which dependent and explanatory variables are jointly determined. For instance, Paldam (2002) draws attention to the seesaw dynamics in which corruption and economic growth feed on each other. For instance, there is a strong negative correlation between country corruption indices and per-capita GDP or per-capita growth rates, but it is not always possible to determine whether corruption generates poverty or whether it is the poverty which is causing corruption. As a consequence, sometimes empirical evidence on the cause-effect relationships is still not conclusive. Nonetheless, some evidence appears to support the strength of a specific causal association. The following empirical evidence, as well as the results from theoretical investigations, presents the current state of the art of the cause-effect relationships of corruption.

Causes of Corruption

A precise and well-defined survey on the causes of corruption can be found in Aidt (2011). Two main factors are considered responsible for corruption and corrupt behavior: institutional or cultural factors and economic drives. On the one hand, weak political and bureaucratic institutions and poor moral virtues encourage the abuse of discretionary power, and on the other hand, economic conditions foster rent extraction. The former are very long-term factors to modify, while the latter can be altered in a shorter period.

Cultural or institutional traditions significantly affect the level of perceived corruption. For

instance, protestant religion is a robust predictor of lower perceived corruption (Treisman 2000). Former British colonies also have significantly lower perceived corruption, which is probably due to the presence of common law legal systems. Overregulation and an excess of bureaucracy are found in cultures endowed by excessive social capital, to such an extent that economic activities eventually rely on relationships and connections. Increasing levels of regulation and bureaucracy are positively associated with perceived corruption and may be part of a deliberate strategy of public officials to increase the willingness to pay of private citizens coming across public administration (Schleifer and Vishny 1999). This is a typical example of reverse causality.

Economic factors feed back on corruption. Economic development increases the delegation process of competences and agencies but also the production of goods and services on behalf of the government. A greater involvement of government in the market economy through higher public spending may increase rents as well as discretionary power about rules and resource allocation. The growth of international trade has provided big corporations with the opportunity to explore new markets and to obtain large orders. However, this expansion has sometimes occurred at the price of large bribes paid to officials and politicians of foreign countries, especially in developing economies (Tanzi 2002).

Corruption may also originate from other causes, such as low press freedom, government involvement in promoting industrial policy, widespread poverty, unfair recruitment and promotion procedures, low wages of public officials, and a high tax burden. Finally, endemic corruption is a cause of further corruption because, where corruption is more prevalent, detection and punishment of corrupt episodes is harder, but also the importance of social stigma and loss of reputation decreases (Andvig and Moene 1990; Soares 2004). This is why the level of corruption appears path-dependent.

Effects of Corruption

The main research question on the effects of corruption is whether corruption turns out to be sand

or grease in the complex economic mechanisms of modern economies. This is a controversial question, which has so far produced no definitive answer.

In principle, corruption may enable individuals to avoid bureaucratic delays in order to smooth and speed transactions. According to this line of thinking, corruption is seen as a second-best solution vis-à-vis the slow and inefficient bureaucracy. Especially in the early stages of economic development, inefficient bureaucracy hinders economic growth that could be smoothed by corrupt practices. In the same way, corruption may introduce competition for scarce governmental resources. For example, corrupt practices may minimize the waiting costs for those who place more value on time or could also facilitate firms' entry into highly regulated economies. However, it must be noted that all the grease arguments have undergone several criticisms and have been considered flawed both conceptually and empirically (e.g., see Kaufmann 1997).

The sand argument seems to be corroborated by more robust evidence. As noted in many studies, corruption acts as an uncertainty- and cost-increasing factor. In a general perspective, the main negative consequence of corruption is the impediment to economic growth, especially for those countries with an unreliable institutional framework. Corruption has a negative impact on economic growth through several channels. In particular, it negatively affects both private and public investments. For instance, it discourages investments by potential innovators because established firms are favored in exchange for bribes. Similarly, corruption undermines foreign direct investments, because it acts as a sort of additional tax on business. It weakens the growth-enhancing effects of public investments because corrupt politicians and public officials may tend to divert public resources toward highly profitable investments such as large-scale and high-cost construction projects, which are subject to low rates of return, rather than small-scale and decentralized projects. Corruption also hampers economic growth by distorting individual incentives. It hinders private investments in education by diverting

individuals' efforts toward rent-seeking activities and induces bureaucrats to inflate regulations and slow down bureaucratic processes in order to receive bribes from their public administration customers. As seen above, the latter consequence is also a cause of further corruption.

Corruption has been especially investigated for many other undesirable distortions that also feed back on economic growth. The list of negative consequences is very large and cannot be exhausted in this brief overview. For extensive surveys on this topic, see, among many, Lambsdorff (1999) and Pellegrini (2011). Nonetheless, it is worth mentioning additional negative effects, such as the growth of the informal economy, the contraction in international trade, the deterioration of market competition, the increment of income inequality and poverty, and not least the increase in pollution and environmental degradation as well as the weakening of environment policies.

Conclusion

The economic analysis can provide an important analytical framework to explain corrupt behavior and explore its implications. On the one hand, the theoretical approach of the principal-agent model allows us to understand the complex informative set of all subjects involved, their incentives, and the consequent effectiveness of the relevant policies to combat corruption. On the other hand, although measuring corruption is very complicated, an increasingly large number of measures of corruption are available to the public at large as well as to scholars. As a result, econometric analysis has made available an extensive empirical evidence on the cause-effect relationship of corruption, which still appears intricate to disentangle, although a few relevant empirical regularities have been revealed. Of course, this brief entry could only illustrate some of the achievements of such a vast and multifaceted phenomenon as is corruption, which is still far from being fully explored and understood in all its forms and consequences.

Cross-References

- [Administrative Corruption](#)
- [Political Corruption](#)
- [Whistle-Blower Policy](#)

References

- Aidt TS (2003) Economic analysis of corruption: a survey. *Econ J* 113(491):F632–F652
- Aidt TS (2011) The causes of corruption. *CESifo DICE Rep* 9(2):15–19
- Aidt TS (2016) Rent seeking and the economics of corruption. *Constit Polit Econ* 27(2):142–157
- Andvig JC, Moene KO (1990) How corruption may corrupt. *J Econ Behav Organ* 13(1):63–76
- Dabla-Norris E (2002) A game theoretical analysis of corruption in bureaucracies. In: Abed GT, Gupta S (eds) *Governance, corruption, economic performance*. IMF, Washington, DC, pp 111–134
- Escresa L, Picci L (2017) A New Cross-National Measure of Corruption. *The World Bank Economic Review* (2017) 31 (1): 196–219
- Jain AK (2001) Corruption: a review. *J Econ Surv* 15(1):71–121
- Kaufmann D (1997) Corruption: the facts. *Foreign Policy* 107(Summer):114–131
- Kaufmann D, Kraay A, Mastruzzi M (2007) Measuring corruption: myths and realities. *Africa region findings & good practice infobriefs*, no. 273, World Bank, Washington, DC
- Klitgaard R (1988) *Controlling corruption*. University of California Press, Berkeley
- Lambsdorff JG (1999) Corruption in empirical research: a review. *Transparency International*, Berlin
- Lambsdorff JG (2005) *How corruption Affects Economic Development, Corporate Governance und Korruption*. In: Aufderheide D, Dabrowski M (eds) *Wirtschaftsethische und moralökonomische Perspektiven der Bestechung und ihrer Bekämpfung*. Duncker & Humblot, Berlin, pp 11–34
- Lambsdorff JG (2007) *The institutional economics of corruption and reform: theory, evidence and policy*. Cambridge University Press, Cambridge
- Lambsdorff J G, Schulze G (2015) What can we know about corruption?. *Jahrbücher für Nationalökonomie u. Statistik* 235/2
- Lisciandra M (2014) A review of the causes and effects of corruption in the economic analysis. In: Caneppele S, Calderoni F (eds) *Organized crime, corruption, and crime prevention*. Springer, New York, pp 187–195
- Macrae J (1982) Underdevelopment and the economics of corruption: a game theory approach. *World Dev* 10(8):677–687
- Malito D (2014) *Measuring corruption indicators and indices*. EUI working papers, no. 13, Robert Schuman Centre for Advanced Studies, Florence
- Mocan N (2008) What determines corruption? international evidence from microdata. *Econ Inq* 46(4):493–510
- OECD (2007) *Bribery in public procurement. Methods, actors and counter-measures*. OECD Publications, Paris
- Paldam M (2002) The cross-country pattern of corruption: economics, culture and the seesaw. *Eur J Polit Econ* 18(2):215–240
- Pellegrini L (2011) *Corruption, development and the environment*. Springer, Dordrecht
- Rose-Ackerman S (1978) *Corruption: a study in political economy*. Academic, New York
- Schleifer A, Vishny R (1999) *The grabbing hand: government pathologies and their cures*. Harvard University Press, Cambridge, MA
- Soares RR (2004) Crime reporting as a measure of institutional development. *Econ Dev Cult Chang* 52(4):851–871
- Svensson J (2005) Eight questions about corruption. *J Econ Perspect* 19(3):19–42
- Tanzi V (2002) Corruption around the world: causes, consequences, scope and cures. In: Abed GT, Gupta S (eds) *Governance, corruption and economic performance*. IMF, Washington, DC, pp 19–58
- Treisman D (2000) The causes of corruption: a cross-national study. *J Public Econ* 76(3):399–457

Further Reading

- Polinsky AM, Shavell S (2001) Corruption and optimal law enforcement. *J Public Econ* 81(1):1–24
- Shleifer A, Vishny RW (1993) Corruption. *Q J Econ* 108(3):599–617

Cost of Crime

David A. Anderson

Department of Economics, Centre College,
Danville, KY, USA

Abstract

The cost of crime guides society's stance on crime and informs decisions about crime-prevention efforts. While early studies focused on crime rates and the direct cost of crime, the aggregate burden of crime involves a much broader pool of information. Counts of crimes do not indicate the severity of criminal acts or the burden of expenditures to deter crime. Beyond aggregating expenses commonly associated with unlawful activity, a thorough examination of the cost of crime covers such

repercussions as the opportunity cost of victims' and criminals' time, the fear of being victimized, and the cost of private deterrence.

Definition

The direct and indirect costs of crime and its repercussions.

Introduction

The cost of crime influences society's legal, political, and cultural stance toward crime prevention and is part and parcel to the benefits of compliance with legal codes. In the ideal state of compliance, there would be no need for expenditures on crime prevention, no costly repercussions of criminal acts, and no losses due to fear and distrust of others. We will not reach that ideal state, but with knowledge of the full cost of crime, we also know the benefit of eliminating any more realistic fraction of that cost.

Early crime-cost studies focused on particular types of crime, geographical areas, or direct repercussions of crime. The aggregate burden of crime involves a much broader array of direct and indirect costs. The cost of crime includes the opportunity cost of time lost to criminal activities, incarceration, crime prevention, and recovery after victimization. The threat of crime elicits private expenditures on locks, safety lighting, security fences, alarm systems, antivirus software programs, and armored car services. The threat of noncompliance with regulations causes myriad federal agencies to dedicate resources to enforcement. And the implicit psychological and health costs of crime include fear, agony, and the inability to behave as desired.

The largest direct outlays resulting from crime in the United States include annual expenditures of \$119 billion for police protection and \$85 billion for correctional facilities (Kyckelhahn 2011). (This and all figures are in 2014 dollars.) Several less-visible costs are also substantial. For example, in a typical year, US citizens spend \$170 billion worth of time locking and unlocking

doors, the psychic cost of crime-related injuries is \$106 billion (Anderson 2011), and computer security issues cost businesses \$82 billion (FBI 2006).

Estimation Methods

Counts of crimes such as the FBI's *Uniform Crime Reports* (UCR) offer no weights on particular crimes according to their severity. In a period with relatively few acts of arson, for example, society can be worse off than before if the severity of those acts is disproportionately great. From a societal standpoint, what matters most is the extent of damage inflicted by crime, which the UCR does not indicate. Measures of the *cost* of crime convey the severity of crimes – an act of arson that destroys a shed carries a different weight than an act of arson that destroys a shopping center.

In the estimation of crime's cost, approaches with a broad scope capture several types of cost shifting that can stem from crime-prevention efforts. A dual analysis of public and private costs captures the potential for public expenditures to shift the burden of prevention away from individuals and firms. The inclusion of many types of crime captures shifts from one type of crime to another. And the consideration of a broad geographical area accounts for the possibility of crime prevention in one area shifting criminal activity to another.

A basic approach is to tally purchases associated with crime, such as expenditures on deterrence, property replacement, medical care, and criminal justice. Market-based estimates are necessarily incomplete because many of crime's costs, including losses of time, health, and peace of mind, are not fully revealed by purchases.

With the contingent-valuation method, investigators use surveys to estimate values for non-market cost components such as fear and pain. These surveys are vulnerable to bias that can result, for example, from the hypothetical nature of survey questions, the objectives of the interviewer who words the questions, or the self-selection of respondents with strong opinions.

Hedonic methods yield estimates of crime-cost components drawn from crime's effect on prices paid for goods or services. For instance, other things being equal, the difference in home prices in areas with low and high crime rates reflects the burden home buyers feel from the greater prevalence of crime.

Cohen (2010) describes a "bottom-up" approach of piecing together each of crime's cost components. Estimates based on market prices, contingent valuation, and hedonic pricing can all play a role in bottom-up calculations. A more holistic "top-down" approach is based on the public's willingness to pay for reductions in crime as stated in responses to contingent-valuation surveys. Anderson (1999, 2011) essentially conducted bottom-up investigations. Examples of top-down studies include Cohen et al. (2004) and Atkinson et al. (2005). Heaton (2010) surveys methods for estimating the cost of crime.

Elements of the Cost of Crime

Crime-cost elements fall into four general categories:

1. Crime-induced production

Crime leads to the purchase of goods and services that would be obsolete in the absence of crime. With the threat of crime, resources are allocated to the production of security fences, burglar alarms, safety lighting, protective firearms, and electronic surveillance, among other examples of crime-induced production. The growing enormity of crime's burden warrants larger outlays for police, private security personnel, and government agencies that enforce laws. As more criminals are apprehended, expenditures on the criminal justice system and correctional facilities grow. If there were no crime, the resources absorbed by crime-induced production could be used to create gains rather than to avoid losses – \$50 spent on a door lock is \$50 that cannot be spent on groceries. The foregone benefits from such alternatives represent a real cost of crime.
2. The opportunity cost of time spent on crime-related activities

Criminals spend time planning and committing crimes, and many serve time in prison. Crime victims lose work time recovering from physical and emotional harm. Virtually everyone beyond early childhood spends time locking and unlocking doors, securing assets, and looking for lost keys. Time is also spent purchasing and installing locks and other crime-prevention devices and watching out for crime, for example, as members of neighborhood-watch groups.
3. Implicit costs associated with risks to life and health

The psychic costs of violent crime include the fear of being injured or killed and the agony of being victimized. Although the costs associated with risks to life and health are perhaps the most difficult to ascertain, a vast literature is devoted to their estimation. Some direct expenditures on crime prevention are made to address these costs, but preventive measures are limited in their ability to deter crime, so a substantial burden of risks to life and health remains.
4. Transfers from victims to criminals

Fraud, robbery, and theft cause a loss to the victim, but to the extent that the victim's loss is the criminal's gain, there is not a net loss to society. For this reason, it is useful for crime-cost reports to break out the component of the cost that is purely a transfer. That way, readers can consider the cost to victims as well as the net cost to society without the transfer component. The effect of theft on production may also be a wash – the use of stolen goods often substitutes for the purchase of legal goods, while it is likely that the victims will replace what they have lost. Thus, the transfer of stolen goods does not necessitate additional production of similar items. On the other hand, if low prices on stolen merchandise entice some people to buy items they would otherwise forego, some of these transfers may necessitate additional production.

Findings

Comparisons of crime-cost estimates over time suggest a growing burden in the United States, partly due to true increases in the cost of crime and partly due to the inclusion of a broadening scope of crime's repercussions. An early study by the President's Commission on Law Enforcement and Administration of Justice (1967) placed crime's cost at \$158 billion. This estimate includes the direct cost of crimes against persons and property, expenditures on illegal goods and services, and public expenditures on police, criminal justice, corrections, and some types of private prevention.

US News and World Report (1974) estimated a \$428 billion crime burden for the United States. That included some private crime-fighting costs, although no breakdown was given. Collins (1994) updated the *US News and World Report* crime study with a cost estimate of \$1.82 trillion. Collins included the value of shoplifted goods, bribes, kickbacks, embezzlement, and other thefts among the costs of crime. Collins also expanded the scope of crime-cost calculations to include pain and suffering and lost wages.

Anderson (1999) estimated a \$2.5 trillion annual cost of crime in the United States, including transfers of \$897 billion worth of assets from victims to criminals. The cost of lost productivity, crime-related expenses, and diminished quality of life amounted to an estimated \$1.6 trillion. Anderson (2011) estimated a \$3.3 trillion annual cost of crime in the United States, including \$1.6 trillion in transfers, \$674 billion worth of crime-induced production, \$265 billion in opportunity costs, and \$756 billion in implicit costs of life and health. This more recent study reflects the crime-related expenditures of the post-9/11 era of heightened security and adds expenditures on investigation services and locksmiths, for which data were previously unavailable.

Estimates of the cost of crime are now available for many countries. For example, Brand and Price (2000) estimate a \$125 billion annual cost of crime in the United Kingdom, Zhang (2008)

estimates a \$110 billion annual cost of crime in Canada, and Detotto and Vannini (2010) estimate a \$53 billion annual cost of crime in Italy. These findings should not be used for international comparisons due to differences in methodology, scope, and data availability.

Conclusion

Counts of criminal acts do not capture the scale of crimes or the expense of crime prevention and victim recovery. A focus on the cost of crime allows investigators to gauge the enormity of crimes, measure the financial burden of public and private prevention efforts, and incorporate both direct and indirect effects of crime. A thorough assessment of the cost of crime includes the value of victim losses, the cost of crime-induced production, the value of time lost due to crime, and the costs associated with risks to life and health. In the United States, as criminals acquire an estimated \$1.6 trillion worth of assets from their victims in a typical year, they generate an additional \$1.7 trillion worth of lost productivity, crime-related expenses, and diminished quality of life.

References

- Anderson DA (1999) The aggregate burden of crime. *J Law Econ* 42(2):611–642
- Anderson DA (2011) The cost of crime. *Found Trends Microecon* 7(3):209–265
- Atkinson G, Healey A, Mourato S (2005) Valuing the costs of violent crime: a stated preference approach. *Oxf Econ Pap* 57(4):559–585
- Brand S, Price R (2000) The economic and social costs of crime, vol 217, Home office research study. Home Office, London
- Cohen MA (2010) Valuing crime control benefits using stated preference approaches. In: Roman JK, Dunworth T, Marsh K (eds) *Cost-benefit analysis and crime control*. Urban Institute Press, Washington, DC
- Cohen MA, Rust RT, Steen S, Tidd ST (2004) Willingness-to-pay for crime control programs. *Criminology* 42:89–110
- Collins S (1994) Cost of crime: 674 billion. *US News World Rep* 17:40

- Detotto C, Vannini M (2010) Counting the cost of crime in Italy. *Global Crime* 11(4):421–435
- Federal Bureau of Investigation (2006) 2005 FBI computer crime survey. Houston FBI – Cyber Squad, Houston
- Heaton P (2010) Hidden in plain sight: what cost-of-crime research can tell us about investing in police. RAND occasional paper. Document OP-279-ISEC
- Kyckelhahn T (2011) Justice expenditures and employment, FY 1982–2007 statistical tables. U.S. Department of Justice. December, NCJ 236218
- President's Commission on Law Enforcement and Administration of Justice (1967) Crime and its impact – an assessment. GPO, Washington, DC
- U.S. News and World Report (1974) The losing battle against crime in America. 16:30–44
- Zhang T (2008) Costs of crime in Canada, 2008. Department of Justice Canada, Ottawa, rr10-05e

Further Reading

- Anderson DA (2002) The deterrence hypothesis and picking pockets at the pickpockets hanging. *Am Law Econ Rev* 4(2):295–313
- Cohen MA (2005) The costs of crime and justice. Routledge, New York

Cost–Benefit Analysis

Jacopo Torriti¹ and Eka Ikpe²

¹University of Reading, Reading Berkshire, UK

²King's College London, London, UK

Abstract

Over the past 30 years, cost–benefit analysis (CBA) has been applied to various areas of public policies and projects. The aim of this essay is to describe the origins of CBA, classify typologies of costs and benefits, define efficiency under CBA and discuss issues associated with the use of a microeconomic tool in macroeconomic contexts.

Definition

Cost–benefit analysis (CBA) is an economic technique applied to public decision-making that attempts to quantify and compare the economic advantages (benefits) and disadvantages (costs) associated with a particular project or policy for society as a whole.

Introduction

Over the past 30 years, cost–benefit analysis (CBA) has been applied to various areas of public policies and projects. Even research on CBA varies significantly and can be classified into two wide areas of work. On the one hand there are studies which have attempted to define the technical-economic reasons underpinning CBA. On the other hand there are studies which carried out empirical evaluations over the performance of samples of CBA.

From a theoretical point of view, CBA has been seen as a tool to increase the quality of regulation and public policy through welfare economics principles and Pareto efficiency. CBA in theory allows for the improvement of social and environmental conditions based on empirical evidence (Sunstein 2002; Koopmans et al. 1964) while improving market competitiveness (Viscusi et al. 1987).

Empirical studies in the area of law and economics have focused on the choice of discount rate, (Dasgupta 2008; Gollier 2002; Lind 1995; Viscusi 2007), the integration of distributional principles (Adler and Posner 1999), the choice of datasets (Morrall 1986; Hahn and Litan 2005), and the performance of different methodologies for monetizing benefits and costs in cases where a market value does not exist (Sunstein 2004; Viscusi 1988). The latter point is of particular interest given the distance between theory and practice and deserves further reflection.

This entry describes the origins of CBA, classifies typologies of costs and benefits, defines efficiency under CBA, and discusses issues associated with the use of a microeconomic tool in macroeconomic contexts.

Origins of CBA

Dupuit, a French engineer, and Marshall, a British economist, defined some of the formal concepts that are at the foundation of CBA. The Federal Navigation Act of 1936 required that the US Corps of Engineers should carry out projects for the improvement of the waterway system when

the total benefits of a project exceeded the costs. This was initiated by Congress, which ordered agencies to appraise costs and benefits when assessing projects designed for flood control as part of the New Deal.

In the 1950s economists tried to provide a rigorous, consistent set of methods for measuring benefits and costs and deciding whether a project is worthwhile. This mainly consisted in applying compensation tests and distributional weights. However, such measures were considered by several economists as a failure (Adler and Posner 1999). Notwithstanding opposition, in the USA, CBA was increasingly applied in an expanding domain of policy areas, often following the rationale that alternative policy appraisal tools were less efficient (Pearce and Nash 1981).

Following some experiences in Scandinavian countries and Canada, the US Executive Order 12291 of 1981 institutionalized CBA as a consistent method for the appraisal of Government policies and regulations, hence marking the beginning of the CBA era (Posner 2000).

Costs

Each type of legislative change imposes various typologies of costs. Private companies, citizens, and public administration can be subject to an increase in costs. The first significant

classification is with regard to private and societal costs. The former consist of what a citizen or household has to pay in relation to a legislative change. CBA is often used by public administrations as an instrument to measure only certain components of private costs. This is particularly the case when legislative change is expected to have impacts on individual categories of companies.

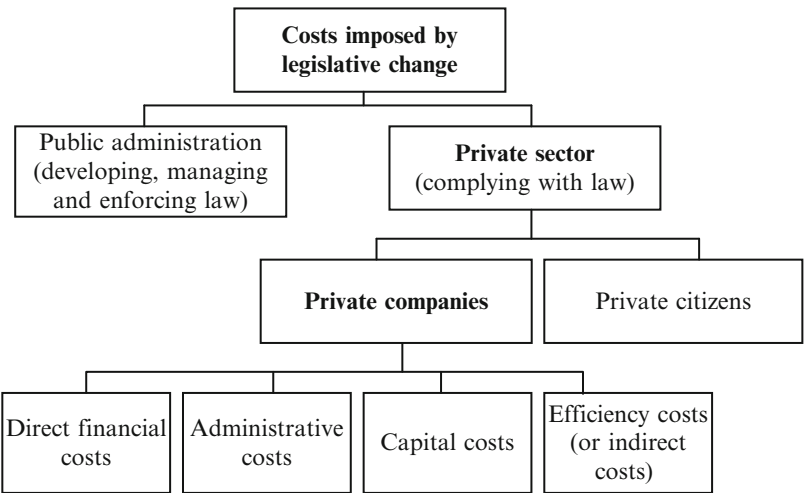
Social costs represent what society as a whole has to pay because of legislative change. They typically include negative externalities and exclude transfer costs among groups of citizens (or companies).

Figure 1 outlines the typologies of costs associated with legislative change. Costs for public administration mean management costs as well as enforcement costs, i.e., costs associated with monitoring and inspections to ensure compliance. On the right of Fig. 1, private costs are divided between costs for private citizens and private companies. The latter are broken down in terms of direct financial costs, administrative costs, capital costs, and efficiency costs.

Benefits

CBA practice suggests that benefits are more problematic to quantify and monetize than costs. The following taxonomy of costs outlines some

Cost–Benefit Analysis,
Fig. 1 Scheme of costs imposed by legislative change



broad categories of benefits ordered from the highest level of monetization and quantification to the lowest:

- Economic benefits valued in the market
- Noneconomic benefits which are not valued in the market that can be quantified and monetized
- Noneconomic benefits that can be quantified but not monetized
- Noneconomic benefits that cannot be quantified

Economic benefits for which a value is provided in the market are undemanding to monetize. An example could come from adding more wind turbines into the energy market. The market benefits are known because it is known both the physical quantity of energy that the extra turbines would provide (i.e., kWh) and the monetary value of the physical quantity (i.e., €/kWh) (Torriti 2010).

The most controversial category of benefits consists of noneconomic benefits which are not valued in the market that can be quantified and monetized. An example of this is reducing health risks. There is no market value for this and neither for saving lives, but the monetary value of the benefit can be seen as a reduction in the risk of dying or catching a disease. Economists have developed four main methods for monetizing non-market values associated with reductions in risk. First, willingness to pay values are based on asking citizens how much they would pay to reduce the likelihood of a specific risk. In practice this is implemented through (i) stated preference surveys, where individuals are asked questions on changes in benefits; (ii) close-ended survey, where respondents are asked whether or not they would be willing to pay a particular amount for reducing risk; and (iii) stochastic payment cards, which offers to respondents a list of prices and associates likelihood matrix describing how likely the respondent would agree to pay the various offered prices. Second, the human capital approach calculates the value of a human life saved, assessing the present value of the worker's earnings over the lifetime. The value is the benefit

associated with reducing loss wages. Third, the cost of illness (or medical costs assessment) method consists of an estimate of the costs to the medical system for treatment due to illness. Fourth, willingness to accept values are based on the wage premiums workers accept for risks. When the wage premium is divided by fatality risk, the result is the value of a statistical life saved.

Examples for the noneconomic benefits that can be quantified but not monetized include the number of fish species saved from extinction. CBA is anthropocentric, and impacts on other animal species are rarely taken into account in monetary terms as part of a policy appraisal, unless this refers specifically to ecological conservation and animal species protection. Nonetheless, handbooks on CBA including the British HM Treasury's (2012) Green Book may contain details about parameters to be used for plant species as part of the policy appraisal.

The category of benefits which cannot be quantified and let alone monetized in a CBA is typically very broad, as it comprises several areas of social benefits. An example is the benefit of improving social justice thanks to a new policy or regulatory change. There might be social indicators which address some of this change, but this is hardly reconciled to a monetary value. In the last US Executive Order 13563 on CBA, it is stated that agencies should take account of "human dignity" and "fairness," values, although these are "difficult or impossible to quantify."

Defining Efficiency Under CBA

CBA is the most comprehensive of a family of economic evaluation techniques that seek to monetize the costs and/or benefits of proposals. Following the classifications in the section above, benefits and costs can be broadly defined as anything that increases human well-being (benefits) or anything that decreases human well-being (costs).

The core efficiency principles of CBA lie in welfare economics (i.e., the branch of economic theory which has investigated the nature of the

policy recommendations that the economist is entitled to make) within the domain of allocative efficiency (Baumol et al. 1977; Perman 2003). Allocative efficiency (i.e., allocation of scarce resources that gives maximum social well-being) is defined via the concept of a “Pareto improvement.” A Pareto improvement is a reallocation of resources (e.g., a decision to develop) that makes at least one individual better off without making anyone worse off. Pareto efficiency/optimality is achieved when it is impossible to make one individual better off without making at least one other individual worse off.

The problem is that if economists restricted their domain of advice to Pareto improvements, they would not be able to advise on much as most decisions involve a trade-off between making someone better off at the expense of making someone else worse off. What could be required is that the individual who gains must compensate the individual who loses, for all of the latter’s loss. If the individual who gains still gains after having paid out the compensation in whatever way (e.g., cash), the move would still be a Pareto improvement. This is not much less restrictive, because actual compensation is rarely paid. As an alternative, economists developed the idea of potential Pareto improvements. The Kaldor compensation test (after Nicholas Kaldor) sanctions a move from one allocation of resources to another, if the winner could compensate the loser and still be better off. In this case, the compensation does not actually have to be paid. Hicks identified a problem with Kaldor’s compensation test – namely, that it could sanction a move from one allocation to another, but it could equally sanction a move in the opposite direction, depending on where the problem starts. Instead Hicks suggested that the loser could compensate the winner for forgoing the move, without being worse off than if the change took place. If no compensation takes place, then the reallocation should be sanctioned.

It took a later paper by Scitovsky (1951) to disentangle the problem. Both rules need to be satisfied, such that a reallocation is desirable if on the one hand the winners could compensate the losers and still be better off and, on the other hand, the losers could not compensate the winners

for the reallocation not occurring and still be as well off as they would have been if it did occur.

Hence, CBA is based on the Kaldor–Hicks efficiency criterion. The benefits should be enough that those that benefit could in theory compensate those that lose out. It is justifiable for society as a whole to make some worse off if this means a greater gain for others. Under Pareto efficiency, an outcome is more efficient if at least one person is made better off and no one is made worse off. This is a stringent way to determine whether or not an outcome improves economic efficiency. However, some believe that in practice, it is almost impossible to take any social action, such as a change in economic policy, without making at least one person worse off (Buchanan 1959). Using Kaldor–Hicks efficiency, an outcome is more efficient if those that are made better off could in theory compensate those that are made worse off, so that a Pareto-improving outcome results. An allocation is defined as “Pareto efficient” or “Pareto optimal” when no further Pareto improvements can be made.

CBA is a tool for judging efficiency in the case where the public sector supply goods or where the policies executed by the public sectors influence the behavior of private sectors and change the allocation of resources (Stiglitz 2000). The concept of efficiency, though, is normally thought on the premise of the market economy. This is particularly controversial when the decision being contemplated involves some cost or benefit, for which there is no market price or which, because of an externality, is not fully reflected in the market price.

Microeconomic Concepts...

As the market economy consists of the spontaneous transaction of goods and services, a characteristic of spontaneous transaction is that there is no individual who loses in the transaction itself. A buyer of goods, who obtains the goods by paying money, is willing to obtain the goods because the amount he or she has to relinquish in exchange for obtaining the goods is in a permissible range. If the amount to be relinquished is

excessive, a person dare not obtain the goods. This assumes that there is an upper limit to the amount of money a buyer is willing to relinquish in exchange for obtaining the goods. In economics, this amount of money is called “willingness to pay” (WTP). No one is willing to buy the goods in the market unless they can be bought at a price lower than WTP. Meanwhile, a seller of the goods, who receives money in exchange for relinquishing the goods, relinquishes them because the amount he or she can get is large enough. A seller will not be willing to relinquish the goods if the amount of money is too small. This assumes that there is a minimum amount of money needed to make a seller willing to relinquish the goods. In economics, this monetary amount is called “willingness to accept” (WTA). No one is willing to sell goods unless they can be sold at a price greater than WTA. In the case where a buyer is the consumer of the goods, WTP for the buyer will depend greatly on the buyer’s subjective appraisal of the goods, in other words, the utility the goods bring. WTA is normally equal to the costs of supplying the goods. The agreement of selling and buying in the market means that a buyer bought at a price lower than WTP and a seller sold at a price greater than WTA. A buyer who bought at a price lower than WTP will have the better subjective utility, and a seller who sold at a price greater than WTA should make a profit from the sale that exceeds the costs. That is to say, market transactions do not fail to bring gain to the parties in a transaction. In a case where everyone makes gain or a person makes gain with no one suffering a loss as a result of change, the change is defined as a “Pareto improvement.” The market transaction is defined as efficient in that it brings about the Pareto improvement.

... Applied to Macroeconomics: The Case of Development Projects

CBA is a method by which this concept of efficiency can be applied to publicly supplied goods as well. Provided that the publicly supplied goods bring utility to people, these people are associated with WTP values for the publicly supplied goods.

The WTP of these people represents the benefit of supplying such goods, whereas WTA is the cost for supplying the goods. Difficulty exists, however, in transferring the efficiency concept for market goods to publicly supplied goods. Mishan (1980), for instance, states that the efficiency concept in the meaning of a Pareto improvement cannot be applied because the goods targeted by CBA are supplied goods of a public nature.

Public goods are publicly supplied because they cannot be adequately supplied through the market. It is the reason why they are defined as “public goods.” When compared with public goods, private goods have the following characteristics. First, no more than one person can use the private goods at the same time because the process of using the goods is a private process. Second, the private goods cannot be used unless a reward is paid. Third, the users can ordinarily decide whether they use the private goods or not and the degree to which they use them. Public goods do not encompass these characteristics at all (or only to some degree). Among these characteristics, there are two relevant points involving the efficiency concept in CBA: that people can use public goods without paying the reward and that they have no choice as to use or nonuse, or as to the degree of usage. Two critical points in terms of difference between private and social CBA are that (i) public suppliers may be concerned with a much broader range of consequences than firms and (ii) public suppliers may not always use market prices in evaluating projects either because the market prices may not exist or because market prices may not represent true marginal social benefits/costs.

The efficiency criteria on which CBAs for public goods are based are relatively different from the WTP–WTA equation explained above because they are supposed to take into account distribution and equity issues. When efficiency conflicts with other values, it is actually impossible to create economic welfare criteria that integrate all values. Mishan (1982) insisted on the idea that only ethical consensus in society could justify the use of efficiency criteria.

An example of this discussion on the impractical use of CBA in the context of the

macroeconomics of public goods comes from development projects. These create both winners and losers. In most of the cases, losers already belong to the poorest and more marginalized members of society. While development projects can bring enormous benefits to society, their costs to the poorest have effects on their health and even their lives (Kanbur 2002). The use of CBA by international organizations for development projects is widespread and often criticized for not considering areas like basic needs approaches, shadow prices, social discount rates, and macroeconomic shocks to public goods (Brent 1998; Devarajan et al. 1997; Kirkpatrick and Weiss 1996). Examples include how nonmarket values associated with water use and human displacement are considered in CBAs for large-scale hydropower projects in developing countries (Mirumachi and Torriti 2012). Some of these challenges are the result of the paucity of data that thereby requires substantial reliance on approximations and wider resource constraints on assessors (ECA SRO-SA 2012).

Economists tried to introduce sensitive weights to compensate the economic estimates of the project with social, health, and environmental factors. CBA can still be employed in order to identify how losers from, e.g., displacement will be economically affected by the change. The method of aggregation, for example, gives a much larger weight to the gains and losses of the poor than those of the rich. Egalitarianism is introduced through the nature of the utility function. In other words, using the method of aggregation, a dollar's loss or gain means more to a poor person than a rich one. The method of aggregation represents a first move in the direction of the compensation principle (Robbins 1932). Countered (Harrod 1938) and improved (Kaldor 1939) by other economists, the idea of the compensation principle is that a policy change could be Pareto improving if it were accompanied by appropriate lump-sum transfers made by tax winners in order to compensate losers. In practice, the complex combination of using a microeconomic technique in a macroeconomic context means that the systematic use of such weights in project appraisal or CBA is rare.

References

- Adler M, Posner E (1999) Rethinking cost-benefit analysis. *Yale Law J* 109:165–247
- Baumol WJ, Bailey EE, Willig RD (1977) Weak invisible hand theorems on the sustainability of multiproduct natural monopoly. *Am Econ Rev* 67:350–365
- Brent RJ (1998) Cost-benefit analysis for developing countries. Edward Elgar, Cheltenham/Northampton, MA
- Buchanan JM (1959) Positive economics, welfare economics, and political economy. *J Law Econ* 2:124
- Dasgupta P (2008) Discounting climate change. *J Risk Uncertain* 37:141–169
- Devarajan S, Squire L, Suthiwart-Narueput S (1997) Beyond rate of return: reorienting project appraisal. *World Bank Res Obs* 12(1):35–46
- Economic Commission for Africa Sub-Regional Office for Southern Africa (ECA SRO-SA) (2012) Cost-benefit analysis for regional infrastructure in water and power sectors in Southern Africa. United Nations Economic Commission for Africa, Addis Ababa
- Gollier C (2002) Discounting an uncertain future. *J Public Econ* 85:149–166
- Hahn R, Litan R (2005) Counting regulatory benefits and costs: lessons for the US and Europe. *J Int Econ Law* 8(2):473–508
- Harrod RF (1938) Scope and method of economics. *Econ J* 48(191):383–412
- Kaldor N (1939) Speculation and economic stability. *Rev Econ Stud* 7(1):1–27
- Kanbur R (2002) Economics, social science and development. *World Dev* 30(3):477–486
- Kirkpatrick CH, Weiss J (eds) (1996) Cost-benefit analysis and project appraisal in developing countries. Edward Elgar, Cheltenham/Brookfield
- Koopmans TC, Diamond PA, Williamson RE (1964) Stationary utility and time perspective. *Econometrica* 32:82–100
- Lind R (1995) Intergenerational equity, discounting, and the role of cost-benefit analysis in evaluating global climate policy. *Energy Policy* 23:379–389
- Mirumachi N, Torriti J (2012) The use of public participation and economic appraisal for public involvement in large-scale hydropower projects: case study of the Nam Theun 2 hydropower project. *Energy Policy* 47:125–132
- Mishan EJ (1980) How valid are economic evaluations of allocative changes? *J Econ Iss* 14:143–161
- Mishan EJ (1982) The new controversy about the rationale of economic evaluation. *J Econ Iss* 16:29–47
- Morrall J (1986) A Review of the record. *Regulation* 2:25–34
- Pearce DW, Nash CA (1981) The social appraisal of projects: a text in cost-benefit analysis. Macmillan, London
- Perman R (ed) (2003) Natural resource and environmental economics. Pearson Education, New York /Harlow
- Posner RA (2000) Cost-benefit analysis: definition, justification, and comment on conference papers. *J Leg Stud* 29(S2):1153–1177

- Robbins L (1932) The nature and significance of economic science. Macmillan, London
- Scitovsky T (1951) The state of welfare economics. *Am Econ Rev* 41:303–315
- Stiglitz JE (2000) Capital market liberalization, economic growth, and instability. *World Dev* 28(6):1075–1086
- Sunstein C (2002) The cost-benefit state: the future of regulatory protection. American Bar Association, Chicago
- Sunstein CR (2004) Lives, life-years, and willingness to pay. *Colum L Rev* 104, 205
- Torriti J (2010) Impact assessment and the liberalisation of the EU energy markets: evidence based policy-making or policy based evidence-making? *J Common Mark Stud* 48(4):1065–1081
- Treasury HM (2012) Accounting for environmental impacts: supplementary green book guidance. HM Treasury, London
- Viscusi WK (1988) Irreversible environmental investments with uncertain benefit levels. *J Environ Econ Manag* 15:147–157
- Viscusi WK (2007) Rational discounting for regulatory analysis. *Univ Chic Law Rev* 74:209–246
- Viscusi K, Magat W, Huber J (1987) An investigation of the rationality of consumer valuations of multiple health risks. *Rand J Econ* 18:465

Counterfeit Money

Elena Quercioli¹ and Lones Smith²

¹Department of Economics, College of Business Administration, Central Michigan University, Mt. Pleasant, MI, USA

²Department of Economics, University of Wisconsin, Madison, WI, USA

Abstract

Counterfeit money is the topic of television, movies, and lore but hardly seen by most of us – for only about one in ten thousand notes is found to be counterfeit, annually, in the USA (Judson and Porter 2003). And while the value of globally seized and passed counterfeit American dollars has exceeded \$250 million in recent years, it is a multi-billion-dollar *potential* crime. Here, we distill a recent model of counterfeit money as a massive multiplayer game of deception. Bad guys attempt to pass forged notes onto good guys. Good guys expend effort to verify the notes, in order to avoid potential losses. The

model sheds light on how the counterfeiting rates, the counterfeit quality, and the cost of attention vary in response to changes in the banknote denomination, the technology, and the severity of the legal penalties for counterfeiters.

Synonyms

Deception games; Passed money; Seized money

Introduction

Fiat currency is paper or coin that acquires value by legal imperative. The longstanding problem of counterfeit money strikes at its very foundation, debasing its value and undermining its use in transactions. Williamson (2002) points out that counterfeiting of private banknotes was common before the Civil War but nowadays is quite rare – in fact, only about one in 10,000 notes are counterfeit.

An important distinction must be drawn between *seized* and *passed* counterfeit notes. Seized notes are confiscated before they enter regular circulation, by the police. Passed notes are those fake notes found at a later stage, once they have entered circulation and exchanged hands among unwitting individuals. Passed notes lead to losses by the public, since they must be handed to the police when discovered, and their origins are invariably lost to history.

In the USA, the Secret Service (USSS) investigates the crime of counterfeiting of the dollar. In so doing, it records all seized and passed notes and reports them, annually, to the Congress (US Department of Homeland Security, USSS Annual Report 2012). Since 2002, the European Central Bank (ECB) also reliably records passed and seized notes for the euro, as does the Bank of Canada for the Canadian dollar. Exploring this data, Quercioli and Smith (2014) uncovered some basic facts about counterfeiting, focusing largely on the US dollar.

The data beginning in the 1960s reveal that, initially, the vast majority of all counterfeit notes were seized prior to circulation. But starting

around 1986, this began to change. Currently, the *seized-passed ratio* (the ratio of the number of seized notes and the number of passed notes) has fallen from around 9 to 1/9. But this shift varies by denomination. In particular, the seized-passed ratio increases in the value of the denomination – a pattern that is replicated by the six-denomination Canadian currency.

Next, the *passed rate* (the number of passed notes over the number of circulating genuine notes) rises in the denomination, initially very steeply, and then less so. Notably, for the European currency, the passed rate rises at first and then dramatically plunges at the 500-euro note. Curiously, however, when we look at passed rates for Federal Reserve Banks (referred to as FRBs, hereafter), we notice that they exhibit the opposite pattern: FRBs disproportionately detect previously missed, *low*-denomination notes. Their share of passed notes, instead of rising in the denomination, falls in it – at least until the \$100 bill is reached.

As well, Quercioli and Smith (2014) observe some other, interesting empirical patterns: The manufacture of fake notes also varies by denomination. Specifically, the \$50 and especially the \$100 notes are forged using more sophisticated methods than lesser notes. By contrast, the fraction of fake notes manufactured using “inexpensive” digital means is clearly skewed towards the low denominations (the \$5s, \$10s, and \$20s).

The early literature on counterfeiting was quite small, mainly focused on the money-and-search framework of Kiyotaki and Wright (1989). This general equilibrium theory assumes – in its only margin of explanation – that the price of money (e.g., its purchasing power) adjusts to accommodate supply and demand shocks. Such shocks could, for example, derive from the threat of counterfeit notes circulating in the economy. But that framework has problems accounting for the existence of counterfeiting, let alone explaining its stylized facts. For example, Nosal and Wallace (2007) find no counterfeiting equilibrium when the cost of production of forged notes is high enough. Green and Weber (1996) also find no equilibrium with counterfeiting when agents

observe a fixed signal of the quality of the money, before trading with each other.

The most sophisticated paper in this thread is Williamson (2002). This paper assumes that banknotes can be forged at a cost and detected at an exogenous chance. Also, as Williamson (2002) remarks, discounting of private banknotes was common before the Civil War, when counterfeiting of the “Confederate dollar” ran rampant. However, nowadays, counterfeiting is a relatively rare phenomenon. Bills may be declined in payment. For example, “No \$100 bills accepted” signs abounded in Ontario after a counterfeiting rash in early 2000 (The Globe and Mail 2002).

The critical, missing variable in all previous literature is the attention that individuals pay to their notes. For instance, Quercioli and Smith (2014) observe that after Canada colorized its notes in 1969–1976, passed rates dropped dramatically across all six denominations of the Canadian dollar. So motivated, they pursue instead a “behavioral” explanation for counterfeiting, that we flesh out here. The first element of the story is a grand *cat and mouse* game of deception played amongst good and bad guys. In essence, illegitimate banknote producers deceive unwitting individuals who, in turn, invest effort to avoid deception. They carefully examine the notes acquired before accepting them. Aside from production and distribution costs, counterfeiters court imprisonment and heavy legal fines when ultimately apprehended by the police. Finally, the police can reduce the passage of bad notes, but do so in a mechanical fashion, aided by the verification efforts of innocent individuals.

In the second element of the story, more careful inspection costs more effort, but reduces the chance of accepting bad money. Inevitably, some good guys are deceived. They then proceed to pass on the bad notes to others – the two parties unwitting partners in a larger, collateral game of deception. The moment a fake note is detected, the holder loses it, by law. This *hot potato* game is one of strategic complements: For the better others check for counterfeits, the more everyone else is motivated to check too, to avoid future losses.

All told, there are two interlinked massively multiplayer games: The cat and mouse game pits

counterfeiters against unwitting individuals and the police, and the hot potato game is the collateral battle fought by unwitting individuals, against each other. The two interlinked games are solved in reverse order. Nash equilibrium is the requisite solution concept – namely, all individuals simultaneously optimize, taking as given others' actions. In the cat and mouse game, Nash equilibrium identifies the verification effort of good guys and the quality level of the forged notes. This yields the verification rate. Next, in the hot potato game, the verification rate identifies the counterfeiting rate.

Building a model of counterfeiting is only part of the task at hand. The next question that needs to be answered is: Does the model shed light on the empirical regularities uncovered? To answer this, we sketch an example of the theory developed in Quercioli and Smith (2014). The full model, data set and discussion of the empirical findings can be found in Quercioli and Smith (2014).

An Economic Model of Counterfeiting

Consider an economy where there are two types of anonymous and unrecognizable agents. Both are risk-neutral. One type of agent is bad, while the other is good. The bad agent can choose to enter the market as a counterfeiter of denomination $\Delta > 0$ notes and can choose the quality $q \geq 0$ of fake notes to produce. For simplicity, we assume his expected production is not a choice variable and simply normalize it to 1. If he enters, his expected profits are zero, after accounting for the anticipated *legal penalty* (or monetary equivalent), Λ , when he is arrested. Counterfeiters also expect to have some *money seized*, and in total, this amounts to $S[\Delta]$, in every period. Bad guys try to distribute everything they produce. To be determined is the *counterfeiting rate* κ , namely, the fraction of all notes that are forged and not authentic.

In this example, good guys simply pass on their notes to other good guys every period, in an unmodeled exchange for goods. If it is fake money, then they lose it. So, note by note, they quickly examine each bill they are

handed – expending *effort* $e \geq 0$ in doing so. In turn, this successfully uncovers fake notes with a probability $0 \leq v \leq 1$, the *verification rate*. This probability v reflects the effort e and quality q of counterfeits. We have:

$$e = qv^B$$

where $B \geq 2$.

Thus, the verification rate rises in e and falls in q . The verification rate acts as an implicit price. In equilibrium, knowing the quality, an effort choice is formally equivalent to a verification choice. When handed a note in a transaction, a good guy does not know whether that note is counterfeit or authentic. We thus use an analytic device, allowing the good guy to act *as if* he chooses the verification rate $\hat{v}$ to minimize his losses. Those will occur when he encounters a bad note, he does not recognize it, and he passes it onto a good guy who does. For this, we must introduce a variable κ to capture the *counterfeiting rate*. All such events are independent, and therefore the probability of the joint event is the product of the respective, individual chances of each separate event. His total expected losses in the transaction are:

$$\kappa(1 - \hat{v})v\Delta + q\hat{v}^B$$

including the effort costs of examining the note.

Take the first-order optimality condition in $\hat{v}$, and then impose the equilibrium assumption that all good guys are likewise optimizing, so that $\hat{v} = v$. Then,

$$-\kappa v\Delta + qB v^{B-1} = 0$$

Inverting this expression yields the *counterfeiting rate*, κ :

$$\kappa = \frac{qBv^{B-2}}{\Delta}.$$

Observe that while κ , the total stock of counterfeit notes, is not an observable variable, the passed rate is. The theory of Quercioli and Smith (2014) struggles to tease out the behavior of this

important unobserved counterfeiting rate from the passed rate – for they are similar but do not coincide. On a “per transaction” basis, it amounts to $\pi = v\kappa$. So, substituting for κ in π :

$$\pi = v\kappa = \frac{qBv^{B-1}}{\Delta}$$

The passed rate admits a nice economic interpretation. It represents *the ratio of the marginal verification cost and the value of the denomination*. Counterfeiters choose to come into the market and select the quality of fakes they wish to manufacture (the European Central Bank has adopted the catch phrase “feel-look-tilt” in its campaign for the security features of the euro, where tilt refers to the hologram). To be specific, producing a counterfeit note of quality q costs:

$$c(q) = Cq^A$$

where $A > 1$.

We simply assume that anticipated, future *legal costs* are fixed at $\Lambda > 0$. Counterfeiters choose quality to maximize their *profits*, while perfect market competition drives their future profits to zero:

$$(1 - v)\Delta - Cq^A - \Lambda = 0$$

Let us call this the $\bar{\Pi}$ -locus. In light of the legal penalties, there is a positive minimal counterfeit note that can be profitably counterfeited $\underline{\Delta} = \Lambda > 0$; near such note, the level of verification and quality vanishes. First-order conditions for optimal quality imply what we next call the Q^* -locus:

$$ACq^A - \frac{\Delta v}{B} = 0.$$

Together, the optimality and zero-profit equations yield the induced (Nash)-equilibrium verification rate:

$$v = \bar{v} \left(1 - \frac{\Lambda}{\Delta} \right)$$

where $\bar{v} = AB/(C + AB) < 1$.

While the verification rate improves as the note value rises, it is bounded to be below one. Intuitively, some fake notes will always pass into circulation.

In summary, the equilibrium choice variables are:

$$q = [(1 - \bar{v})(\Delta - \Lambda)/C^2]^{\frac{1}{A}} \text{ and } e = qv^B = C^{-2/A}(1 - \bar{v})^{\frac{1}{A}} \bar{v}^B \Delta^{-B}(\Delta - \Lambda)^{B+\frac{1}{A}}$$

We see here that effort and quality ramp up in the note, but effort progresses faster than quality. This raises the verification rate.

Finally, we substitute q into our earlier expression for the counterfeiting rate and find:

$$\kappa = BC^{-2/A}(1 - \bar{v})\bar{v}^{B-2}\Delta^{1-B+\frac{1}{A}}(\Delta - \Lambda)^{B-2+\frac{1}{A}}$$

Unlike our earlier expression, this formula expresses the counterfeiting rate solely as a function of the exogenously given variables and can be used for predictions.

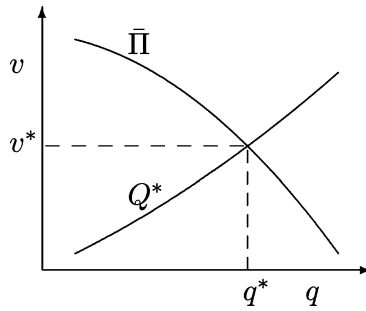
Altogether equilibrium is described by a quadruple of variables (e, q, v, κ) that simultaneously yield equilibrium in the two separate, independent sub-games, linked by the verification and counterfeiting rates. From the cat and mouse equilibrium, drawn on the left hand side of Fig. 1, effort, quality, and thus the verification rate are all determined. Intuitively, one can think of the verification rate as fixing the counterfeiting supply curve K^S – an infinitely elastic curve. (This is so because the homogeneous counterfeiters do not care at all about the counterfeiting rate prevailing in the economy). The “hot-potato” equilibrium then pins down the counterfeiting rate, on the right hand side of Fig. 1, and yields the “derived” demand curve $\kappa = \frac{qBv^{B-2}}{\Delta}$ for counterfeit notes, K^D . It oddly has a positive slope. The reason is that the commodity in question, counterfeit money, is a “bad” and not a “good.” So, the verification rate is the price that needs to be paid to discourage illegitimate, counterfeiting activities.

Notice that this equilibrium is *stable*: if the verification rate is too low, below $\bar{v}$, counterfeiters will find it profitable to enter the market – as detection is not very effective. The counterfeiting rate surges. On the other hand, if good guys

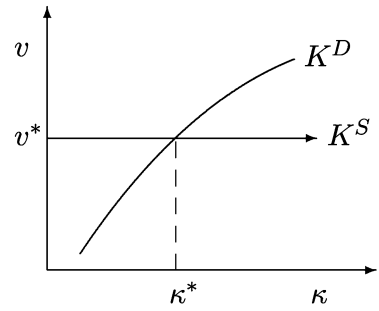
Counterfeit Money,

Fig. 1 Equilibrium in the counterfeiting model

1. Cat and Mouse Game Equilibrium

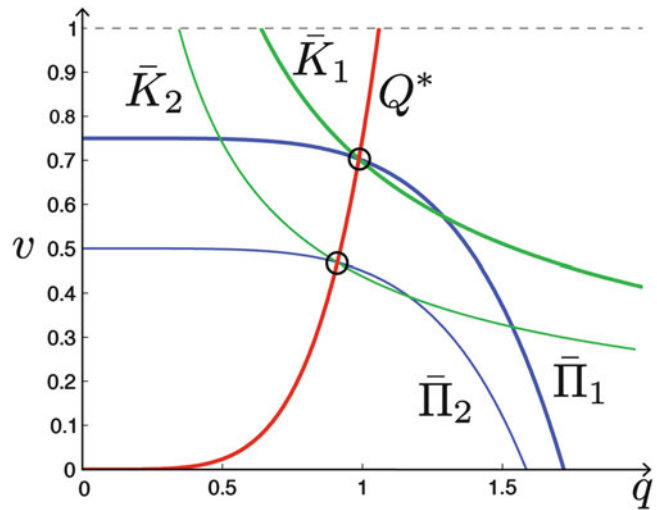


2. Counterfeiting Equilibrium



Counterfeit Money,

Fig. 2 Raising the legal penalty lowers the verification rate and quality



perceive the counterfeiting rate to be below (above) κ , verification ultimately rises (falls). This is so because the good guys realize their errors and readjust their attention level accordingly.

Equilibrium Predictions

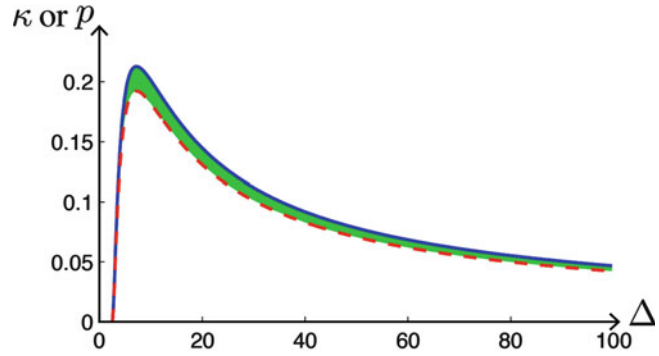
We produce and plot the equilibrium variables. The model exhibits equilibrium feedback. We illustrate one such result of Quercioli and Smith (2014). Suppose the authorities crack down on counterfeiting, thereby rising the (anticipated) legal penalty that counterfeiters court. Then counterfeiters would all exit, unless the cost structure somehow improved. The only variable that can change, here, for a given denomination value, is the effort verifiers expend examining notes. In fact, this must fall so much that – even though counterfeiters produce lower quality notes – the

resulting verification rate is lower (see Fig. 2). We find that greater legal penalties strongly “crowd-out” verification effort.

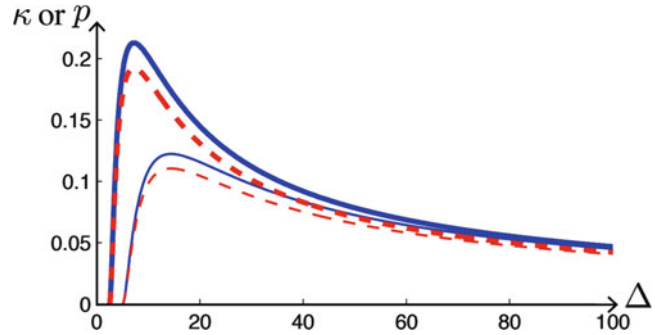
Next, consider the behavior of the counterfeiting and the passed rate. Figure 3, plots them as a function of the denomination. Crucially, this rate is first zero, then rising and eventually falling. This reflects some simple truths. For progressively higher denominations the earlier zero-profit locus $\bar{\Pi}$ and the optimal quality locus Q^* shift right, thereby raising quality; but the zero-profit curve moves further to the right and thus verification rises too, see Figure 7, (bottom) in Quercioli and Smith (2014). The inflated verification rate and quality require a greater verification effort too. So, the verification rate, verification effort, and quality all rise in the denomination Δ .

But notice that the counterfeiting rate moves ambiguously. For it is the quotient of the marginal

Counterfeit Money,
Fig. 3 Increased legal
 costs lower verification



Counterfeit Money,
Fig. 4 Increased legal
 penalty lowers the passed
 rate and counterfeiting rate



verification cost and the denomination value, and both of these rise; however, the first rises proportionately faster at the outset, since it starts at zero. Approaching the lowest note $\Delta \rightarrow \underline{\Delta}$ that is still profitable to counterfeit, the counterfeiting rate and passed rate both vanish, since the marginal verification costs of counterfeiting vanish. Loosely, people choose to pay little attention when handed these notes. So, both the counterfeiting rate and passed rate must vanish in this limit too, in order to justify this as an optimal choice. Such inverse reasoning is typical of the strategic analysis in Quercioli and Smith (2014). At large enough notes $\Delta \rightarrow \infty$ (a limit that does not obtain for the existing US dollar denominations), this turns around, since the marginal cost of increasing quality rises without bound.

Finally, it is worth observing that the passed rate reaches a maximum at a lower denomination than the counterfeit rate, since their quotient is the verification rate, which is increasing in the note. This is important to keep in mind, in that the most counterfeited note might well exceed the most passed note and certainly is not less.

Now, let us combine our last two lines of thinking, and consider what happens to these whole curves with a rise in the legal penalty Λ . We have seen that *at any given denomination*, this depresses both the verification rate and the quality (see Fig. 2). But our first formulas for the passed rate and the counterfeiting rate are increasing in both the verification rate and the quality. Consequently, both the passed rate and counterfeiting rate curves fall, as seen in Fig. 4.

For a slightly different perspective, observe how Fig. 2 included a third type of curve, namely, a *constant counterfeiting rate locus* $\bar{K}$. This is derived from the optimization of innocent verifiers and consists of all verification rates that would be optimal for a constant counterfeiting rate and the specified quality. Intuitively, this is downward sloping, since a lower verification rate is optimal for a higher quality, holding fixed the counterfeiting rate. This is nicely sandwiched horizontally between $\bar{\Pi}$ and Q^* . Figure 2 shows that the increased legal costs results in a lower quality, and thus the new constant counterfeiting locus $\bar{K}_2$ lies below $\bar{K}_1$.

Quercioli and Smith (2014) also used this graphical apparatus to flesh out predictions for improvements in the counterfeiting technology or in the ease of verification. For instance, cheaper counterfeiting technology lowers the verification rate and raises the counterfeiting rate. In each case, there are feedback effects that undermine but do not overwhelm the intuitive, first-order shifts.

Empirical Evidence via Seized and Passed Money

The total value of *passed money* every “period” is $P[\Delta]$. In fact, the length of the period falls in the velocity of circulation of money. Quercioli and Smith (2014) explore this nuance more carefully, but here it is assumed, for simplicity, to be the average length of time it takes for a note to change hands. Define total counterfeit money produced as $C[\Delta]$. In a steady state, this production must be balanced by the total outflow of seized $S[\Delta]$ and passed $P[\Delta]$ money:

$$C[\Delta] = S[\Delta] + P[\Delta]$$

Also, the amount of passed money in circulation is in balance, and thus its inflow – namely, fakes that escape the attention of the first verifier – must balance its outflow:

$$(1 - v)C[\Delta] = P[\Delta]$$

Combining the two expressions reveals the importance of the seized-passed ratio:

$$\frac{1}{1 - v} = 1 + \frac{S[\Delta]}{P[\Delta]}$$

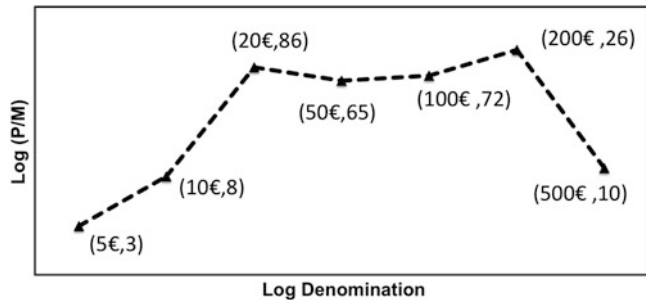
One can now understand that the seized-passed ratio could only have fallen over time if the verification rate had done so too. Quercioli and Smith (2014) identify a digital counterfeiting revolution – the advent of digital color printers, for instance – that strongly suggests a fall in the costs of counterfeiting. That, in turn, would have depressed the verification rate. Additionally, they document the fact that the *counterfeit-passed ratio* $C[\Delta]/P[\Delta]$ rises in the note, but less than proportionately so. For as argued, the verification rate rises in the denomination; on the other hand, higher denominations demand higher quality, in order to deflect increasingly attentive verifiers. All in all, this raises the average costs of counterfeiting. For example, if \$100 bills passed half as often as \$50 bills, then counterfeiters of \$100 bills would lose money. This explains why the counterfeiting revenues, namely, the chance that a note passes times the denomination, must rise in the denomination. In other words, the inverse counterfeit-passed ratio rises in the denomination but less than proportionately, as claimed. We conclude that the model makes sense of the data patterns we observe.

Next, Fig. 5 reports data for the euro counterfeits passed rates from 2002 to 2013 as initially rising and then falling. These data are consistent with the earlier plot depicted in Fig. 3. While the US dollar denominations do not rise high enough to show this, the euro denominations certainly do: notice how the passed rate plummets, starting at the €200 and €500 denominations.

As mentioned in the section “[Introduction](#),” a curious inverse piece of evidence for the costly

Counterfeit Money,

Fig. 5 The average ratio of passed counterfeit notes to money circulation (in millions) for the euro currency from 2002 to 2013, by the seven euro denominations



attention model is provided by counterfeit money that is missed by verifiers and regular banks. Commercial banks typically pass damaged notes to FRBs. Let the *FRB-ratio* be defined as the counterfeiting rate at FRBs divided by the average counterfeiting rate. Quercioli and Smith (2014) show that this ratio *falls* in the note, except for the \$100 note. So, even though \$1 is the poorest quality counterfeit, it is disproportionately often found at the FRBs. The reason is that this note secures the lowest verification rate and so it is missed systematically by banks and individuals. Eventually, it surfaces at the FRBs.

One can equally well use the Quercioli and Smith (2014) model to estimate the underlying parameters of the model. Our earlier expression for the seized-passed ratio rearranges to:

$$v = \frac{S[\Delta]}{S[\Delta] + P[\Delta]}.$$

In fact, this is an upper bound on the true verification rate, since the police undoubtedly intercept some counterfeit notes. Calculating the right-hand-side ratio for the US dollar seized and passed data in Quercioli and Smith (2014) reveals that the verification rate upper bounds range from 0.25 (for the \$5 note) to 0.31, 0.43, 0.52, and 0.54 (for, respectively, the \$10, \$20, \$50, and \$100 notes). This increasing verification rate agrees with the model predictions. It is also interesting to notice the implications of this finding for the current literature on counterfeiting. It unequivocally refutes any assumption that agents observe fixed authenticity signals for notes, as contended, for example, by Williamson (2002). For the verification rate is clearly not fixed, but varies by denomination.

But, what about the so-called street price of a counterfeit note? Logically, this must – at most – equal the average cost of production of a counterfeit note. The conjectured average costs from Quercioli and Smith (2014) by and large agree with the available anecdotal evidence on this. For instance, in one recent case, a Mexican counterfeiting ring sold counterfeit \$100 notes at 18% of their face value to distributors, who then resold the counterfeit notes for 25–40% of their

face value. The money was transported across the border by women couriers.

In modeling counterfeiting, an important question is whether one can estimate the unobserved counterfeiting rate. As we know, its observable manifestation is the passed rate $p[\Delta] = v[\Delta]\kappa[\Delta]$. Of course, this passed rate is on a “per transaction” basis and not annual. For now, let us assume that the annual velocities of notes are all equal, say, to 10. In this case, the actual passed rates are all $10 p[\Delta]$. Given the upper bounds discussed above on verification rates (by denominations), we can compute the lower bounds for the counterfeiting rate. Dividing annual passed rates by 10 and then dividing by the verification rates yields the following lower bounds for the counterfeiting rate – for, respectively, the \$5, \$10, \$20, \$50, and \$100 notes – 0.87, 2.517, 1.733, 1.013, and 1.019 per 100,000. In fact, velocities fall in the denomination – intuitively \$1 bills are spent far more often than \$100 notes – and thus the observed annualized passed rates should fall relative to the per-transaction rates. That skews the numerical implications of the theory. See Quercioli and Smith (2014) for more details.

The Social Costs of Counterfeiting

We conclude with a final word on the costs imposed on society as a whole by counterfeiting. A passed counterfeit note incurs one – and only one – counterfeiting cost but many verification costs until it is eventually discovered. If it survives inspection with chance $(1 - v)$, then it will be inspected on average $1/v$ times. Let us approximate (thereby overstating) the verification cost by the marginal verification cost times v . Since the counterfeiting rate κ is the ratio of the marginal verification cost and Δ , an upper bound on the total verification expenses for the Δ note in its life is given by $v \times (\text{marginal verification cost})/v = \kappa\Delta$. Now, compare this to the insights of Tullock (1967). He predicts that parties to a transfer, or a theft, of D dollars should be collectively willing to spend up to D to affect the transfer, or the theft. So, the stochastic nature of

counterfeiting holds the actual rent-seeking costs – namely, *the social costs of this crime* – to a tiny fraction of their value.

References

- Green E, Weber W (1996) Will the new \$100 bill decrease counterfeiting? Fed Reserve Bank Minneap Q Rev 20(3):3–10
- Judson R, Porter R (2003) Estimating the worldwide volume of U.S. counterfeit currency: data and extrapolation. Finance and Economics Discussion Paper, Board of Governors of the Federal Reserve System
- Kiyotaki N, Wright R (1989) On money as a medium of exchange. J Polit Econ 97(4):927–954
- Nosal E, Wallace N (2007) A model of (the threat of) counterfeiting. J Monet Econ 54(4):994–1001
- Quercioli E, Smith L (2014) The Economics of Counterfeiting. SSRN ID 1325892:1–24. Conditionally accepted at Econometrica
- The Globe and Mail (2002) The unwanted bill: has the C-note lost its currency?. , Saturday, 4 May 2002. www.globeandmail.com
- Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. Western Econ J 5(3):222–232
- U.S. Department of Homeland Security (2012) United States Secret Service Annual Report
- Williamson S (2002) Private money and counterfeiting. Fed Reserve Bank Richmond Econ Q 88(3):37–57

Counterfeiting Models: Mathematical/Economic

Andrea Di Liddo

Department of Economics, University of Foggia, Foggia, Italy

Abstract

Counterfeiting and piracy are illicit activities infringing IPR (intellectual property rights). The market for counterfeit can be divided into two important submarkets. In the primary market, consumers purchase counterfeit products believing they have purchased genuine articles (deceptive counterfeiting). In the secondary market, consumers knowingly buy counterfeit products (nondeceptive counterfeiting).

Counterfeiting has, obviously, consequences on genuine producers and consumers; nevertheless, it can have general socioeconomic effects.

There is a considerable body of theoretical and empirical literature on the mechanisms of counterfeit trade and on the economic and social effects of counterfeiting. A number of the methodological papers are undertaken within the framework of operations research and game theory.

Synonyms

Fake; Forgery; Imitation

Introduction

Counterfeiting and piracy are illicit activities infringing IPR (intellectual property rights).

Counterfeit trademark goods and pirated copyright goods are well defined in the Trade-Related Aspects of Intellectual Property Rights (TRIPS) Agreement, signed in Marrakesh, Morocco, on 15 April 1994. Counterfeit trademark goods are goods which cannot be distinguished in their essential aspects from genuine trademark goods and which thereby infringe the rights of the owner of the trademark. Pirated copyright goods are copies made without the consent of the right holder and that constitute an infringement of a copyright or a related right.

In the present context, according to Abalos (1985), the word “counterfeit” does not cover trade practices as gray market sales, parallel sales, diverted sales, passing off, counterfeiting of money, or money substitutes.

The market for counterfeit can be divided into two important submarkets. In the primary market, consumers purchase counterfeit products believing they have purchased genuine articles (deceptive counterfeiting). The products are often substandard and bring health and safety risks that can be very serious. In the secondary market, consumers knowingly buy counterfeit products (nondeceptive counterfeiting) (OECD 2008).

Typical counterfeited products are luxury goods; pharmaceuticals; and automotive and electrical components. However, almost any product can be counterfeited.

The Economic and Social Consequences of Counterfeiting

Counterfeiting has, obviously, consequences on genuine producers and consumers; nevertheless, it can have general socioeconomic effects.

A number of detailed and enlightening reports and books about the consequences of counterfeiting and piracy have been published by international organizations and scholars. Among them we mention OECD (2008), OECD (2016), and Chacharkar (2013).

The latest estimate from the International Chamber of Commerce (ICC) indicates that the total value of counterfeiting and piracy could reach the staggering level of \$2.8 trillion by the end of 2022 (ICC 2017).

The extent of counterfeiting is higher in developing economies also because of the relatively weak enforcement of IP.

Counterfeiters often are linked to organized crime and corrupt government officials in charge of countering their illegal activities.

Usually counterfeit products are cheaper than genuine goods but are of inferior quality. Counterfeiting of medicines, airplane, and auto parts has a detrimental effect on the health and safety of the public. The World Trade Organization estimates that approximately 70 percent of all medicines sold in some African countries are counterfeit.

The owners of the intellectual property may suffer loss of revenues from price pressures, royalties, sales, and brand dilution. Moreover, costs will increase because of legal liability and a higher number of warranty claims.

Counterfeiting harms not only individual consumers and businesses but also the society as a whole. Counterfeiting has effects on tax revenues, government expenditures, and, as a consequence of bribery, the effectiveness of public institutions. Governments also suffer additional costs associated with customs; judicial proceedings; increasing of public awareness; and handling of seized goods.

Counterfeiting also discourages innovation. R&D resources are diverted from creating new technologies for consumers to build up more effective methods to deter counterfeiters.

Destruction of counterfeited items is costly and results in bulky waste; moreover, shoddy fakes

can seriously damage the environment (e.g., in the case of chemicals industry).

Counterfeiting affects also employment. Employment shifts from genuine firms to infringing ones, where workers live in less healthy conditions and receive lower wages.

Strategies to Fight Counterfeiting

Governments and industries fight counterfeiting on a number of fronts, both independently and, equally importantly, with each other through multilateral institutions and on a bilateral and regional basis. A comprehensive multilateral legal framework has been established within the World Trade Organization (WTO).

A good and enforced system of IPR is necessary to counter counterfeiting and piracy, but it is not sufficient. Measures to limit the free circulation of fakes together with address efforts to fight organized crime are of undoubtedly utility. Advertising campaigns to heighten consumer awareness have often been conducted. The contribution of genuine producers to fight counterfeiting is crucial because of their experience and knowledge and it efficiently complements government action.

Modeling Counterfeiting

There is a considerable body of theoretical and empirical literature on the mechanisms of counterfeit trade and on the economic and social effects of counterfeiting. A number of methodological papers are undertaken within the framework of operations research and game theory. Two excellent literature reviews are provided by Staake et al. (2009) and Cesareo (2016).

According to Staake et al. (2009), contributions can be classified as follows: public perception of counterfeiting and its relevance to management theory and practice; qualitative and quantitative investigations on the consequences of counterfeiting for manufacturers of genuine goods and their supply chain partners; supply-side investigations concerning the production settings, tactics, and motives of illicit actors, and the ways in which their products enter the licit supply chain; demand

side investigations focusing on customer behavior and attitudes in the presence of counterfeit goods; managerial guidelines; and legal and legislative issues concerning different options for IP rights enforcements.

Cesareo (2016), in her book, identify, analyze, and systematize the available research on counterfeiting and piracy published over a 35-year time span (1980–2015).

Among methodological papers, Grossman and Shapiro (1988a, b) can certainly be considered milestones.

Grossman and Shapiro (1988a) focus on the effects of deceptive counterfeiting. They develop an equilibrium model of counterfeit-product trade, incorporating into their analysis both the direct effects of counterfeiting and the induced effects on the behavior of legitimate producers. Their analysis is conducted in a dynamic, two-country model with imperfect quality information and brand-name reputations.

Grossman and Shapiro (1988b) investigate non-deceptive counterfeiting. They develop an equilibrium model of the market for status goods, i.e., those goods for which the mere use or display of a particular branded product confers prestige on their owners. The counterfeiting of a status good, then, deceives not the individual who purchases the product but rather the observer who sees the good being consumed and is mistakenly impressed. The efficacy of two alternative policies is investigated: enforcement and confiscation; tariffs on low-quality goods. Interestingly, the authors prove that it is not true in general that stricter enforcement is welfare-improving.

Supply Side Investigations

Researches dedicated to the supply-side issues of the counterfeit market are of great importance for understanding the way the illicit market operates and how licit brand owners can fight illicit producers.

Qian (2014) provides a theory for brand-protection strategies to reduce counterfeiting under weak IPR. His model incorporates two layers of asymmetric information that counterfeits can incur:

counterfeiters fooling consumers and buyers of counterfeits fooling other consumers. One of the theoretical predictions of this study is that counterfeit entry induces incumbent brands to introduce new products. Moreover, better channel management complements a company's own enforcements against counterfeits.

Cho et al. (2015) prove that the effectiveness of anticounterfeiting strategies depends critically on whether a brand-name company faces a non-deceptive or deceptive counterfeiter. Therefore, firms and governments should carefully consider a trade-off among different objectives in implementing an anticounterfeiting strategy.

Dual channel supply-structures have been considered in Zhang and Zhang (2015) as a means of mitigating counterfeiting activities. Adopting the vertical differentiation model, consumers' utility toward the brand name product is described as a function of the price and of the perceived quality. The main finding is that the brand name company should continue to sell, sometimes exclusively, through the general channel despite deceptive counterfeiting.

Yao (2005) considers a market where a monopolist sells a genuine luxury product and counterfeiters illegally copy the product and can enter and exit the market freely. Fines paid by caught counterfeiters are supposed to be pocketed by the monopolist and pegged to the price of the genuine product. Surprisingly, the author shows that it is not always true that the presence of counterfeit products hurts genuine producers. Similar results are found in Di Liddo (2015), but with fines not pegged to the price of the true branded items, and in Buratto et al. (2015) where a dynamic framework is adopted.

Demand Side Investigations

An important field of research addresses awareness, purchase intentions, demographic characteristics, or the attitudes of counterfeit consumers.

Eisend and Schuchert-Güler (2006) review selected studies on the determinants of consumers' intention to purchase counterfeit products and provide a theoretical concept in order to explain the motives when purchasing such goods.

Gentry et al. (2006) investigate product counterfeiting from a consumer search perspective. They prove that factors positively affecting purchase decisions of counterfeits are their low prices, the low investment risks when buying low-cost fake, and, in Western markets, the fun of showing imitation products to friends. In China especially, the potential loss of face when exposed as a counterfeit consumer negatively affects the decision to purchase counterfeit goods.

In order to curb counterfeiting growth, most countries enact codes to punish those counterfeiting firms that are caught. Nevertheless, in Italy and in France purchasers of counterfeit products are also fined and authorities have the right to confiscate their counterfeit items when found. Yao (2015) shows that imposing such penalties reduce demand and hence profit of the legitimate producer under some situations. Under uniform distribution of consumers in product quality estimation, social welfare is reduced. Consequently, counterfeit-purchase penalties employed in some countries are not recommended.

Consumer demand for counterfeit luxury brands is often viewed as unethical, but the demand is also robust and growing. Employing in-depth interviews, Bian et al. (2016) identify the psychological and emotional insights that both drive and result from the consumption of higher involvement counterfeit goods. Moreover, they uncover the coping strategies related to unethical counterfeit consumption.

Legal Issues and Legislative Concerns

A wide range of industries agree that there is severe problem with the protection of IPR throughout the world. The book by Chaudhry and Zimmerman (2013) aims to give the most complete description of various characteristics of the IPR environment in a global context. Authors believe that a holistic understanding of the problem must include consumer complicity to purchase counterfeit products, tactics of the counterfeiters as well as actions by home and host governments, and the role of international organizations and industry alliances.

Cross-References

- Copyright
- Imitation
- Innovation
- Intellectual Property: Economic Justification
- Optimization Problems

References

- Abalos RJ (1985) Commercial trademark counterfeiting in the United States, the third world and beyond: American and international attempts to stem the tide. *Boston Coll Third World Law J* 5:151–182
- Bian X, Kai-Yu W, Smith A, Yannopoulou N (2016) New insights into unethical counterfeit consumption. *J Bus Res* 69:4249–4258
- Buratto A, Grosset L, Zaccour G (2015) Strategic pricing and advertising in the presence of a counterfeiter. *IMA J Manag Math* 27:397–418
- Cesareo L (2016) Counterfeiting and piracy. A comprehensive literature review. Springer, Heidelberg
- Chacharkar DY (2013) Brand imitation, counterfeiting and consumers. Centre for Consumer Studies Indian Institute of Public Administration, New Delhi
- Chaudhry PE, Zimmerman A (2013) Protecting your intellectual property rights. Springer, New York
- Cho SH, Fang X, Tayur S (2015) Combating strategic counterfeiters in licit and illicit supply chains. *Manuf Serv Oper Manage* 17:273–289
- Di Liddo A (2015) Does counterfeiting benefit genuine manufacturer? The role of production costs. *Eur J Law Econ* 3:1–45
- Eisend M, Schuchert-Güler P (2006) Explaining counterfeit purchases: a review and preview. *Acad Mark Sci Rev* 10:1–25
- Gentry JW, Putrevu S, Shultz CJ II (2006) The effects of counterfeiting on consumer search. *J Consum Behav* 5:245–256
- Grossman GM, Shapiro C (1988a) Counterfeit-product trade. *Am Econ Rev* 78:59–75
- Grossman GM, Shapiro C (1988b) Foreign counterfeiting of status goods. *Q J Econ* 103:79–100
- ICC (2017.) <https://cdn.iccwbo.org/content/uploads/sites/3/2017/02/ICC-BASCAP-Frontier-report-2016.pdf>. Last accessed 30 Mar 2017
- OECD (Organization for Economic Cooperation and Development) (2008) The economic impact of counterfeiting and piracy. OECD Publishing, Paris
- OECD/EUIPO (2016) Trade in counterfeit and pirated goods: mapping the economic impact. OECD Publishing, Paris
- Qian Y (2014) Brand management and strategies against counterfeiters. *J Econ Manag Strateg* 23:317–343
- Staake T, Thiesse T, Fleisch E (2009) The emergence of counterfeit trade: a literature review. *Eur J Mark* 43:320–349

- Yao JT (2005) How a luxury monopolist might benefit from a stringent counterfeit monitoring regime. *Int J Bus Econ* 4:177–192
- Yao JT (2015) The impact of counterfeit-purchase penalties on anti-counterfeiting under deceptive counterfeiting. *J Econ Bus* 80:51–61
- Zhang J, Zhang RQ (2015) Supply chain structure in a market with deceptive counterfeits. *Eur J Oper Res* 240:84–97

Counterterrorism

Marie-Helen Maras

John Jay College of Criminal Justice, City
University of New York, New York, NY, USA

Abstract

The European Union has developed several policies and actions to combat terrorism. The EU's counterterrorism strategy is fourfold. It seeks to prevent terrorism, protect people and property against terrorism, pursue terrorists, and respond to terrorism. To achieve this, the EU has implemented measures that seek to disrupt terrorists' operations, prevent radicalization and recruitment of terrorists, secure critical infrastructures and borders, impede terrorists' plans and actions, cut off terrorists' funding and support, improve information sharing, enhance cooperation among agencies, and coordinate responses in the event of a terrorist attack.

Definition

Counterterrorism refers to the measures taken to prevent, deter, pursue, and respond to terrorism. These measures are also designed to protect individuals and property.

Counterterrorism

There is no universally accepted definition of terrorism. However, there are certain elements of terrorism that most governments, politicians,

policymakers, practitioners, and academicians agree upon. More specifically, it is generally accepted that those who engage in terrorism use coercive tactics (e.g., threats or the use of violence) “to promote control, fear, and intimidation within the target nation or nations for political, religious, or ideological reasons” (Maras 2012, p. 11). The policies and measures implemented to combat this threat are a form of counterterrorism. Counterterrorism includes a plan of action with specific measures designed to deal with the threat of terrorism.

The EU's counterterrorism strategy is fourfold. Firstly, the EU seeks to prevent terrorism. Here, prevention focuses on reducing the opportunities for engaging in terrorism. Opportunities for terrorism can be reduced by increasing the effort involved, increasing the risks of failure, reducing the rewards of terrorism, and removing temptations, provocations, and excuses for terrorism (Maras 2012, 2013). Most economic analyses of terrorism are based on Becker's (1968) rational choice approach to crime (e.g., Landes 1978; Frey 2004; Enders and Sandler 2006), which holds that an individual will engage in crime if the expected utility of that illicit activity exceeds the gains of engaging in other activities. Pursuant to this approach, terrorists are viewed as rational actors who pursue identifiable goals, decide whether or not to act by examining possible courses of actions in terms of costs and benefits, and assess each action's probability of success or failure. In light of this, countermeasures should make it more difficult for terrorists to achieve their objectives and thus increase the perceived costs of so doing. If the effort required to succeed in a task is raised high enough, it is believed that the terrorists might give up on that task or take longer to execute their operations. Accordingly, an effective counter strategy should focus on ways to frustrate criminals by making it more difficult and risky to commit crime and by reducing its rewards (Frey and Luechinger 2007). Another essential element in preventing terrorism is the implementation of measures targeting the radicalization (i.e., the process whereby individuals adopt extremist views) and recruitment of terrorists (Maras 2012, 2014). There are many different forums within which

radicalization and recruitment take place, for example, prisons and the Internet. The EU counterterrorism prevention strategy seeks to target these environments.

Secondly, the EU aims to protect against terrorism. As part of its counterterrorism strategy, the EU seeks to reduce its vulnerability to terrorist attacks and minimize the impact of an attack should it materialize. Protection involves the collective action to secure critical infrastructure, transportation sectors (e.g., airport, rail, and maritime sectors), and borders. Several new technologies, measures, and programs have been implemented to enhance critical infrastructure protection and border security.

The EU has designated the following sectors as critical infrastructure (CI): communications and information technology, finance, health care, energy, food, water, transport, government facilities, and the production, storage, and transport of dangerous goods (e.g., chemical and nuclear materials) (European Commission 2004). To protect CIs, the European Programme for Critical Infrastructure Protection (EPCIP) was implemented, which identifies critical infrastructure, determines interdependencies of critical infrastructure, analyzes existing vulnerabilities, and provides solutions to protect CIs (European Commission 2007). To facilitate EPCIP, critical infrastructure protection (CIP) expert groups were created, processes to enable the sharing of information on CIP were developed, and the Critical Infrastructure Warning Information Network (CIWIN) was implemented. The CIWIN is a warning network that sends rapid alerts about risks and vulnerabilities to critical infrastructure. Relevant public and private stakeholders are required to share information on critical infrastructure interdependency, the securing of critical infrastructure, vulnerabilities and threats to critical infrastructure, and risk assessments of critical infrastructures.

To protect borders, the passenger name record (PNR) agreement was implemented, which called for the collection and storage of information on all passengers traveling in and out of the EU (European Commission 2010). The type of data obtained and recorded includes contact

information, billing information, the travel agency where the ticket was purchased, and any available Advance Passenger Information (API) (e.g., date of birth and nationality) (Directive 2004/82/EC). The Visa Information System (VIS) was also created. This database contains visa application information and biometrics of individuals required to have a visa to enter the Schengen Area (European Commission 2013b). Additionally, the Schengen Information System (SIS) was developed. This system holds information on missing persons and stolen, missing, or lost property (e.g., cars, firearms, and identity documents). It further includes information on individuals who are not authorized to stay or enter into the EU or are suspected of being involved in serious crime (European Commission 2013a). SIS is used by law enforcement, judicial authorities, customs agents, and visa-issuing authorities. SIS II was implemented on April 9, 2013, and includes enhanced functionalities such as the ability to link different alerts (e.g., a stolen vehicle and missing persons alert) and engage in direct queries on the system (European Commission 2013b). Moreover, the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union (Frontex) was set up to improve the management of the external borders of the EU. Frontex is responsible for the control and surveillance of borders as well as for facilitating the implementation of existing and prospective measures concerning the management of external EU borders.

Thirdly, the EU engages in the pursuit of terrorists across borders. This part of the EU's counterterrorism strategy focuses on impeding terrorists' activities (i.e., their abilities to plan and organize attacks, to obtain weapons, to receive training, and to finance operations), improving information gathering and analysis, enhancing police and judicial cooperation, and combating terrorist financing. For instance, to interdict terrorists, the EU Action Plan for Enhancing the Security of Explosives, the European Bomb Database, and the Early Warning System for Explosives and Chemical, Biological, Radiological and Nuclear (CBRN) material were

developed. Here, the counterterrorism strategy focuses on cutting off terrorists' access to materials that can be used in an attack. The Prüm Treaty facilitated and streamlined the exchange of information and intelligence between law enforcement agencies of the EU Member States. It was created in order to improve EU-wide access to and the exchange of information. Additionally, the Hague Programme removed national borders in data collection, storage, and use, thereby creating an EU-wide right of use of data (Balzacq et al. 2006). To facilitate the transfer of person and evidence between Member States, the European Arrest Warrant (Council Framework Decision 2002/584/JHA) and the European Evidence Warrant (Council Framework Decision 2008/978/JHA) were created and implemented. Furthermore, Directive 2005/60/EC was implemented to prevent the use of the financial system for money laundering and terrorist financing.

Finally, the EU seeks to respond to terrorism. To coordinate responses to acts of terrorism, the EU has implemented several measures. Among the most prominent of which are the Crisis Coordination Committee (CCC) and the ARGUS system (a rapid alert system). The CCC evaluates and monitors the development of crises or emergency situations. Specifically, it identifies issues relevant to the situation and options for decisions and actions. ARGUS links all relevant services of the European Commission during a crisis or an emergency (European Commission 2014). It enables the exchange of real-time information in the event of a crisis and/or foreseeable (or imminent) threat that requires action on the European Community level (European Commission 2005). Europol was established by the Treaty on European Union (Maastricht Treaty) of 1992 and also plays a vital role in responses to terrorism by facilitating information exchange between agencies. This agency is primarily concerned with disrupting criminal and terrorist networks and assisting Member States in the EU in their investigations of criminals and terrorists (Europol 2011).

Overall, in its fourfold strategy, the EU seeks to harmonize existing measures and actions aimed at combating terrorism in order to effectively and efficiently counter this threat. These

harmonization attempts, however, need to be cost-effective. Cost-benefit analysis can be used to determine whether counterterrorism resources are best allocated to protect life, human rights, and property. Impact assessments on major policies are used to determine if such an efficient allocation of resources is occurring.

Cross-References

- Cost-Benefit Analysis
- Impact Assessment

References

- Balzacq T, Bigo D, Carrera S, Guild E (2006) Security and the two-level game: the treaty of Prüm, the EU and the management of threats. CEPS working document no 234
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Council Framework Decision 2002/584/JHA of 13 June 2002 on the European arrest warrant and the surrender procedures between Member States
- Council Framework Decision 2008/978/JHA of 18 December 2008 on the European evidence warrant for the purpose of obtaining objects, documents and data for use in proceedings in criminal matters
- Directive 2004/82/EC on the obligation of carriers to communicate passenger data (2004) OJ L 261
- Directive 2005/60/EC of the European Parliament and of the Council of 26 October 2005 on the prevention of the use of the financial system for the purpose of money laundering and terrorist financing
- Enders W, Sandler T (2006) The political economy of terrorism. Cambridge University Press, Cambridge
- European Commission (2004) Communication from the commission to the council and the European parliament of 20 October 2004 – critical infrastructure protection in the fight against terrorism. COM (2004) 702 final. http://europa.eu/legislation_summaries/justice_freedom_security/fight_against_terrorism/133259_en.htm
- European Commission (2005) Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions – commission provisions on “ARGUS” general rapid alert system. COM (2005) 662 final. <http://eur-ex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52005DC0662:EN:HTML>
- European Commission (2007) Communication from the commission of 12 December 2006 on a European programme for critical infrastructure protection. COM (2006) 786 final (7 June 2007). http://europa.eu/legislation_summaries/justice_freedom_security/fight_against_terrorism/133260_en.htm

- European Commission (2010) On the global approach to transfers of Passenger Name Record (PNR) data to third countries. COM (2010) 492 final (21 Sept 2010). http://europa.eu/legislation_summaries/justice_freedom_security/fight_against_terrorism/jl0043_en.htm
- European Commission (2013a) Schengen Information System (SIS). http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/schengen-information-system/index_en.htm
- European Commission (2013b) Visa Information System (VIS). http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/visa-information-system/index_en.htm
- European Commission (2014) ARGUS – a general European rapid alert system. http://ec.europa.eu/health/preparedness_response/generic_preparedness/planning/argus_en.htm
- Europol (2011) Protecting Europe. <https://www.europol.europa.eu/content/page/protecting-europe-21>
- Frey BS (2004) *Dealing with terrorism: stick or carrot?* Edward Elgar, Cheltenham
- Frey BS, Luechinger S (2007) Decentralization as a response to terror. In: Brück T (ed) *The economic analysis of terrorism*. Routledge, Abington
- Landes WN (1978) An economic study of U.S. aircraft hijacking: 1961–1976. *J Law Econ* 21:1–31
- Maras M-H (2012) *Counterterrorism*. Jones and Bartlett, Burlington
- Maras M-H (ed) (2013) *The CRC Press terrorism reader*. CRC Press, Boca Raton
- Maras M-H (2014) *Transnational security*. CRC Press, Boca Raton

Court of Justice of the European Union

Roland Vaubel

Universität Mannheim, Mannheim, Germany

Abstract

Quantitative studies are surveyed which indicate that the Court is biased towards political centralization. The econometric evidence shows that this bias is not due to a lack of political independence but to self-selection and vested interest. Various reforms are discussed which would reduce self-selection and vested interest, notably the requirement of judicial experience, delegation of judges from the highest national courts, and the establishment of a separate “Subsidiarity Court.”

Definition

The Court of Justice of the European Union is the highest court of the European Union.

Before 2009, the Court of Justice of the European Union used to be called the European Court of Justice or, as in the pre-2009 treaties, simply “the Court of Justice.” Its predecessor was the Court of Justice of the European Coal and Steel Community. The Court has its seat in Luxembourg. It is composed of one judge from each member state. The judges are appointed by common accord of the governments for a term of 6 years. Reappointment is possible and frequent. The mean term length has been 9.3 years (Voigt 2003). Every 3 years, some of the judges are replaced. Cases are decided by the full court (27 judges), the Grand Chamber (11 judges), or chambers of three or five judges. Decisions are taken by simple majority except in chambers of three judges or concerning the expulsion of judges. Dissenting opinions are neither written nor published. The Court interprets the European treaties and the secondary legislation of the European Union at the initiative of (i) the other EU institutions; (ii) national governments; (iii) national courts (so-called preliminary rulings); (iv) national parliaments, or chambers thereof, if they claim that the principle of subsidiarity has been infringed by a legislative act; (v) the staff of the EU institutions; and (vi) a natural or legal person if a specific act is addressed, or is of direct and individual concern, to that person or if a regulatory act which is of direct concern does not entail implementing measures. There is also a General Court which, in some cases, serves as a court of first instance. Further details are set out in Art. 19 TEU, Articles 251–281 TFEU, and the Protocol on the Statute of the Court of Justice of the European Union.

The Court has often been called a “motor of integration” – both with respect to market integration and political integration, i.e., centralization. This general impression has been confirmed by several empirical studies. Stein (1981), in a seminal paper, showed that none of the

signatories of the Rome Treaty filed an observation in favor of any of the Court's major centralizing moves, while each of the member states opposed the Court in at least one of them. Jupille (2004, p. 98f.) demonstrated that the Court significantly favors the Commission at the expense of the Council. In 2008, Carruba, Gabel, and Hankla (CGH) analyzed the distribution and effects of such observations in a sample of 3,176 issues over a period of 11 years. Their data shows that the Commission and the net balance of the Council disagreed on 199 (6.3%) of the issues and that the Court sided with the Commission rather than the Council on 137 (68.8%) of these issues (Sweet and Brunell 2010, p. 28). Indeed, comparing the coefficients in CGH's regressions, the probability of a ruling in favor of the plaintiff is *significantly* higher when the Commission is the plaintiff or submits an observation favoring the plaintiff and *significantly* lower when the Commission is the defendant or submits an observation favoring the defendant (Vaubel 2009a, p. 216f.). Using the CGH data, Sweet and Brunell (2010) also show that regardless of the net position of the Council, the Commission's observations significantly affect the Court's decision in the desired direction. By contrast, if the Commission does not file an observation, the balance of the Council does not have a significant effect.

The Court's centralist and centralizing bias has been criticized by many scholars (e.g., Schermers 1974; Stein 1981; Philip 1983; Rasmussen 1986; Weiler 1991, 1999; Bzdera 1992; Burley and Mattli 1993; Garrett 1993; Neill 1995; Bednar et al. 1996, 2001; Garrett et al. 1998; Pitarakis and Tridimas 2003; Sweet 2004; Voigt 2003; Josselin and Marciano 2007; Höreth 2008; Vaubel 2009a, b, c). A court should not propagate a political program. It ought to be an impartial and objective interpreter of the law.

How can the Court's bias towards centralization be explained? The economic approach to law distinguishes between preferences and constraints. Is the Court constrained to decide as it does? The independence of the judges is protected in many ways:

- They are free to decide as they like.
- They enjoy immunity from legal proceedings even after they have left the Court. The immunity of an individual judge may not be waived except by the full Court (Art. 3 Statute).
- “A judge may be deprived of his office or of his right to a pension or other benefits in its stead only if, in the unanimous opinion of the judges and advocates-general of the Court of Justice, he (!) no longer fulfils the requisite conditions or meets the obligations arising from his (!) office” (Art. 6 Statute).
- The number of judges can only be increased by adding new member states or amending the Treaties (Art. 19 TEU).
- The Court's independence cannot be abolished or reduced except by amending the Treaties and the Statute, i.e., by a unanimous decision of the member states including their parliaments and/or electorates.

As for potential conflicts of interest, the Treaties emphasize that judges “shall be chosen from persons whose independence is beyond doubt” (Art. 19 TEU and Art. 253 TFEU). The Statute (Art. 4) adds that “the judges may not hold any political or administrative office” and “may not engage in any occupation, whether gainful or not.” Moreover, “before taking up his (!) duties each judge shall, before the Court of Justice sitting in open court, take an oath to perform his (!) duties impartially” (Art. 2 Statute). The fact that the judges are biased in favor of centralization is a violation of their oath.

The sole constraint which may affect their behavior is the provision that in order to be reappointed they require the support of their home government and the common accord of all member governments (Art. 19 TEU and Art. 253 TFEU). Term limitations and reappointments are extremely unusual for the judges of a supreme or constitutional court. Do they constrain the Court? The governments may not know how the individual judges – especially their own appointee – have voted in the chamber. The oath which the judges have to take also obliges them “to preserve the secrecy of the deliberations of the Court.” However, if one or more governments

dislike the Court's decisions, they may refuse to reappoint all judges seeking reappointment.

If the reappointment constraint were effective, it would favor the Council rather than the Commission. It would be a barrier to centralization, not a centralizing force. Thus, the Court's centralizing bias cannot be explained by a lack of independence. Moreover, a cross-section analysis of 41 national constitutional or supreme courts (Vaubel 2009a) shows that the share of central government in public expenditure is significantly larger in federal states if the independence of these supreme judges is strongly protected by the constitution. Constitutional judges do not centralize because they are dependent but because they are independent. They are an independent centralizing force, preferring to centralize.

In the same vein, the study shows that public expenditure is significantly more centralized when constitutional amendment is difficult, requiring a supermajority in parliament or several votes. Thus, the courts are more centralist than the constitutional legislator. They centralize more than the people or their parliamentary representatives want.

The only effective constraint on the centralizing tendencies of constitutional or supreme courts is legislative override. However, as Vaubel (2009a) and Sweet and Brunell (2010) have argued in some detail, legislative override is extremely difficult in the European Union. Treaty amendments require unanimity among the member governments and ratification by all parliaments of the member states. Revisions of secondary legislation always require a proposal from the Commission and usually the assent of the European Parliament, both of which share the centralist preferences of the judges. Moreover, the Council has to muster a qualified majority, in some cases even a unanimous vote. So far, there has been only one case of legislative override in the EU – the “Barber Protocol” in the Treaty of Maastricht (Vaubel 2009a; Sweet and Brunell 2010).

Finally, is the Court effectively constrained by the threat of noncompliance as some authors think? Sweet and Brunell (2010) reject this view: “The EU's legal system is organised to deal with non-compliance: member state non-compliance will generate legal actions and non-compliance

with any important ECJ ruling will generate new litigation, and new findings of non-compliance” (p. 9).

If the centralist bias of the judges is neither effectively constrained nor due to constraints, it must be attributed to their preferences. Why do they prefer more centralization than the governments, the national parliaments, and the voters do? Two explanations have been discussed (Vaubel 2009a).

The first is the self-selection hypothesis: The experts eligible for appointment to the EU Court are lawyers who believe in centralization rather than subsidiarity. That is why they have been specializing in EU law. They have studied what they cherish, not what they detest.

The second explanation is the vested interest hypothesis: The EU judges, like the Commission and the members of the European Parliament, are interested in centralization at the EU level because it increases their influence and prestige. The larger the powers of the European institutions and the larger therefore the extent of EU legislation, regulation, and administration, the more important and interesting are the cases that the EU judges will be entitled to decide. For example, constitutional courts have to adjudicate interinstitutional disputes at the same level of government. As long as the policy competence belongs to the member states, these disputes are not decided by the EU Court but by the national constitutional courts. But once the competence is transferred to the European level, the European Court is in charge.

These explanations are compatible with each other, and both are compatible with the evidence.

What can be done against centralization by the Court? There are three possible avenues for reform: facilitate legislative override, impose effective constraints on the judges, or alter the preferences of the judges.

Posner and Yoo (2005) argue that independent courts “can be effective only in an institutional setting where external agents such as executive and legislative branches of government . . . correct their errors” (p. 56). Thus, whenever the European Court reinterprets secondary law, the Council may be given the right to reverse the decision without a proposal from the Commission and without the

assent of the European Parliament. Alternatively, if a second chamber consisting of delegates of the national parliaments is added, as the European Constitutional Group (2004) has suggested, this chamber may be given the right of legislative override. Whenever the Court reinterprets the Treaties, the parliaments of the member states or this second chamber may be entitled to correct the judgment.

Could the judges be prevented from passing centralizing judgments in the first place? Again, three proposals come to mind.

First, as Weiler (1999, p. 131) suggests, a qualified majority of the judges may be required to overturn the legislation of the member states.

Second, the Court could be required to publish its voting record or any dissenting opinions. This would enable the governments to better evaluate the performance of their judges. However, judges ought to be independent.

Third, the judges might not be proposed and appointed by the governments of the member states. A survey of 18 modern democracies (Brouard and Hönnige 2010) shows that not a single one has a constitutional court whose judges are exclusively selected by government (s). In four countries, all judges are exclusively chosen by elected parliamentarians, in five countries the majority of the judges is chosen in this way, in six countries all judges have to be accepted by parliament, in one country (Germany) one half of the judges is selected by elected parliamentarians, and in the remaining two a minority of the judges is chosen in this way. In a supranational organization like the European Union, selection by a committee of the national parliaments or by a second chamber of the European Parliament might be appropriate.

Finally, how can the preferences of the judges be changed? Self-selection may be reduced by requiring judicial experience. In the past, only a minority of the lawyers appointed to the Court had previously served in a judicial function in their home country (Kuhn 1993, p. 195). The current president of the Court is a former professor and cabinet minister from Greece. According to the treaties, the judges shall be chosen from

persons “who possess the qualifications required for appointment to the highest judicial offices in their respective countries or who are jurisconsults of recognised competence” (Art. 253 TFEU). Moreover, “a panel shall be set up in order to give an opinion on the candidates’ suitability to perform the duties of Judge . . . The panel shall comprise seven persons chosen from among former members of the Court of Justice and the General Court, members of national supreme courts and lawyers of recognised competence, one of whom shall be proposed by the European Parliament” (Art. 255 TFEU). But this panel has merely a consultative role (Art. 253 TFEU).

The European Constitutional Group (2004) suggests that the EU judges should not only have judicial experience in their home country but also be drawn from its highest court (provided that they have gained judicial experience before being appointed to it). They would be delegated for a term of 8 years after which they would return to their national constitutional court. This would not only minimize self-selection but also improve the competence of the European judges and the integration of EU and national constitutional law. At present, cooperation between the EU Court and the national constitutional courts is also undermined by the preliminary reference procedure (Art. 267 TFEU) which enables the lower courts of the member states to appeal directly to the European Court, bypassing the highest national courts.

The Court’s vested interest in centralization is due to the fact that it is responsible for two tasks at the same time: (i) the task of allocating powers between the member states and the European Union and (ii) the task of interpreting EU law within those powers. The solution, therefore, is to separate these tasks and to have two European courts: one court that has no power other than adjudicating cases concerning the division of labor between the member states and the EU – call it the Subsidiarity Court – and one court that decides all other cases. This, too, has been suggested by the European Constitutional Group (2004). The judges of the Subsidiarity Court would be delegated from the highest courts of the member states. This arrangement would

neither invalidate the Court's decisions nor deprive the judges of their independence. It would correct their biased incentives. This is the economic approach.

References

- Bednar J, Ferejohn J, Garrett G (1996) Politics of European federalism. *Int Rev Law Econ* 16:279–294. [https://doi.org/10.1016/0144-8188\(96\)00020-8](https://doi.org/10.1016/0144-8188(96)00020-8)
- Bednar J, Eskridge W, Ferejohn J (2001) A political theory of federalism. In: Ferejohn J, Rackove J, Riley J (eds) *Constitutional culture and democratic rule*. Cambridge University Press, Cambridge, pp 223–270
- Brouard S, Hönnige C (2010) Constitutional courts as veto players. Lessons from Germany, France and the U.S. In: Paper presented at the Midwest Political Science Association conference, Chicago, Apr 2010
- Burley A-M, Mattli W (1993) Europe before the court. A political theory of legal integration. *Int Organ* 47:41–76. <https://doi.org/10.1017/S0020818300004707>
- Bzdera A (1992) The court of justice of the European community and the politics of institutional reform. *West Eur Polit* 15:122–136. <https://doi.org/10.1080/01402389208424925>
- Carruba CJ, Gabel M, Hankla C (2008) Judicial behavior under political constraints: evidence from the European Court of Justice. *Am Polit Sci Rev* 102:435–452. <https://doi.org/10.1017/S0003055408080350>
- European Constitutional Group (Bernholz P, Schneider F, Vaubel R, Vibert F) (2004) An alternative constitutional treaty for the European Union. *Public Choice* 91:451–468
- Garrett G (1993) The politics of legal integration in the European Union. *Int Organ* 49:171–181. <https://doi.org/10.1017/S0020818300001612>
- Garrett G, Kelemen D, Schulz H (1998) The European Court of Justice, national governments and legal integration in the European Union. *Int Organ* 52:149–176. <https://doi.org/10.1162/002081898550581>
- Höreth M (2008) Die Selbstautorisierung des Agenten: Der Europäische Gerichtshof im Vergleich zum US Supreme Court. *Nomos*, Baden-Baden
- Josselin J-M, Marciano A (2007) How the court made a federation of the EU. *Rev Int Organ* 2:59–76. <https://doi.org/10.1007/s11558-006-9001-y>
- Jupille J (2004) *Procedural politics: issues, interests and institutional choice in the European Union*. Cambridge University Press, Cambridge
- Kuhn B (1993) *Sozialraum Europa: Zentralisierung oder Dezentralisierung der Sozialpolitik?* Schulz-Kirchner, Idstein
- Neill SP (1995) *The European Court of Justice: a case study in judicial activism*. European Policy Forum, London
- Philip C (1983) *La cour de justice des communautés Européennes*. Presses Universitaires de France, Paris
- Pitarakis J-Y, Tridimas G (2003) Joint dynamics of legal economic integration in the European Union. *Eur J Law Econ* 16:357–368. <https://doi.org/10.1023/A:1025366909016>
- Posner EA, Yoo JC (2005) Judicial independence in international tribunals. *California Law Rev* 93:3–74
- Rasmussen H (1986) *On law and policy in the European Court of Justice*. Nijhoff, Dordrecht
- Schermers H (1974) The European Court of Justice: promoter of European integration. *Am J Comp Law* 22:444–464. <https://doi.org/10.2307/838965>
- Stein E (1981) Lawyers, judges and the making of a transnational constitution. *Am J Int Law* 75:1–27. <https://doi.org/10.2307/2201413>
- Sweet AS (ed) (2004) *The judicial construction of Europe*. Oxford University Press, Oxford
- Sweet AS, Brunell T (2010) How the European Union's legal system works – and does not work: response to Carruba, Gabel and Hankla. Faculty Scholarship series paper, vol 68. Yale Law School
- Vaubel R (2009a) Constitutional courts as promoters of political centralisation: lessons for the European Court of Justice. *Eur J Law Econ* 28(3):203–222
- Vaubel R (2009b) The European institutions as an interest group, vol 167, Hobart paper. Institute of Economic Affairs, London
- Vaubel R (2009c) The European constitution and interjurisdictional competition. In: Meessen KM (ed) *Economic law as an economic good*. Sellier European law publishers, Munich, pp 369–381
- Voigt S (2003) Iudex calculat: the ECJ's quest for power. *Jahrbuch für Neue Politische Ökonomie* 22:77–101
- Weiler JHH (1991) The transformation of Europe. *Yale Law J* 100:2403–2483. <https://doi.org/10.2307/796898>
- Weiler JHH (1999) *The constitution of Europe*. Cambridge University Press, Cambridge

Courts Voluntary Networks

Sylwia Morawska¹, Joanna Kuczevska² and Przemysław Banasik³

¹Warsaw School of Economics, Collegium of Business Administration, Warszawa, Poland

²Faculty of Economics, University of Gdansk, Sopot, Poland

³Faculty of Management and Economics, Gdansk University of Technology, Gdańsk, Poland

Abstract

Although the legal framework for their establishment and operation is identical, courts are marked by diversity. And it is not just about the differences arising from the court's place in the

hierarchy. Courts not only differ in the tangible resources (depending on size) and intangible resources (the knowledge and skills of employees) they possess but also in the organizational culture and the ability to learn (Banasik and Brdulak 2015) and the reputation they have, as well as in the network of contacts (Banasik and Morawska 2016). Courts are embedded in the dense structure of relations with the environment (Czakon 2007), including with other courts. The interorganizational cooperation between courts takes place not only within hierarchical, i.e., regulatory, networks (regulated courts networks) with regard to the tasks imposed by the legislature but also within heterarchical, i.e., voluntary, networks (voluntary courts networks). Courts should strive to harmonize the services they offer, as opposed to companies, where resources together with the core competencies built upon them serve to build a competitive advantage in the market. Courts do not compete for customers on their products or services. Jurisdiction is determined by regulations. What is more, the citizen has the right to the same services in each court. What may help standardization are voluntary courts networks, where courts will exchange good practices, managerial and organizational. Networking can also contribute to the organizational efficiency of courts of general jurisdiction through the rational use of resources and the harmonious interaction of all the elements of the organization.

Networking in Public Management

The rapidly developing science and management practice showed little interest in public sector organizations and the way they were managed. Meanwhile, in recent years, there has been an increased need for theoretical grounds for management in the public sector and in the units forming this sector, as part of the developing a specific discipline of management sciences, i.e., public management. Although management of

public organizations becomes, both theoretically and empirically, the object of research investigations and explorations by many scientific fields and disciplines and by practitioners, knowledge in this field is still insufficient. Studies on the behavior of public organizations worldwide in particular relate to public administration (central and local government), schools, or health organizations. In this context, there is a clear cognitive gap as regards the functioning of courts. Courts have so far been considered almost exclusively as a legal entity, with scientific papers concerning them dominated by the law faculties at universities.

Meanwhile, in the judiciary, we can observe phenomena similar to those that take place in other public organizations or in the private sector. This applies *inter alia* to the networking of courts of general jurisdiction. The network approach has been widely analyzed in the framework of the co-management paradigm. It emerged in the late 1980s and early 1990s of the twentieth century. The issue of network management has been introduced into considerations on management in the public sector by (Heclo 1978; Heclo and Wildavsky 1974). It was taken up by public management theoreticians (Marin and Mayntz 1991; Kooiman 1993; Scharpf 1994; Sorensen and Torfing 2007; March and Olsen 1995; Kickert et al. 1997; Rhodes 1997; Pierre and Peters 2000; Hill and Hupe 2002). However, there is no research on networking the judiciary. The judiciary has the potential to take advantage of the mechanisms of network cooperation. Within its framework, interorganizational cooperation can assume different relationships and interactions (Banasik 2015). Interorganizational cooperation is possible in auxiliary and basic (judicial) activities. Interorganizational cooperation and the possibility of its implementation in the core activities require in-depth research. It is characterized by specificity resulting from the fact that it concerns the administration of justice, where judges are independent. Interorganizational networks in the area of the core activities in civil, criminal, economic cases, etc. in horizontal systems can promote unification of views within homogeneous factual circumstances by obtaining consensus

within different interpretational views and thus build confidence in the judiciary. Creating interorganizational bonds in the auxiliary activities primarily serves the transfer of good practices, managerial and organizational.

Networking in the Judiciary: Research Results

Studies relating to the phenomenon of networking in courts of general jurisdiction in Poland were inspired by the results of pilot implementation of new management methods in courts of general jurisdiction undertaken under the project *PWP Edukacja w dziedzinie zarządzania czasem i kosztami postępowań – case management*. The pilot program was introduced in 60 courts of general jurisdiction. Its task was to change the managerial and organizational practices used in courts. The pilot program was also intended to unleash creativity, both in the managers and in the employees: justices and administrative staff of the courts. It served to initiate cooperation between courts, breaking the hierarchical subordination, within the voluntary hierarchical network. This was an additional unplanned result of the pilot program. The pilot program involved the courts of different hierarchy. Initially, the network consisted of 11 courts. At the end of the project, there were 65 courts cooperating within the network. On completion of the pilot program, courts continued to identify good managerial and organizational practices and within the framework of the resulting network exchanged information and knowledge on the possibilities of implementing management solutions in other courts. The pilot implementation of modern methods of managing the courts was in the nature of a research project of particular importance for the justice administration sector. It became a kind of “experimental field” for testing management methods and techniques. In view of the experiences gained by the public administration when implementing business practices, it was assumed that managerial and organizational “good practices” will be identified in courts, and practice from the business and

public administration sector will only support this process. The experiences gained by the public administration were to protect the judiciary against taking hasty decisions with regard to adapting solutions that in practice do not work in the public sector. In cooperation with the managerial staff from the pilot courts, the project developed 24 optimal organizational and technical solutions in the area of managing human resources/finances and information/knowledge (“good practices”), and by way of an experiment, up to 15 of them were implemented in each court taking part in the project on the basis of the implementation path developed. Working groups were set up in the pilot courts – staff groups made up of presidents and directors of the courts (60 presidents and 60 directors). The groups met once a month at the time of the pilot program. Their aim was to exchange information and knowledge between the presidents and directors of the pilot courts, i.e., mutual learning. In addition, for each practice proposed by external experts, thematic working subgroups were established, including the presidents and directors of the courts and employees implementing the practices in the primary pilot program.

During the pilot program, a network was established between the courts. It served the exchange of knowledge between particular entities and the creation of solutions aimed to improve management. It was also a platform for sharing good practices in the substantive area of action. Such a nature of action shares the reasoning represented by (Mandell and Keast 2008, 2009), who believe that the main purpose of the network is to connect its members, facilitating joint activities and learning, and consequently to create new solutions to existing problems. Research into the network formed between the courts shows that it is a highly integrated structure, as evidenced by the density on the level of 0.773, which indicates that on average almost 80% of the actors had the opportunity to cooperate with all the others during work on the project. Similar conclusions can be drawn from the level of the actor’s average extent, which indicates that in the course of the project, each of the courts had the opportunity to work

with representatives from approximately 50 other institutions. Also, a high average rate of clustering (a measure of how large a portion of the neighbors of a given actor also neighbor on each other) points to a strong integration of the entities involved in the project. Research indicates that the network created new relationships between the courts, which are in no direct hierarchical dependency relation.

To sum up, the network created as a result of the pilot program:

1. It is voluntary and cooperative (members of the network remain independent and interact only as needed, and the links between them are loose and sporadic; these networks assume a loose form of cooperation, their basic aim being to share knowledge).
2. There is no separate management unit, there is no strategic center, and the organizations take decisions on equal terms.
3. It is at its beginning stage of development, where relations are being built, standards are being established, and courses of action are being set.
4. The networks created as a result of the pilot program are regulated by their participants; thus they are in the nature of heterarchical networks.
5. The network under examination did not develop a strategic center. The network that was created was initiated by 11 presidents of courts of general jurisdiction taking part in the first project.

It follows from the analysis of the results of qualitative research that in the judiciary the development of network collaboration is facing serious constraints. As for the voluntary heterarchical networks, none of the presidents wants to act as a strategic center. Adopting this role by the president may cause a conflict between the president and the Minister of Justice, as in this case the president becomes a visionary and strategist for the community of courts. In this respect therefore, he poaches on the preserve of the Minister of Justice. The centrality and popularity of the strategic center of the network create a potential impact on the individual members of the network

and then in turn on the entire network. In addition, acting as a strategic center is a major managerial challenge, because it goes beyond the formal limits of the court. The courts participating in the network remain independent in formal and legal terms. The hierarchical tools used in managing the court are of no use here. Given the restrictions and threats to the strategic center, it seems that the court and its managing president can play the role of an initiating unit, but managing the network as such will require co-decision on the part of the courts co-participating in the network. In heterarchical networks, leadership is developed rather as a result of mutual activities by network participants (Müller-Seitz 2012).

Considerations regarding the strategic center in the newly created voluntary network are particularly important, because on completion of the pilot program the network is slowly dying away. Research shows that the leader of the voluntary courts network determines its existence. There is also need for an outside entity in order for the voluntary courts network to function properly. Its role should be limited to network management and the sharing of knowledge bases, while creating networking initiatives, and should be dependent on partners – the courts. In the case of voluntary courts networks, the present authors are in favor of a mixed solution, i.e., leadership combining the qualities of leadership based on consensus and rotary leadership depending on the unique resources held by the partners. In leadership based on consensus, the decision-making process and the goals are the result of joint activities by the members. Rotary leadership refers to a situation where every network participant has a chance to be the leader for some time. Interorganizational networks in the area of the judiciary are aimed at creating and exchanging knowledge, as well as finding new ways of solving problems.

Cross-References

- [Cooperative Game and the Law](#)
- [Heterarchy](#)
- [Horizontal Effects](#)

References

- Banasik P (2015) Organizacja wymiaru sprawiedliwości w strukturze sieci publicznej – możliwe interakcje. *E – mentor* 2(59):56. <http://www.e-mentor.edu.pl/artykul/index/numer/59/id/1171>
- Banasik P, Brdulak J (2015) Organisational culture and change management in courts, based on the examples of the Gdańsk area courts. *Int J Contemp Manag* 14:33–50
- Banasik P, Morawska S (2016) The Courts' public image – the desired direction of change. *Int J Court Adm* 8(1):2–11
- Czakon W (2007) *Dynamika więzi międzyorganizacyjnych przedsiębiorstwa*. Wydawnictwo Akademii Ekonomicznej im. Karola Adameckiego w Katowicach, Katowice
- Heclo H (1978) Issue networks and the executive establishment. In: King A (ed) *The new American political system*. American Enterprise Institute for Public Policy Research, Washington, p 103
- Heclo H, Wildavsky A (1974) *The private government of public money*. Macmillan, London
- Hill M, Hupe P (2002) *Implementing public policy: governance in theory and practice*. Sage, London
- Kickert WJM, Klijn EH, Koppenjan JFM (eds) (1997) *Managing complex networks*. Sage, London
- Kooiman J (ed) (1993) *Modern governance: new government – society interactions*. Sage, London
- Mandell MP, Keast R (2008) Evaluating the effectiveness of interorganizational relations through networks. Developing a framework for revised performance measures. *Public Manag Rev* 10(6):715–731
- Mandell M, Keast RL (2009) A new look at leadership in collaborative networks: process catalysts. In: Raffel J, Leisink P, Middlebrooks A (eds) *Public sector leadership: international challenges and perspectives*. Edward Elgar, Cheltenham, pp 163–178
- March JG, Olsen JP (1995) *Democratic governance*. The Free Press, New York
- Marin B, Mayntz R (eds) (1991) *Policy networks: empirical evidence and theoretical considerations*. Campus Verlag, Frankfurt-am-Main
- Müller-Seitz G (2012) Leadership in Interorganizational networks: a literature review and suggestions for future research. *Int J Manag Rev* 14:428–443
- Pierre J, Peters BG (2000) *Governance, politics and the state*. St. Martin's Press, New York
- Rhodes RAW (1997) *Understanding governance: policy networks, governance, reflexivity and accountability*. Open University Press, Buckingham
- Scharpf FW (1994) Games real actors negative coordination in embedded negotiations. *J Theor Politics* 6(1):27–53
- Sorensen E, Torfing J (eds) (2007) *Theories of democratic network governance*. Palgrave Macmillan, Basingstoke/New York

Craft Guilds

David Dolejší

Faculty of Social and Economic Studies, Jan Evangelista Purkyně University, Ústí nad Labem, Czech Republic

Abstract

Craft guilds were formal professional organizations in medieval and early modern Europe that operated in specific areas of industrial production. The primary objective of these organizations was to protect interests of their members. To achieve this goal, craft guilds engaged in a variety of activities from which obtaining legal monopolies over local production and trade within their crafts was the most important.

Definition

Craft guilds were geographically restricted government-licensed organizations of professionals who were specialized in certain areas of industrial production and who were concerned with securing welfare of their members mainly through possessing exclusive rights over the matters of their professions. These cooperatives were found in various forms across the world including China, Japan, the Middle East, Latin America, and North Africa, but their histories are mostly connected with medieval and early modern Europe.

European Craft Guilds

In premodern Europe, craft guilds were a prevailing force in most industries for more than 500 years. Guilds first appeared in antiquity but their heyday came about later in the late medieval period, and in some economies, they continued until the early modern period. Guilds' end came with the rise of national states as governments passed legislation that abolished them.

Premodern professionals formed guilds in almost all spheres of production and trade representing a large part of population. Nearly all urban areas of Europe, and sometimes rural areas, had industries controlled by these cooperatives. In most cases, there were guilds of manufacture producers. These organizations included traditional professions such as furriers, smiths, cobblers, or tailors but also nontraditional professions such as soap makers, dyers, and saddlers. Aside from these manufacturers, there were also widespread guilds of food producers. These included among others brewers, butchers, and bakers. Finally, there were guilds of servicemen, which included most notably merchants and retailers but also occupations such as barbers, painters, surgeons, and drivers. Besides single-craft guilds, some guilds combined affiliated professions in one organization creating multi-occupational guilds.

Although guilds could emerge in any industry, the number of guilds varied from city to city. Some towns had a relatively low density of guilds around one guild per 1000 inhabitants. Others had a relatively large density of guilds over five guilds per 1000 inhabitants. Also the number of craftsmen in each guild varied from a few men to over a 1000 (Lucassen and Prak 1998; Ogilvie 2014).

Activities of Craft Guilds

In medieval and early modern cities, the absence of traditional kinship networks and modern state institutions exposed the newly rising class of professionals to a variety of uncertainties typical to the premodern era. The response was to establish a new form of a network, one that was based on occupational affiliation rather than family ties, a craft guild. Indeed, despite the great variation between regions and even within the same city, the underlying purpose of the guild was everywhere the same: to secure stable standards of living for its members in the face of new risks brought as a result of market development. Craft guilds engaged in a variety of activities in order to achieve this goal.

Private Activities

What made craft guilds different from other forms of urban cooperatives and social networks was their emphasis on economic relationships. Most importantly, guilds laid down conditions for controlling trade and production within a particular geographic location (Ogilvie 2007, 2014). They held the monopoly over production in particular crafts as well as the monopsony over hiring workforce and purchasing inputs. Guilds' market monopolies were fortified by government privileges. Guilds regularly bargained with sovereign authorities over legal privileges, which guaranteed their members exclusive rights over local markets. These privileges enabled guild members to control economic activities of nonmembers within the city and its close neighborhood. Nonmembers were either excluded from trade and production within the particular location or their economic activities were regulated. In either case, guild members were clearly in a favorable position. By obtaining legal privileges, guilds provided their members with economic rents at the expense of nonmembers and the society in general.

However, guilds' concerns did not stop with obtaining economic rents. Guilds used their legal rights over local markets to reduce the transaction costs associated with premodern production and trade (Gustafsson 1987; Epstein 1998; Epstein and Prak 2008). In these respects, craft guilds were able to enforce contracts within their crafts, maintain reputation of their products and services, and provide skill transfers across time and places through a system of apprenticeship and journeymanship.

Furthermore, local experiences indicate that craft guilds were extensively active in areas of life other than economic (Epstein and Prak 2008; Lucassen et al. 2008; Richardson 2005; Prak et al. 2006). Many guilds engaged in the spiritual salvation of their members. The guilds' religious endeavors included organizing members' funerals, maintaining altars, lighting candles, and managing masses and prayers, all secured from collective resources. Guilds served as insurance networks. They provided resources and collective support to masters and their families in difficult situations, especially in sickness, widowhood, or old age. Guilds and their members participated in local politics. The guilds' influence, although limited, was

exhibited through their members, who regularly held positions in local and sometimes national governments.

Public Activities

Although craft guilds were primarily concerned with the interest of guild members, they were government-licensed organizations. Guilds' existence was directly linked with the privileges they obtained from local or national authorities. But authorities did not recognize guilds for free. In exchange for legal rights, governments demanded services from their recipients in return.

Before the state established strong public institutions, craft guilds provided means for civil administration (Hickson and Thompson 1991; Prak et al. 2006). Collection of fiscal revenues was among the most important functions. Guilds' knowledge of local conditions and expertise in craft fitted this function perfectly. Guild members were required to collect a variety of industrial and commercial taxes as well as provide loans and pay dues to their rulers.

Craft guilds also provided a number of civic tasks that benefited the public at large. Guilds' assemblies functioned as courts of the first instance in matters of their professions. Guilds also provided fire services. City charters demanded that craft guilds acquired their own fire equipment such as buckets, hooks, and water pumps and if necessary to engage in firefighting. Guild members also had to patrol inside city walls and finance military activities of their rulers. In some cases, guild members were even required to take part in the defense of their towns.

funerals, participated in local governance, patrolled on city streets, or pioneered fire services. Missing either of these aspects of guilds may lead to the wrong conclusions. Including private and public dimensions of guilds' activities into the picture is, therefore, important for understanding the role of these organizations in history.

References

- Epstein SR (1998) Craft guilds, apprenticeship, and technological change in preindustrial Europe. *J Econ Hist* 58(3):684–713
- Epstein SR, Prak MR (eds) (2008) *Guilds, innovation and the European economy, 1400–1800*. Cambridge University Press, Cambridge
- Gustafsson B (1987) The rise and economic behaviour of medieval craft guilds an economic-theoretical interpretation. *Scand Econ Hist Rev* 35(1):1–40
- Hickson CR, Thompson EA (1991) A new theory of guilds and European economic development. *Explor Econ Hist* 28(2):127–168
- Lucassen J, Prak MR (1998) Guilds and society in the Dutch Republic. In: Núñez CE (ed) *Guilds, economy and society*. Universidad de Sevilla, Sevilla, pp 63–78
- Lucassen J, de Moor T, van Zanden JL (eds) (2008) *The return of the guilds*. Vol. 53. *International Review of Social History*, Supplement 16
- Ogilvie S (2007) 'Whatever is, is right'? Economic institutions in pre-industrial Europe. *Econ Hist Rev* 60(4): 649–684
- Ogilvie S (2014) The economics of guilds. *J Econ Perspect* 28(4):169–192
- Prak MR, Lucassen J, Soly H, Lis C (eds) (2006) *Craft guilds in the early modern low countries: work, power and representation*. Ashgate, Aldershot
- Richardson G (2005) Craft guilds and Christianity in late-medieval England: a rational-choice analysis. *Ration Soc* 17(2):139–189

The Impact of Craft Guilds

Craft guilds were multifunctional organizations. Therefore, it is difficult to pin down their clear impact on society. The traditional literature argues that craft guilds had a mostly negative impact on the economy because of guilds' monopoly of trade and production. However, the new literature on craft guilds tries to moderate this view by showing that craft guilds also had a positive impact on the economy as they provided their members with social insurance, organized

Creative Commons and Culture

Joëlle Farchy¹ and Pierre-Carl Langlais²

¹Faculty of Economics, Pantheon-Sorbonne University Paris 1, Paris, France

²Université Paris IV – Sorbonne, Paris, France

Inspired by a specific philosophy, the “free” movement first emerged in the software and academic research fields and has since enjoyed

considerable economic success. Originally, the development of free software was in keeping with practical considerations. Computer specialist Richard Stallman, the “father” of free software, one day in the late 1970s became enraged at his rebellious desktop printer, because he was unable to access the program that controlled it or get it to respond to his commands. The real story begins in 1983, when he sent out a message announcing the creation of a comprehensive software package compatible with the proprietary system, UNIX. He intended to make it available to anyone who wished to use it. . . for free. This project was called GNU. In 1991, Linus Torvalds, a Finnish student, perfected Linux, the first version of an entire operating system based on GNU. The Free Software Foundation (FSF), a nonprofit association, was established in 1985 by Stallman to ensure the logistical structuring and financing of the “free” project, GNU, which then gave rise to GPL (General Public License). In the software field as in that of scientific publications, digital technology led to the emergence of systems combining intellectual property rights and extended tolerances for certain types of use.

In the field of culture, initiatives are born in a specific intellectual context, which we begin by briefly reviewing, and the revolutionary ambitions of pioneers have evolved with the development of Creative Commons licenses (1). These licenses, which originally were but one form of free license among many, have gradually gained a monopolistic position for numerous reasons, and in their 15 years of existence, their use has greatly increased (2). The purpose of this article is to make an initial assessment of the use of these licenses based on the scattered statistical data that is available. This appraisal provides an overview in terms of the works available (3), the varied practices of creators and consumers of works (4), and different economic models (5).

A Revolutionary Initial Goal

“Free culture” (the term was used by Lessig for the first time in 2004 but has its roots in earlier movements) is the heir to a double movement – a

community ideal characteristic of the digital utopias of the late 1960s and the sharing practices promoted by theoreticians of the commons. The convergence between these two movements occurred in several American university centers in the 1990s.

Digital utopias are a first breeding ground for thinking that emphasizes community ideals and self-regulation. Starting in the 1960s, certain American counter-culture movements began using digital technologies. Computers, until then more often associated with a public and private managerial planning, appeared as a vector of emancipation. New Communism, notably, based on the idea of small, self-managed communities, influenced an entire generation of entrepreneurs and intellectuals (Turner 2006) but did not, however, contradict the development of intellectual properties and use restrictions. In contrast, the principle of commons, based on Elinor Ostrom’s founding work (1990) on self-managed agricultural communities and initially far from digital technology, stressed the sharing of resources and collaborative governance. The 2009 Nobel Prize in Economy Winner highlighted the sustainability of forms of governance that are neither public nor private, where resource allocation is not based not on the individual attribution of property titles and collaborative management of community-pool resources determines the use conditions of these resources.

In the mid-1990s, Commons and digital technology combined and took shape within a small network of academic institutions. The Berkman Center (Harvard University) in particular brought together a number of actors, ideas, and movements closely linked to the sharing of works and information on the Internet: free software (Richard Stallman of MIT was a frequent visitor there), the decentralized networks theory (Yochai Benckler, a Harvard-educated professor at New York School of Law), a positive definition of the public domain (James Boyle, also a Harvard graduate and professor at Duke Law School), and legal analysis of the laws and uses of cyberspace (Lawrence Lessig). The desire to revolutionize, through practice, the copyright rules to facilitate the conditions of the sharing of works gave rise to a plethora of initiatives. The end of the 1990s

marked the golden age of free licenses in the cultural field; the vast majority of licenses were created between 1997 and 2001 (De Filippi and Ramade 2013).

In 2001, Creative Commons was but one free license among others in an already busy sector. Similar initiatives existed: Wikipedia adopted the GNU's GFDL; the Free Art license responded to national specificities in countries with a copyright tradition. On May 16, 2002, a press release announced the launch of a new organization, Creative Commons, in which American lawyer Lawrence Lessig would play a key role by contributing to the project's reorientation... and success. Under his leadership, one of the organization's initial key goals was to unite enough people and circles to establish itself as a cultural and social movement, and not just another free license among dozens of others. The radical opinions voiced by early activists gradually gave way to a more conciliatory approach so as to broader future coalitions.

Lessig (2001) developed a new understanding of culture as a common in the digital age. For the latter, the protection that intellectual property affords creators was not be called into question but rather balanced by user rights. Advocates of free licenses simply wish to see copyright used differently to promote the sharing and reuse of commons, meaning that although owners of copyrights retains the legal monopoly, they can use their power to authorize their use rather than forbid it. Licenses offer creators contractual tools that favor sharing and collective creation in the digital world. Free licenses may be used and shared openly, but only under particular terms and conditions, the contours of which must be determined by the individual rights holder who define the availability of their works and the terms of their use by way of "private ordering." Born under revolutionary auspices, Creative Commons gradually built a balance between the anarchic dissemination of works on the Internet and complete control. Creative Commons licenses are thus a revisited approach to intellectual property as a bundle of rights rather than a revolutionary system – in other words, a complement rather than an alternative to copyright.

The Gradual Emergence of the Creative Commons Monopoly

The success of Creative Commons licenses in the cultural field hinges on three strategic choices that have made them indispensable tools for expanding use rights: the implementation of a flexible, modular design, easy access visual symbols and the aim of adapting to national legal specificities despite their international scope.

A Flexible, Modular Design

In 2001, creators who wanted to share their productions could choose between several licenses ranging from total liberalization (WFTPL) to protection against enclosures (GFDL) to full retention of Moral rights (Free Art license). Most of these licenses were monolithic however, offering but a single legal framework to be taken or left. In contrast, Creative Commons offered a modular structure, leaving users free to choose the option best suited to their specific needs. While the core of the initial 2002 version remains unchanged, many adjustments have taken place over the years. In 2016, four options (allocation, identical sharing, noncommercial, and nonderivative) combined to create six separate licenses, to which the public domain was added (see Box 1).

Box 1 *Creative Commons* Licences in 2016



CC-BY (By Yourself): The

reuser shall give appropriate credit, provide a link to the license, and indicate if changes were made. This is a default option since 2004.



CC-BY-SA (*Share Alike*):

The reuser shall give appropriate credit and distribute the contributions under the [same license](#) as the original (in case of remix or transform of the original work).



CC-BY-NC

(*Noncommercial*): The reuser shall give
(continued)

Box 1 Creative Commons Licences in 2016

(continued)

appropriate credit and may not use the material for commercial purposes. The extent of “commercial use” remains currently fuzzy.

**CC-BY-ND**

(*Nonderivative*): The reuser shall give appropriate credit and may not distribute [any] modified material.

**CC-BY-NC-SA**

(*Noncommercial/Share Alike*): The reuser shall give appropriate credit, may not use the work for commercial purpose, and shall publish every derivative work under the same license.

**CC-BY-NC-ND**

(*Noncommercial/Nonderivative*): The reuser shall give appropriate credit, may not use the work for commercial purpose, and may not distribute any modified material.

**CC0 (Public Domain):** The

work is dedicated to the public domain by waiving all rights to the work worldwide under copyright law. The license is especially used by database as it allows an effective removal of sui generis database right.

that is easily understood by nonprofessionals (a kind of summary).

- Finally, based on this simplified webpage, users can access the detailed terms of authorization in the event of legal dispute.

International Goals Through Adaptation to National Legal Frameworks

While most licenses were exclusively designed for American legislation, internationalization was a priority in the creation of Creative Commons. Less than 4 months after the official announcement of its creation, the organization set up an office in Berlin to coordinate volunteer efforts to develop national versions. This was facilitated by preexisting transnational networks within the legal community, with 71% of its members from legal organizations (Dobusch and Quack 2008). In 2004, the license was transposed in 12 countries. The internationalization process continued at a steady pace until the end of the 2000s before slowing down. In 2016, the Creative Commons movement had representatives in 79 countries, with licenses adapted to 60 legislations. The complexity of the transposition process thus supplanted the original, nearly unattainable goal of a single, universal license.

Fifteen years after the first version (the project, launched in 2001, culminated in the publication of the first usable license and the organization’s launch in 2002), CC licenses have crowded out most other free and acquired licenses in the cultural field, attaining a virtually monopolistic position.

Three Levels of Accessibility

Creative Commons offer not only a choice of options but also of different levels of access. In contrast to the relatively heavy terms of other licenses, which require a full reinterpretation of the specialized legal text, CCs are divided into three levels of accessibility (Loren 2007):

- At the level of content itself (and its reworking and reuse). A small set of codified symbols indicates which option was chosen.
- These icons are linked to a simplified webpage that presents the terms of the license in a way

A Large Offer Concentrated on Certain Platforms and for Certain Types of Content

In absence of an internal indexing system, statistical data on works, creators, and consumers of works is still quite patchy. Three types of sources can be used:

- Internal sources within the organization (such as State of the Commons or the Metrics Project). Since 2014, Creative Commons has

published a “State of the Commons” regarding uses on the main platforms (Flickr, YouTube, and Wikimedia), supplemented by a breakdown of off-platform web pages that use these licenses on Google (317 million pages in 2015). Previously, several statistical projects were done, including the article by Cheliotis et al. (2007) which marked the starting point.

- The data collected by certain Web actors, companies (e.g., Flickr and YouTube) or communities (e.g., Wikimedia Commons or Wikipedia).
- Occasional assessments from the scientific literature.

In the reviews conducted by the Creative Commons organization as well as researchers, the work is systematically equated to a “web page” with a valid link to a license. This semantic shift, though it considerably facilitates statistical collection, is nonetheless a simplification. Under this reserve, by 2015, there would have been 1.1 billion of what we will call “publications” under Creative Commons circulating on the web, a third on Flickr alone (350 million) and about 15% on various Wikimedia projects (140 million Wikipedia pages and 20 million images on Wikimedia Commons, the image library of Wikimedia projects).

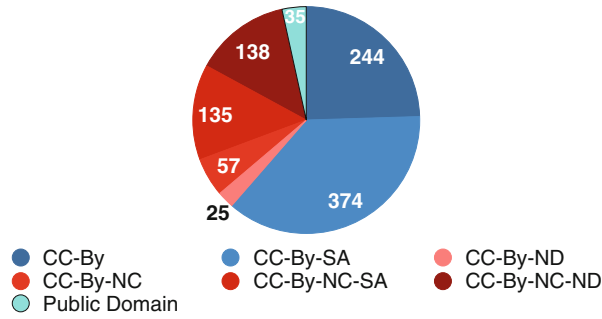
Two periods stand out: a rapid take-off starting in 2004 (and particularly in 2005) and a period of stable growth after 2010. In January 2007, there were 2.5 million publications (Kim 2007); 10 million by mid-February, 14 million by March, and 53 million by September. The end of the period covered by the Metrics Project on the contrary shows a relative stabilization, with 357 million publications in 2010 and 422 million in May 2011.

Although the movement is part of a widely internationalized approach, as Cheliotis et al. (2007) note, the use of CC tools reveals strong national disparities in the 33 countries where the licenses have been adapted. Economic development is the first factor that explains these differences; the magnitude of the digital divide has led to underrepresentation of China and South American countries compared to European countries and

certain “highly-connected” Asian countries, such as Taiwan or South Korea. Social and cultural factors also play a role. For Cheliotis et al. (2007), the countries with the most widely distributed licenses and the most open options have a common profile: they are developed countries with strong digital infrastructures where the sharing of works, legal or illegal, is socially accepted (e.g., Sweden and Spain). In contrast, France, Belgium, and Italy form a kind of “Latin cluster,” with a combination of strong license growth and more restrictive options. More recently, Latonero and Sinnreich (2014) found that growth was weaker in the United States than in Europe. The data collected in the 2015 edition of the State of the Commons on 317 million web pages also highlights the increasing internationalization of licenses: some countries in South America and Asia, such as Brazil and Thailand, have growth rates comparable to those in Europe. However, the distinctive characteristics of certain countries remain; Hungary, for example, has roughly one web publication per capita on Google (off-platform) under Creative Commons, versus only one in ten in France.

In the vast collection of publications under CC that exists today worldwide, certain publishing platforms clearly stand out: nine platforms publish 72% of the pages with link to a Creative Commons license, and only three of them, Flickr, Wikipedia, and Libre.fm, have 55% of all pages. The overall share of the eight platforms already enumerated in 2014 increased from 66% to 68% in 2015 (State of the Commons 2015).

The breakdown of licenses by type of activity also provides interesting information, with (in decreasing order) photographs (39 million), articles, stories and documents (47 million), videos (18 million), audio recordings (4 million), scientific articles (1.4 million), open educational resources (76,000), and an “other” heading (multimedia, 3D, etc.) of 23,000 (State of the Commons 2015). While the new growth drivers tend to be noncultural productions, scientific or educational, most notably, photographs and musical recordings dominate among cultural productions. The use of Creative Commons for the production of videos is only in its infancy (the CC-By license, for example, was only introduced



Creative Commons and Culture, Fig. 1 Breakdown of licenses in millions of publications. The most open licenses are indicated in shades of *blue* (From public domain to CC-BY-SA). The most “closed” licenses are indicated in shades of *red* (From CC-BY-NC-ND to CC-BY-SA) (Source: Data from State of the commons 2015)

on YouTube in 2011) and has not yet been studied specifically.

In the field of photography, after a rapid take-off on Flickr between 2004 and 2006 (26 million photographs at the end of 2006 versus 300,000 in 2004), the number of new photographs under CC reached a plateau in late 2008 (39 million). Figure 1 remained steady at the end of 2009 and 2010 and then fluctuated sharply (24 million at the end of 2014 and 48 million at the end of 2015).

Wechsler (2010) observed the same pattern for music platforms. On Soundclick and Jamendo, initial growth was rapid until 2005–2006 and then stabilized. Certain music categories are also better represented: electronic music authors and classical music performers, respectively, represented 28% and 26% of audio publications under Creative Commons (Wechsler 2010). Concerning electronic music, a similar phenomenon was observed on Jamendo (Bazen et al. 2014). Two complementary explanations are put forth by the researchers. According to Wechsler, these two niche markets have fairly weak marketing prospects compared to rap or pop, for instance. Bazen, Bouvard, and Zimmermann insist, however, on the highly digitized production conditions of electronic music to explain this phenomenon.

Thus, the use of Creative Commons often reflects shared affinities or reluctance within a community of creators who, each in their own way, make strategic choices by opting for Creative Commons licenses and certain options more specifically.

A Continuum of Practices between Active and Passive Users

The very concept of users and free license users raises questions. A user is at once the “creator” who proposes original or derivative works under a CC license, the “consumer” who uses them or the platform that hosts them. However, few studies address these concepts in all their aspects.

Creators: To Each Their Choice

The studies available provide some information about creators who make their works available under Creative Commons licenses – a practice that is still marginal. For Latonero and Sinnreich (2014), only 5.3% of Internet users in the United States posted content under CC licenses. The overall figure for Internet users for the eleven other countries studied (the notion of contribution may, in this case, refer to creators of cultural works, wiki project contributors, etc.) was 17%. In a poll for Creative Commons (2009), only 6% of all American creators of original web publications had published content under CC in the last 12 months; among all American creators of web publications derived from original publications, only 3.3% of reworked publications had been done under CC in the last 12 months. The term user in this case does not refer to consumers of publications but to those who use them creatively (remixing, sampling, contribution. See the methodology presented in Creative Commons (2009, pp. 24–25)). Though being a minority in terms of the penetration rate among contributors, licenses

play an important role in some of them by meeting specific and changing needs.

Several studies emphasize the fact that creators of content under Creative Commons fall into a new intermediate category between amateurs and professionals. Wechsler (2010) notes that creators of music under CC are significantly less motivated by commercial gains than by social more social or political incentives (sharing of works, commitment to a change in intellectual property legislation, etc.).

Bazen et al. (2014) draw attention to close ties between amateurs and professionals on the Jamendo music platform. Here the majority of creators are amateurs, with professional musicians representing albeit a nonnegligible minority (22% among music groups and 18.5% among solo artists), contrary to the myth that Creative Commons only concern those who practice art as a hobby. The use of licenses reflects neither the dissolution of professional practices nor to the formalization of amateur practices, but rather an unprecedented equilibrium marked by the relaxing of the strict boundaries between “amateur” and “professional” (Bazen et al. 2014).

Beyond the hybridization of professional and amateur practices, the choices creators make from the different options available to them among CC licenses are not without consequence.

To clarify the more or less extensive nature of the various usage rights, a first indicator distinguishes the most open licenses, which allow for modification and commercial use, from the others. The ratio of “open” licenses to “closed” ones (noncommercial, no change possible) has increased considerably: according to internal Metrics data, between 2003 and 2010, after initial phase of fluctuation of 20% and 30%, the ratio jumped to 40% in 2010. The State of the Commons, which took over starting in 2014, shows that the most open licenses accounted for 56% of publications in 2014 and 65% in 2015. CC-By-SA and CC-By predominated here (with 374 and 244 million works in 2015, respectively). By also integrating works in the public domain, licenses allowing for commercial reuse and modifications thus represented almost two-thirds of the works identified (653 million out of 1 billion),

a sharp contrast to a decade ago when the most open licenses were but a small minority. In the two cultural sectors where Creative Commons licenses are used most – photography and music – the preference today is for using the most open licenses.

In addition, Cheliotis (2009) notes recurring segmentation between two extremes: the first is characterized by the maintenance of purely artistic control over works through the use of noncommercial licenses, the second by a more altruistic relinquishing of rights (CC-By and CC-By-SA license), including for commercial uses. Bazen et al. (2014) also identified a divide within the population studied: a good quarter of creators intend to put up as few barriers as possible to the circulation of their work (BY), while half use more specific strategies. Cheliotis explains this divide as being driven by motives that are more pragmatic than ideological, given that creators with a low expectations with regard to the market are those most willing to opt for more open licenses. Wechsler (2010) reaches similar conclusions: among the creators interviewed, no self-proclaimed professional recommended the use of licenses that authorize commercial use. This population is therefore not representative of all Internet users who publish under Creative Commons licenses, nonprofessionals being more likely to use open licenses.

Beyond these divisions, a single creator often moves from one sphere to another depending on the situation: 33% of creators with more than one album to their credit on Jamendo used several CC licensing options (Wechsler 2010). Similarly, 50% of Soundclick’s contributors publish their works based on classic copyright rules or using the various Creative Commons licenses depending on the case.

The Audience: An Unknown Variable

Most surveys on Creative Commons users focus on creators, which may seem paradoxical given that one of the objectives of these licenses is precisely to increase the diffusion of works. There is no evaluation/assessment of the overall audience of works under Creative Commons. In 2013, Wikimedia projects reached the

500 million mark for one-time visitors (per month), and Flickr the tens of millions (14 million for the United State alone). The number of one-time visitors is no longer calculated due to the audience's switchover to mobile devices (and the lack of identification measures to protect users' privacy).

These figures offer us a scale of magnitude but do not allow us to calculate a total audience, given the considerable overlap among one-time visitors between the sites and the fact that content under Creative Commons is designed to circulate freely. Google search results often include extracts from Wikipedia; photographs from Wikimedia Commons and Flickr are often used on press sites. A somewhat active user is or will be regularly exposed to content under Creative Commons without necessarily even knowing it.

Latonero and Sinnreich (2014) consider a specific audience segment, that of users who do targeted searches for content under Creative Commons or in the public domain, implying that they know of the existence of these licenses and their implications. The survey showed that 13% of Internet users in the United States belong to this "insider" audience segment, and that this figure jumps to 31% on average in the 11 other countries studied. Another study (Creative Commons 2009) focuses on an even more active and informed segment – that of users who not only consult works but use them to create new works (remixes, articles, etc.). A survey conducted by Greenfield for Creative Commons shows that 6% of respondents used them this way. In keeping with the hypothesis of cultural remix defended by Lawrence Lessig (2009), the population of active users overlaps with the American creator population also surveyed by Greenfield: only 28% had never published under Creative Commons licenses. The Creative Commons audience is thus a concentric circle with different levels of initiation in terms of licenses and different levels of use (from passive use to recreation).

The various licensing options are thus as many potential models of dissemination to suit the different individuals' pragmatic expectations.

Different Economic Models for Diverse Uses

The idea of a new form of collaboration between economic activities based on free culture has been widely discussed in the literature. Using software programs as support for his idea, Yochai Benkler (2002) theorized a cooperative alternative system called "commons-based peer production" specifically for the digital world. This model is a complement to those established since Ronald Coase's, the contract and the market. In this model, groups of individuals successfully collaborate on large-scale projects by following signals that are neither price- nor hierarchically-based. When it comes to producing information, this mode of production offers systematic advantages over the other two models; even though individuals do not directly reap the benefits of their participation in the collective project, their efforts have a greater impact than they would on the marketplace or in the firm, as commons-based peer production serves both to identify people best-suited to any given aspect of a project and to allocate resources to those able to put them to the best use according to optimal matching logic.

Beyond purely cooperative models based on volunteerism and free contributions, we find within the "free" world different business models of commercial service activity. Foray et al. (2006) underline that an understanding of the economics of open source should not be based on the marked dichotomy between analyses in terms of intrinsic and extrinsic motives, or in terms of community versus market-based economy. Rather, contemporary models of production and distribution of information products are hybrids models. "The hybrid economy "will dominate the architecture for commerce on the Web. It will also radically change the way sharing economies function. The hybrid is either a commercial entity that aims to leverage value from a sharing economy, or it is a sharing economy that builds a commercial entity to better support its sharing aims" (Lessig 2009, p. 177).

Market enterprises relying mainly on Creative Commons have had limited thus far. "The record labels supporting CC licenses are niche players. . .

Except for limited experiments, CC licenses have not yet been adopted by independent or major labels” (Wechsler (2010, p. 122)). However, a hybrid economy has been created to support and/or complement business models in the digital economy. Digital technology has effectively led researchers to identify new business models that are better suited to the new conditions of production and distribution of information goods, particularly given that copyright laws are becoming almost impossible to enforce (Varian 2005). Two-sided market models based on advertising revenues occupy a key position in this economy. We find them in organizations that use CC licenses: 50% of Jamendo’s advertising revenues (by number of pages viewed) were paid back to creators (Russi 2011). YouTube allows users to enhance content available under CC-By by associating it with advertisements. However, the perception of advertising revenue for covers is still complex: most intermediaries that register titles in the identification program of YouTube works (ContentID) do not accept works under free licenses: “Many of the online music distribution company. . . [do not accept them] into their distribution with YouTube’s Content ID system.” Although they sometimes exist, two-sided market models do not dominate in the CC license economy the way they do in the digital economy overall.

The literature on the specific subject of CC licenses reveals a multitude of economic models. To the extent that CC license users are both the creators and final consumers of works and platforms (see *supra*), whose aspirations are inherently heterogeneous, we propose three types of models that in each case highlight the choices of these three main categories of economic agents. However, these are only ideal types, whereas any given organization or economic agent typically combines the various models.

Fueling Celebrity Capital: The Choice of Still-Unknown Creators or Fan Communities

As mentioned above, CC licenses are widely used by a new intermediate category somewhere between amateurs and professionals. For many of them, greater publicity is more important than

direct remuneration because the cultural economy is, in fact, an economy of brand recognition and reputation-building. A survey of artists who have opted for CC licenses (Wechsler 2010) shows that they often use them as a launch pad to create a buzz around their work: “Releasing music under a CC license is another way to get more publicity. By legalizing sharing, artists enable their fans to promote their music” (p. 129). “CC licenses may increase the diffusion of music and boost artists’ reputations. These benefits may then be monetized in the form of more concerts and/or increased donations from fans” (p. 189).

CC are particularly suited to nonprofessionals who wish to make their work known, with little or no expectation of remuneration, and artists aspiring to become professionals. In this economy, symbolic remuneration – pride in collective participation within the community and recognition by peers – is strong. Becoming part a professional circuit can likewise be a powerful incentive for contributors, as drawing the attention of producers, employers, and public financial institutions in the hopes of funding future creations has become essential in an economy rich in cultural goods supply.

Communities of fans also form on the basis of sharing and celebrity. The Wikia website, created by Wikipedia cofounder Jimmy Wales, has become a major documentary reference in cultural industries. Collaborative encyclopedias fed by fans exhaustively cover certain fictional worlds, like that in Star Wars (160,000 articles put as on Wikipedia CC-By-SA license), and help build its reputation.

Harnessing End Users’ Willingness to Pay

In so far as diffusion under CCs does not necessarily mean free access to works, a monetary contribution from end users may be requested. Certain forms of funding such as crowdfunding, which is based on prepurchase by consumers, are little used. Erickson et al. (2015) estimate the percentage of content under CC licenses on these platforms at 1%. End users of content under CC licenses are solicited mainly through voluntary donations or purchases of additional freedoms. On Magnatune and Jamendo, consumers can buy

music albums for the price they wish based on a “Pay what you want” formula (Russi 2011). For Wechsler (2010), consumers would be willing to pay an additional fee for works under Creative Commons licenses if they had the possibility of modifying them, for example. On the Beatpick platform, works are sold without DRM (Russi 2011). Others, like those devices used in the free software economy, for an additional fee offer the possibility of more open licenses by removing certain constraints, such as the obligation to use the same license for derivative works. In 2007, the Creative Commons organization designed a specific mechanism for managing the combining of several licenses: CCPlus, for example, is often used in the field of musical creation. A special symbol indicates that content users have the right to use content under a more open license (for example, allowing for commercial reuse): “The architecture of the CC-Plus scheme enables a commercial economy to co-exist with, or be grafted onto, the sharing economy created by the Creative Commons system” (Russi 2011, p. 106).

Adhering to the Values of a Community: The Choice of Institutions More So than that of Individuals

For Himanen (2001), contributors’ motives for choosing free licenses hinge neither on a restrictive hierarchical structure nor on monetary incentives, but rather are based on a passionate relationship to works, flexibility of scheduling, appreciation for their “free” nature, and symbolic remuneration. Individual commitment is driven by an appreciation for the notion of give and take. In this model, there is no compulsory direct remuneration but rather remuneration through emotional dynamics and a relationship with the larger community.

Economic motives are sometimes secondary for creators. On Jamendo, although 22% chose CCs because the platform requires it, 20% did so because they saw them as a way to create a buzz, and more than 60% adhered to the idea of sharing that CC licenses promote (Bazen et al. 2014, p. 19). Furthermore, CC are particularly well adapted to scalable works, which are by nature perfectible, and can combine the works of different

contributors. They were largely designed to facilitate the job for those who reuse and adapt the earlier versions of works of their colleagues more so than to market those of the original authors.

Beyond the individual choices of creators and/or end users, the use of CC licenses is by and large a choice made by organizations like Wikipedia and Flickr, which host, as we have seen, the vast majority of publications under CC license (see Supra) and impose more than propose the use of free licenses for works available on their sites. The hybrid models somewhere between market and cooperative coordination that are the strength of these organizations can also become Achilles’s heel of this free culture economy by forcing the cohabitation of demands sometimes too contradictory and communities sometimes too different. As Lessig predicted: “No one builds hybrids on community sacrifice. Their value comes from giving members of the community what they want in a way that also gives the community something it needs” (Lessig 2009, pp. 177–178).

References

- Bazen S, Bouvard L, Zimmermann J-B (2014) Jamendo et les Artistes: Un nouveau Modèle pour l’Industrie Musicale?
- Benkler Y (2002) Coase’s penguin or Linux and the nature of the firm. *Yale Law J* 112(Winter):369–446
- Cheliotis G (2009) From open source to open content: organization, licensing and decision processes in open cultural production. *Decis Support Syst* 47(3):229–244
- Cheliotis G, Chik W, Ankit G, Tayi KG (2007) Taking stock of the creative commons experiment monitoring the use of creative commons licenses and evaluating its implications for the future of creative commons and for copyright law. Singapore Management University
- Creative Commons (2009) Defining “noncommercial” (Unsigned collective report)
- Creative Commons (2015) State of the commons. <https://stateof.creativecommons.org/2015/data.html>
- De Filippi P & Ramade I (2013) Les licences Creative Commons: Libre Choix ou Choix du Libre? In Christophe Masutti Camille Paloque-Berges Benjamin Jean (dir.), *Histoire et cultures du Libre*, Framabook, 2013, pp 341–378
- Dobusch L, Quack S (2008) Epistemic communities and social movements: transnational dynamics in the case of creative commons. Social Science Research Network, Rochester

- Erickson K, Heald PJ, Homberg F, Kretschmer M & Mendis D (2015) Copyright and the value of the public domain: an empirical assessment. Report commissioned by the Intellectual Property Office. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/561543/Copyright-and-the-public-domain.pdf
- Foray D, Thoron S, Zimmermann J-B (2006) Open software: knowledge openness and cooperation in cyberspace. In: Brousseau E, Curien N (eds) *Internet and digital economics*. Cambridge University Press
- Himanen P (2001) *L'éthique hacker et l'esprit de l'ère de l'information*. Exils, Paris
- Kim M (2007) The creative commons and copyright protection in the digital era: uses of creative commons licenses. *J Comput-Mediat Commun* 13(1):187–209
- Latonero M, Sinnreich A (2014) Tracking configurable culture from the margins to the mainstream. *J Comput-Mediat Commun* 19(4):798–823
- Lessig L (2009) *Remix: making art and commerce thrive in the hybrid economy*. Penguin Books, New York
- Lessig L (2001) *The future of ideas: the fate of the commons in a connected world*. Random House, New York
- Loren LP (2007) Building a reliable semicommons of creative works: enforcement of creative commons licenses and limited abandonment of copyright. *George Mason L Rev* 14:271–328
- Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge/New York
- Russi G (2011) Creative commons, CC-plus, and hybrid intermediaries: a Stakeholder's perspective. *Int Manag Rev* 7:103
- Turner F (2006) From counterculture to cyberculture – Stewart brand, the whole earth network, and the rise of digital utopianism. University of Chicago Press, Chicago
- Varian HR (2005) Copying and copyright. *J Econ Prospect* 19(2):121–138
- Wechsler J (2010) *Openness in the music business – how record labels and artists may profit from reducing control*. PhD dissertation, Technical University of Munich

Creativity

Christian Barrère

Faculty of Economics, Laboratoire R.E.G.A.R.D.S,
University of Reims, Reims, France
ISMEA, Paris, France

Abstract

First of all creativity approach considers the creative behavior of judges and jurists in the

legal processes. Afterwards it studies the specificities of legal tools, mainly IPR, regarding creative products.

Definition

From the 1960s, cognitive sciences and economics developed a new paradigm centered on creativity which no more dedicated to a marginal role, mainly in the artistic field, but appearing as a key component of human thinking and of social activity. For economists creativity is related to two conditions:

- It is a mental ability to create, i.e., to introduce a new thing (opus, idea, representation, etc.) where previously there was nothing similar (Sternberg 2006), as the image of God's creation.
- Goods based on creativity are mainly produced by the human brain, an idiosyncratic input, and do not result from the use of generic and standard economic resources as energy, equipment, and labor.

This new paradigm has two main consequences on the development of law and economics.

Firstly, it leads to change the basic model of the economic agent, from the standard homo oeconomicus, using his substantive rationality, to a creative person. That raises the issue of creativity in the making of law and in the evolution of legal systems. What are the ways according to which law is evolving and what is the role of creativity in this process? Is it playing differently in the common law and the civil law systems? Can the judicial decision be a creative one? Under what circumstances and within which limits?

Secondly, the creativity paradigm focuses on the evolution from an industrial society, based on the use of natural and human energy, to a creative economy based on creativity, on the use of the human brain and imagination (Potts 2011). The Lisbon Strategy, defined in 2007, decided to make European Union the most creative economy in the world. Thus, creativity has a strong value implying protection. Moreover, creativity appears as cumulative and noncumulative knowledge. Creativity close to science works as an input for the

production of new knowledge that is more developed, so that the old knowledge melts, over time, into the new one, which ceases to have a value as such. On the other hand, creativity close to arts gives noncumulative products that will not be replaced by better quality amenities in the future (Picasso is not a “better” painter than Poussin and Botticelli) and constitute creative heritages which are not fungible unlike the heritages of cumulative knowledge. So their value hugely matters. Thus, the protection of the value of creativity is a new key issue for economic development. Nevertheless, the specificities of creativity challenge the standard system of intellectual and industrial property rights:

- *Defining the entitlement.* The first problem in defining property rights is to identify all the resources (present and past including heritage), which currently have creative effects or can produce some effects in the future, to identify all the producers of resources, to separate their contributions, and to distribute rights among them so as to give each producer exclusive rights in their resources and control over the effect of their creative contribution. All that is often very difficult because creative production is frequently a team production and uses new and old creativity accumulated in heritages.
- *Organizing a market for IPRs.* In order to define property rights that can be transferred through a market process, resources have to be evaluated. But non-separability and nonadditivity, the dominance of creativity, and the difficulties of measurability disrupt the evaluation of the effects of the resources. Idiosyncrasy is a characteristic of creativity that makes it even more difficult to infer the value of used resources from the value of their effects.
- *Enforcing IPRs.* The third problem is to enforce the definition, entitlement, and transfer of property rights. Piracy and opportunistic behavior result from difficulties in identifying resources and in entitling them in order to define exclusive rights.
- *Justifying IPRs.* The difficulties to clearly define and value the productive resources and their holders imply difficulties to justify the present

distribution of property rights. Normative problems arise. Is it fair to give the main part to the creator? Or to the owner of the firm? As usual in the field of intellectual property rights, the distribution of monetary earnings is far from contributing to human happiness or social development – is it fair that Einstein’s income had been so small compared to that of Bill Gates?

Cross-References

► [Innovation](#)

References

- Potts J (2011) Creative industries and economic evolution. Edward Elgar, Cheltenham
- Sternberg RJ (2006) The nature of creativity. *Creat Res J* 18(1):87–98
- Further Reading**
- Barrère C, Delabruyère S (2011) Intellectual property rights on creativity and heritage: the case of the fashion industry. *Eur J Law Econ* 32(3):305–339
- Benghozi PJ, Santagata W (2001) Market piracy in the design-based industry: economics and policy regulation. *Econ Appl LIV*(3):121–148
- Cohen JE (2007) Creativity and culture in copyright theory. *UC Davis Law Rev* 40:1151–1205
- Fauchart E, von Hippel E (2008) Norms-based intellectual property systems: the case of French chefs. *Organ Sci* 19(2):187–201
- Madison MJ, Frischmann BM, Strandburg KJ (2010) Constructing cultural commons in the cultural environment. *Cornell Law Rev* 95:657–710
- Towse R (2001) Creativity, incentive, and reward: an economic analysis of copyright and culture in the information age. Edward Elgar, Cheltenham

Credibility

Jeong-Yoo Kim
Economics, Kyung Hee University, Seoul,
Republic of Korea

Abstract

Two kinds of credibility will be distinguished, depending on the source of information. The

first credibility, which will be called type I credibility, is regarding information about the player's intention and the second credibility, which will be called type II credibility, is regarding information about the player's hidden characteristic. The former problem of credibility occurs when a player makes a promise or a threat (in a certain period of time) that he will do something in the future and then something happens between the two points of time. Then, the promise or the threat may not be credible, since the player will not have the incentive to keep it any more. The latter occurs when the informed player tries to pretend to be a better type by exploiting the uninformativeness of the other player. Then, the message conveying his information may not be credible as long as the incentive to pretend exists.

Definitions

Cheap talk	A costless, nonbinding, and nonverifiable message of a player.
Credible threat	A threat is credible when a player would find it in his interest to carry out the threat whenever he is called upon to do so.
Discovery	The legal process by which civil litigants are entitled to demand relevant evidence from each other.
Empty threat	A threat is empty if it is not credible.
Private information	Information possessed by fewer than all the players in a game.
Separating equilibrium	An equilibrium in which each possible type chooses a different action so that the type is revealed ex post.
Signaling	Choices by those who possess private information that convey information.
Subgame	Any part of a game that meets the following three conditions: (i) it starts from a single decision node, (ii) it includes all the nodes that follow the node, and (iii) if it

includes a node in an information set, it also includes all other nodes in the information set.

Type

Hidden characteristic of a player.

Introduction

Information is widely dispersed throughout the economy, and its distribution is quite heterogeneous among economic agents. No one knows everything about the economy. Some agents may know better in a field, but others may know better in other fields. As such, there are informed players and uninformed players about almost anything in social interactions. Then, uninformed players may seek the information they need from the informed players, but the informed ones do not necessarily have an incentive to provide true information to the uninformed ones. For example, a mechanic may quote a fairly high price for a repair even though he found the problem only minor. A doctor may recommend his patients unnecessary medical tests and treatments. Presidential candidates often try to rope in votes by churning out meaningless pledges that are hardly feasible. Besides, it is not clear whether it is socially desirable always to be honest. As a typical example, it is controversial whether it is better to tell a patient of a cancer diagnosis. This draft addresses the issue of credibility, that is, when the messages of informed players are credible, when it is socially desirable to be credible, and what kind of legal and institutional devices are needed if they do not match, i.e., if informed players have an incentive not to be credible even if social efficiency requires credibility.

In this entry, I distinguish two kinds of credibility, depending on the source of information. The first credibility is regarding information about the player's intention (future choice) and the second credibility is regarding information about the player's hidden characteristic. The former is called the issue of dynamic inconsistency in literature. It occurs when a player makes a promise or a threat (in a certain period of time) that he will do something in the future and then something happens (new information is realized or the other player makes some decision) between the

two points of time. Then, the promise or the threat may not be credible, since the player will not have the incentive to keep it any more. The latter is called the issue of adverse selection. It occurs when the informed player tries to pretend to be a better type by exploiting the uninformativeness of the other player. Then, the message conveying his information may not be credible as long as the incentive to pretend exists. I will call the former type I credibility and the latter type II credibility.

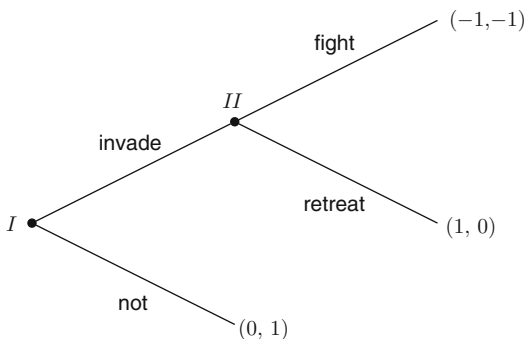
This entry is organized as follows. In the next section, I explain the issue of credibility more formally by using the terminology of game theory and examples in general economics. In section “[Applications to Law and Economics](#),” I provide its applications to law and economics. Section “[Conclusion](#)” contains some future research directions and concluding remarks.

Where Credibility Comes From?

Consider the following simple game (Fig. 1). Two countries are in a war. Country 1 has two options: invade (I) or not (N). If country 1 invades, two options are available to country 2: either fight against it (F) or retreat (R). If country 1 does not invade, however, the game ends. As is well known, the game has two Nash equilibria (NE): (I, R) and (N, F). In particular, the second NE is noteworthy. “Not invade (N)” is optimal for country 1 given country 2 chooses to play F, since his invasion would entail “fighting (F),” whereas his opposite choice simply ends the game. In other words, country 1 finds invading against his

interest because of country 2’s threat of fighting. However, this strategy F involving the threat of fighting is never credible, although it constitutes NE. It could be credible if this game is played via a third-party referee who collects the strategies that each player submits to him and plays out the strategies by dictating them to play according to their submissions. In this scenario, once a player submits his strategy, say F, it is a commitment. He could not change it later, because the referee plays out the game. In this dynamic game, however, country 2 has a chance to change his strategy after the game begins. He can rethink at his decision node and change his mind after country 1 invades. Obviously, country 2 will prefer R to F, once country 1 invades. In other words, the strategy F which is a threat to fight if country 1 invades is not credible and thus called an “empty threat.” The ex post optimal choice for country 2 when the other has already invaded is different from his ex ante optimal choice, because country 1’s invasion has changed the situation. What country 2 should care about once invasion has occurred is a small subgame starting from his decision node, not the whole game. This dynamic inconsistency problem is quite universal. In dynamic games, a NE may involve a strategy which is not credible. Generally, a threat is credible if and only if it is a rational choice in every subgame, implying that it must survive backward induction. This argument is due to Selten (1965). Is there any way to make an empty threat credible? One could think of various commitment devices. One well-known example is to eliminate an option, for example, by burning the bridge (leading to the retreat).

For the issue of the second kind of credibility, consider a game situation in which a player possesses private information. This is called a game of incomplete information. Suppose the informed player sends a message (regarding his private information) to the uninformed player and then the uninformed player responds to it. This special game of incomplete information is called a signaling game. In a signaling game, the informed player moves first, so his choice can be a signal of his private information, and the uninformed player may be able to infer the unobservable



Credibility, Fig. 1 Type I credibility

information from observing the signal. Can the signal of the informed player be credible? Can it convey meaningful message regarding his private information? Of course, it can, if the signaling costs differ across types (i.e., values of private information). For instance, suppose that a firm wants to recruit workers of higher ability and more able workers can be educated at lower costs. Since a more able worker will try to signal his high ability by choosing to be more educated but a less able worker cannot do so, a higher education can be a credible signal for higher ability. This argument is due to Spence (1973). On the other hand, plain talks such as “I am a very able person” or “I am not able” involve no signaling cost so that the signaling costs of the two messages cannot differ for any type. So, the next question is whether costless signals such as talks can be credible at all. If the messages under consideration are “I am able” and “I am not able,” the message of “I am able” can never be credible, because any worker would prefer this message to the other message, as long as both messages are equally cheap (costless) and both of them are taken seriously. However, what if the values of private information are not vertical but horizontal in the sense that the private information involves two kinds of ability that need nice matching between a firm and a worker, for instance, academic work and gardening? In this case, both a firm and a worker want good matching so that there are common interests. Then, a worker will not have an incentive to lie, even if academic work is more respected. For example, in a game illustrated in Fig. 2, the informed player who is good only at gardening will not say to the firm “I apply for academic work.” Generally, if players have

common interests, even cheap talk messages can be credible. This argument is due to Crawford and Sobel (1982).

Applications to Law and Economics

The issue of credibility can occur in many legal situations in which a (potential) defendant (D) and a (potential) plaintiff (P) are involved with strategic interaction sequentially, but I focus only on civil/criminal procedures. I will use the following notation:

w = P's damage amount

q = D's liability (=P's winning probability at court)

c_p = P's litigation cost

c_d = D's litigation cost

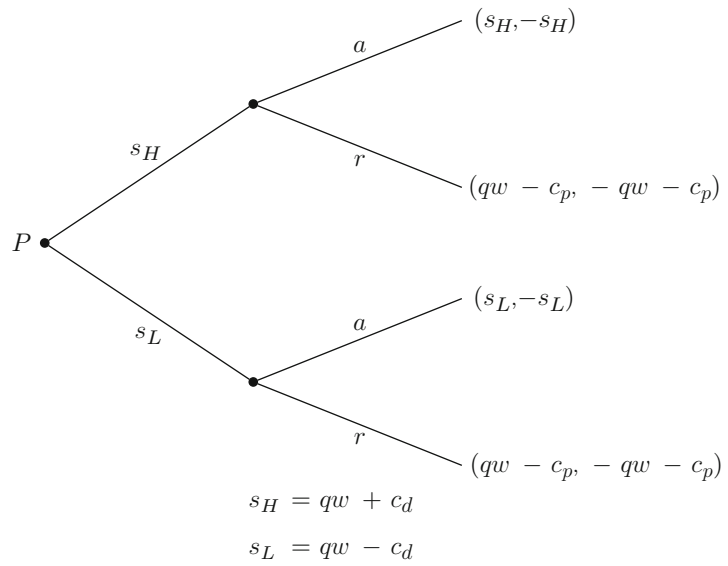
Model 1 Consider a game illustrated in Fig. 3. This is a simple game of pretrial negotiation. P has only two options: either high settlement demand ($s_H = qw + c_d$) or low settlement demand ($s_L = qw - c_p$). If a settlement demand is rejected, they go to trial, causing each party to bear their respective litigation costs. This game has many NE. For instance, it is a NE that P makes a low demand s_L due to D's threat to reject any demand higher than s_L . In this NE, P makes a low demand s_L which is accepted by D for sure. However, D's threat is not credible, because D will accept a high demand any how if it is assumed that indifference in payoffs is resolved in favor of settlement. Thus, the only sensible outcome is that P makes a high demand which is accepted by D for sure. This is an example of type I credibility.

Model 2 The game provided in Fig. 4 is a modification of Fig. 3. Now, P's damage amount is his private information and is either w or 0. If the damage amount is 0 or more generally very low, the case is called frivolous suit. Since the informed party moves first in this game, it is a signaling game and the settlement demand made by P has a signaling effect. A high-type P - (of damage amount w) has no incentive to make a low demand, but a low-type P randomizes between a high demand and a low demand; in

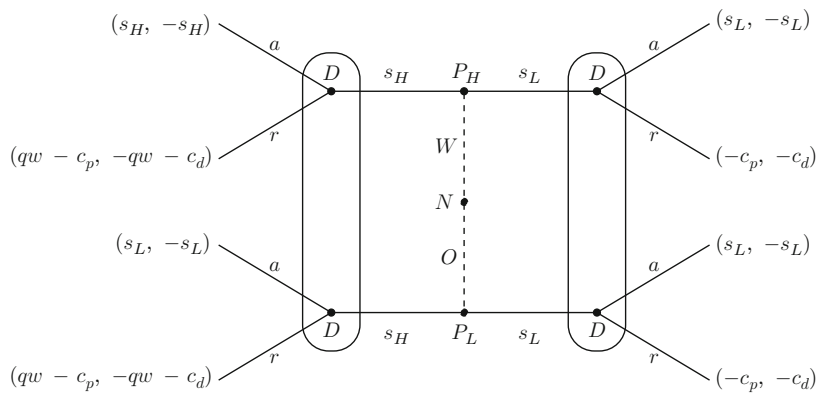
		Receiver's actions	
		t_1	t_2
Student's types	H	3, 3	0, 0
	T	0, 0	2, 2

Credibility, Fig. 2 Type II credibility

Credibility, Fig. 3 Type I
credibility in pretrial
negotiation



Credibility, Fig. 4 Type II
credibility in pretrial
negotiation

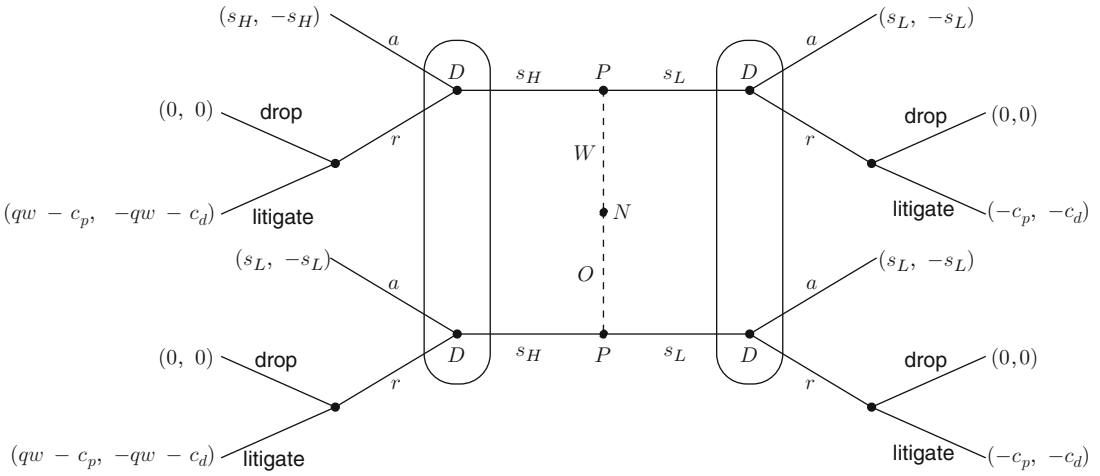


other words, he bluffs in order to extract a positive settlement amount at the risk of costly litigation. Thus, a high demand may not be credible in the sense that it could come from an undamaged P, while D can be sure that a low demand comes from a low-type P implying a frivolous suit. This is an example of type II credibility.

Model 3 The outcome of Fig. 4 relies on the implicit assumption that P commits to litigation when his demand is rejected. However, a low-type P has no reason to proceed to costly litigation, since his expected payoff at court is negative. He would rather drop the case once his demand is rejected. In other words, P's threat to go to trial if

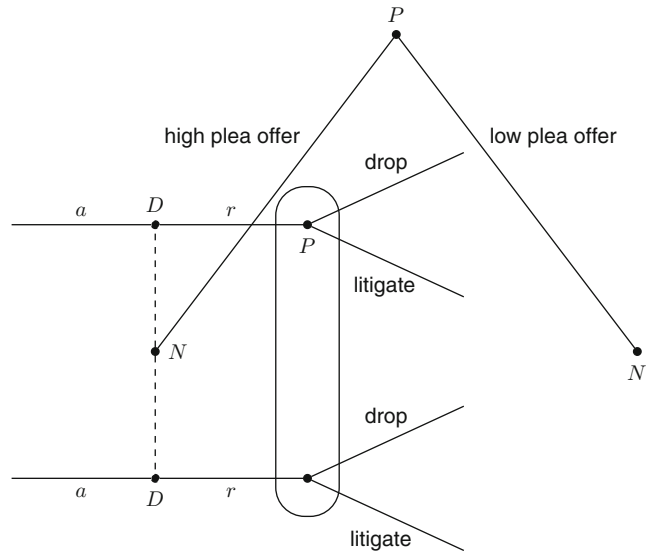
his demand is rejected is not credible. Taking this into consideration, I slightly modify Figs. 4 and 5 by incorporating P's option to withdraw the case. With the option, a low-type P always withdraws the case once his demand is rejected. This increases his expected payoff when he makes a high demand, which will make him choose a high demand with a higher probability. Both types of people are involved in this model.

Model 4 The issue of credibility that occurs when the option to drop the case is added becomes clearer in plea bargaining illustrated in Fig. 6. The defendant is either guilty (G) or innocent (I). The defendant knows his type G or I, but the



Credibility, Fig. 5 Both types of credibility in pretrial negotiation

Credibility, Fig. 6 Both types of credibility in plea bargaining



prosecutor does not. The prosecutor cares about both type I error and type II error. Grossman and Katz (1983) and Reinganum (1988) both identified a separating equilibrium in which a guilty defendant accepts the prosecutor's plea bargaining offer and an innocent one rejects. But both rely on a critical assumption that the prosecutor surely goes to trial once the defendant rejects his offer. In fact, it is not a credible threat to go to trial, because the prosecutor should know that the type who rejects his offer is innocent in the separating equilibrium. The prosecutor who cares

about type I error as well as type II error should drop the case rather than proceed to trial. Grossman and Katz claim that a separating equilibrium cannot be obtained without the prosecutor's commitment to go to court. However, Kim (2010) shows that a (semi)separating equilibrium can be obtained without commitment power if mixed strategies are allowed. It can be a mixed strategy equilibrium that the defendant rejects the prosecutor's plea offer with some probability and the prosecutor chooses to litigate with some probability.

Model 5 Go back to Fig. 3 and consider a version of incomplete information game in which D has private information (about q) instead of P. In equilibrium, if P believes that D's liability is high, he will make a high demand which will be rejected with some positive probability. However, if D is actually a low type, he knows that they will end up with an inefficient outcome of costly trial after he rejects the high demand made by P. Is there any way for D to convince P that he is actually a low type? Suppose that D can use a cheap talk message before the game is played. Then, I can demonstrate that this pre-play communication can induce a more efficient outcome by avoiding some costly trial even though a plain talk is just a costless signal. Consider the following strategies. A high type of D announces "H" (i.e., "I am highly liable for your losses"), and a low type announces "L." P makes a high demand s_H if he observes the message "H" and makes a low demand s_L otherwise. The high demand is always accepted and the low demand is rejected with some probability α . This communicative outcome can be an equilibrium if $s_L = Lw + c_d$ and $s_L < s_H < Hw + c_d$. A high (bad) type of D cannot imitate a low (good) type by saying "L," since it might induce costly trial with some probability, so he would be better by losing s_H for sure rather than $Hw + c_d$ with probability $1 - \alpha$ if α is low. In this cheap talk game, common interests between P and D (predisposition to avoid costly trial) enable costless communication messages to be credible. Credibility of cheap talk in pretrial negotiation is discussed in Kim (1992, 1996).

Disparity between the equilibrium outcome and the efficient outcome is obvious. As far as some information problem is present, the efficient outcome in which all legal disputes are resolved out of court could be hardly achieved even if various costly or costless signals are used. Is there any way to improve efficiency? If voluntary decentralized decisions cannot ensure efficiency, we could resort to institutional means such as binding contracts or law. For example, the government may adopt a rule imposing a penalty for misrepresenting information (e.g., mandatory

discovery rules). Also, reputational concerns may alleviate the credibility problem.

Conclusion

The credibility issue is important in many legal situations beyond pretrial bargaining. For example, should we trust the announcement of the government to strictly regulate certain behavior and impose a severe penalty on it? Since regulation itself is costly to the government, it is optimal for the government that people believe the announcement of the government and discipline their behavior, but if the government knows this, it has no reason to enforce the costly regulation. It is dynamically inconsistent. As such, the government announcement need not to be credible unless the government cares about its long-term reputation or it is a legal commitment. I believe that readers can find more interesting examples in which a lack of credibility weakens the policy of the government.

References

- Crawford V, Sobel J (1982) Strategic information transmission. *Econometrica* 50:1431–1451
- Grossman G, Katz M (1983) Plea bargaining and social welfare. *Am Econ Rev* 73:749–757
- Kim J-Y (1992) Does cheap talk matter in pre-trial negotiation? *Seoul J Econ* 5:301–315
- Kim J-Y (1996) Cheap talk and reputation in repeated pre-trial negotiation. *Rand J Econ* 27:787–802
- Kim J-Y (2010) Credible plea bargaining. *Eur J Law Econ* 29:279–293
- Reinganum J (1988) Plea bargaining and prosecutorial discretion. *Am Econ Rev* 78:713–728
- Selten R (1965) Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft* 12:201–324
- Spence M (1973) Job market signaling. *Q J Econ* 87:355–374

Credit Creation

► Credit Expansions

Credit Expansions

Juan Ramón Rallo
OMMA Center of Studies, Madrid, Spain

Abstract
Credit means deferred payment. Therefore, credit expansion means deferring more payments. The process of granting more credit inside the economic system can foster economic coordination, but it can also lead to intertemporal imbalances.

Keywords
Credit expansion; banks; interest rates; business cycle

Synonyms

Credit creation; Credit supply

Definition

Credit expansion is the process by which economic agents grant credit. Credit expansion can be either coordinated or uncoordinated.

Credit and Its Types

Credit (or debt) is the contractual right of an economic agent (the creditor) to receive a future economic good or service from another (the debtor). Though the nature of such good or service can vary, in monetary economies the repayment or valuation of credit in terms of money is increasingly common. In accounting terms, the monetary value of a debt is a *financial liability*.

Debts repayable in money arise when the debtor promises to deliver money to the creditor in the future. The source of that promise can be either:

- 1. A previous transfer of money from the creditor to the debtor, which gives rise to a loan, otherwise called *financial credit*. In accounting terms:
or
- 2. A previous delivery of goods or services other than money from the creditor to the debtor, which gives rise to an outstanding account, otherwise called *trade credit*. In accounting terms:

Credit from a loan			
Creditor balance sheet at t = 1		Creditor balance sheet at t = 2	
Assets	Liabilities + equity	Assets	Liabilities + equity
Treasury: €10,000	Equity: €10,000	Credit: €10,000	Equity: €10,000
Debtor balance sheet at t = 1		Debtor balance sheet at t = 2	
Assets	Liabilities + equity	Assets	Liabilities
0	0	Treasury: €10,000	Debt: €10,000

Credit from an outstanding account			
Creditor balance sheet at t = 1		Creditor balance sheet at t = 2	
Assets	Liabilities + equity	Assets	Liabilities + equity
Commodities: €10,000	Equity: €10,000	Credit: €10,000	Equity: €10,000
Debtor balance sheet at t = 1		Debtor balance sheet at t = 2	
Assets	Liabilities + equity	Assets	Liabilities + equity
0	0	Commodities: €10,000	Debt: €10,000

The Rate of Interest

The common note of financial and trade credit is that both imply a deferred payment from the debtor and thus both are unsettled transactions: the debtor has a pending obligation to fulfill. As such, every credit involves two essential features: maturity and risk. All credit has a maturity because its repayment is due in the future; it also involves risk as future repayment depends on the ability and willingness of the debtor to fulfill its obligation.

Consequently, granting credit implies that the creditor accepts a deferred payment from the debtor, i.e., the creditor is willing to wait for and bear the risk inherent in future repayment. This is the reason why credit usually comes at a price. That price is called the rate of interest. The rate of interest can thus be understood as the difference in value between a spot, non-risky payment and a deferred, risky payment (Knight 1921; Fisher 1930).

Banks

As every credit transaction involves a particular interest rate, opportunities for arbitrage through intermediation will emerge. In every market where price differentials do appear, arbitrageurs can reduce them by taking a long position (buying side) in the relatively cheap good or asset and a simultaneous short position (selling side) in the relatively expensive good or asset (Fekete 1996).

Banks are the main arbitrageurs in the credit market. They go into debt at low interest rates in order to lend at higher interest rates. As intermediaries, banks specialize in granting credit by obtaining it, lending not their own funds but those of other economic agents. Therefore they do not grant credit out of nothing but out of their own liabilities and match every credit with a debit.

Bank balance sheet	
Assets	Liabilities + equity
Commercial credit: €100,000	Demand deposits: €100,000
Mortgage: €60,000	Covered bonds: €60,000
Buildings: €35,000	Equity: €40,000
Cash reserves: €5000	

Banks’ intermediation serves a useful purpose in coordinating creditors and debtors. However, banks can also distort such coordination. Creditors and debtors have certain preferences about the maturity and risk they wish to assume. When they bargain directly with each other, they tend to reach an optimum agreement that matches both agents’ needs. When, by contrast, they each bargain separately with a financial intermediary, the agreed sets of conditions need not be mutually

compatible: banks can offer creditors maturity and risk conditions that are inconsistent with those offered to debtors. They might do so tempted by the profit opportunity implicit in the interest rate differential between usually low-interest, short-term, safer debts and usually high-interest, long-term, riskier debts.

Credit Expansion

The granting of new, previously inexistent credit is known as credit expansion (since credit is tantamount to debt, credit expansion could also be termed debt expansion). A *coordinated* credit expansion occurs when the sets of conditions offered to creditors and debtors are fully compatible. An *uncoordinated* credit expansion, by contrast, occurs when the sets of conditions offered to both agents are *not* fully compatible.

Coordinated Credit Expansion

Coordinated credit expansions occur when risk and maturity conditions offered by financial intermediaries to lenders and borrowers coincide. This implies that financial intermediaries aim to match the maturity and risk of their assets and liabilities, bringing into mutual consistency their cash outflows and inflows. In other words, banks only grant long-term loans when they hold enough long-term financing sources and only grant loans to risky borrowers when their creditors have willingly acquired risky bank liabilities (such as subordinated bonds).

Since credit was defined as the contractual right of a creditor to receive an economic good or service from a debtor, coordinated credit expansions imply that financial intermediaries can modify in the aggregate neither the future preferences for goods and services among lenders nor the availability of those future goods and services among borrowers. Therefore, in the aggregate, preference for resources among lenders and availability of resources among borrowers should be matched in their maturity and risk profiles. Historically, the prescription for matching banks’ assets and liabilities has been known as the *golden rule of banking* (Hübner 1854).

The most common bank liability is the demand deposit (or, alternatively, the banknote). A demand deposit is a debt callable at sight by the creditor. Due to its immediate and safe convertibility into money, demand deposits are generally used as money substitutes: they can be endorsed to third parties as payment in a given transaction. Economic agents' demand for cash balances is partly satisfied by the supply of banks' demand deposits, thereby transforming hoarding into a means of funding new credits through banking (Selgin 1988). However, if banks wish to expand credit by adequately matching the preferences of both lenders and borrowers, the assets backing demand deposits should be highly liquid, i.e., easily convertible into cash.

Some authors have gone as far as to defend *full-reserve banking* as the only banking practice that can adequately coordinate the preferences of creditors and debtors. Under full-reserve banking, demand deposits can only be issued against an equal sum of cash reserves (Fisher 1935; Friedman 1960). In other words, the reserve ratio (cash reserve divided by demand deposits) must be 100 %. In fact, those authors usually refer to the *money multiplier* as those demand deposits issued in excess of cash reserves. For instance, in the following example, the reserve ratio is 20 % (cash reserves amount to 20 % of all demand deposits) and, therefore, the money multiplier is 5 (the money multiplier is inverse to the reserve ratio).

Bank balance sheet	
Assets	Liabilities + equity
Commercial credit: €90,000	Demand deposits: €100,000
Cash reserves: €20,000	Equity: €10,000

Nonetheless, other economists recognize the possibility of backing demand deposits with other assets of practically equal liquidity such as short-term trade credit (Adam Smith 1776; Melchior Palyi 1936). If financial intermediaries could not issue demand deposits against assets other than cash, then demand deposits would no longer be proper instruments for financial intermediation, i.e., for coordinating creditors and borrowers: they would just become custody deposits (Huerta de Soto 1998).

There certainly seems to be some scope for banks to fund their expansions of short-term credit by increasing their demand deposits without undermining the coordination among lenders and borrowers. However, funding long-term and risky credits with demand deposits (or other kinds of current liabilities) would lead banks to an uncoordinated credit expansion.

Uncoordinated Credit Expansion

Uncoordinated credit expansions occur when risk and maturity conditions offered by financial intermediaries to lenders and borrowers do not coincide. This implies that financial intermediaries mismatch the maturity and risk of their assets and liabilities, throwing their cash outflows and inflows into mutual inconsistency. In other words, banks engage in long-term lending by issuing short-term debt or invest in high-risk assets by issuing debt to risk-averse savers.

Bank balance sheet	
Assets	Liabilities + equity
Mortgages: €100,000	Demand deposits: €100,000
Junk bonds: €50,000	Senior debt: €40,000
	Equity: €10,000

The problems of uncoordinated credit expansion have been generally recognized: (1) banks become exposed to bank runs unless their liabilities are protected by deposit insurance schemes and by institutional lenders of last resort (Diamond and Dybvig 1983); (2) the financial system as a whole becomes fragile and unstable despite the existence of the previously mentioned fire walls (Fekete 1983; Minsky 1986); (3) the real economy runs the risk of entering a business cycle due to a provision of credit for investment that is far in excess of real savings (Mises 1912; Hayek 1931).

The law of large numbers might make it easy to think that an individual bank can avoid the adverse consequences of its own asset and liability mismatching. However, this is certainly not possible for the whole banking system. A widespread uncoordinated credit expansion endogenously gives rise to the financial instability and business cycle that ultimately undermine the very operation of the law of large numbers (Huerta de Soto 1998).



Financial regulation is, in fact, increasingly acknowledging the dangers involved in maturity and risk mismatching. Basel III, for instance, requires banks to keep a minimum capital buffer in order to absorb potential losses (their common equity and Tier I capital should at least be equal to 4.5 % and 6 %, respectively, of their risk-weighted assets). Moreover, Basel III imposes liquidity constraints: the Liquidity Coverage Ratio (banks must have enough liquid assets to cope with 1 month of net cash outflows) and Net Stable Funding Ratio (aimed at ensuring that banks hold a minimum amount of stable funding based on the liquidity characteristics of their assets and activities over a one-year horizon). Other authors, however, suggest that precisely the absence of government intervention would result in a competitive environment conducive to a natural and much more effective self-regulation (Selgin and White 1994).

Conclusion

Whichever the proposed solutions to avoid them, it is clear that uncoordinated credit expansions tend to distort a financial system. Therefore, they should not be confused with *coordinated* credit expansions, which, by matching creditors and debtors' preferences, increase the number of economic exchanges in a sustainable manner.

Cross-References

- [Banks](#)
- [Hayek, Friedrich August von](#)

References

- Diamond D, Dybvig P (1983) Bank runs, deposit insurance, and liquidity. *J Polit Econom* 91(3):401–419
- Fekete A (1983) Borrowing short and lending long. Committee for Monetary Research and Education, Charlotte
- Fekete A (1996) Towards a dynamic microeconomics. *Laissez-Faire* n°5. Universidad Francisco Marroquín, Guatemala, pp 1–14
- Fisher I (1935) 100% money: designed to keep checking banks 100% liquid; to prevent inflation and deflation;

- largely to cure or prevent depressions; and to wipe out much of the national debt. The Adelphi Company, New York
- Fisher I (1930) The theory of interest. The Macmillan Company, New York
- Friedman M (1960) A program for monetary stability. Fordham University Press, New York
- Hayek F (1931) Prices and production. Routledge and Sons, London
- Hübner O (1854) Die Banken. Hübner, Leipzig
- Huerta de Soto J (1998) Dinero, crédito bancario y ciclos económicos. Unión Editorial, Madrid
- Knight F (1921) Risk, uncertainty, and profit. Houghton Mifflin, Boston
- Minsky H (1986) Stabilizing an unstable economy. Yale University Press, New Haven
- Mises L (1912) Theorie des Geldes und der Umlaufsmittel. Duncker and Humblot, Munich
- Palyi M (1936) Liquidity Minnesota bankers association. Minneapolis
- Selgin G (1988) The theory of free banking. Rowman & Littlefield, Totowa
- Selgin G, White L (1994) How would the invisible hand handle with money. *J Econ Lit* 32(4):1718–1749
- Smith A (1776) The wealth of nations. W. Strahan and T Cadell, Scotland

Credit Supply

- [Credit Expansions](#)

Credit Trading

- [Emissions Trading](#)

Credit: Rating Agencies

Alessandro Romano
China-EU School of Law, China University of Political Science and Law, Beijing, China

Abstract

Credit rating agencies (CRAs) are pivotal players in financial markets, and in fact their conduct has attracted the attention of scholars, media, and policy analysts. A very common claim is that CRA behavior contributed to the

explosion and the propagation of the recent financial crisis. This entry sketches the functioning of the market for ratings and explores the market failures by which it is characterized. Moreover, this entry briefly presents some of the proposals advanced by the law and economics literature to induce CRAs to issue accurate ratings.

Introduction

The main activity of credit rating agencies (CRAs) is providing investors and regulators with certifications of the quality of financial assets. Any lender is interested in knowing what is the likelihood that the borrower will honor his debt, and ratings serve exactly this need. In a nutshell, ratings are informed opinions on the probability that the lender (issuer of the financial asset) will not repay the borrower (investor). In other words, ratings are an estimate of the probability of default (PD) of a given bond. In some cases, ratings also account for other factors, like the expected magnitude of the losses associated with a possible default of the issuer (loss given default (LGD)). When ratings are not accurate, they can be either inflated or deflated. A rating is inflated when the CRA overestimates the creditworthiness of the rated bond. Instead, a rating is deflated when the creditworthiness of the issuer is underestimated. CRAs can potentially be useful actors on financial markets because they help reducing information asymmetries between the issuers of the rated assets and regulators and investors. However, over the last decade CRAs have been at the center of a very heated debate both at the policy level and on the scientific literature as, allegedly, they have played a crucial role in the recent financial crisis. In particular, many scholars and policy makers have argued that CRAs issued inflated ratings thus contributing to the explosion and the propagation of the financial crisis (White 2010). This entry investigates how much truth there is behind these accusations, what are the reasons that might have lead credit rating agencies to inflate their ratings, and what are the proposals advanced by the law and economics literature to induce CRAs

to issue accurate ratings. Two preliminary caveats are required. First, ratings can be divided in two broad categories: solicited ratings and unsolicited ratings. Solicited ratings are requested by the issuer that pays a fee to be rated by the CRA and provides the CRA with relevant information. Instead, unsolicited ratings are spontaneously issued by the CRA that does not receive any fee and are generally based on information available to the public. Because solicited and unsolicited ratings are associated with drastically different incentives for the parties involved, these two kinds of ratings cannot be analyzed together. This entry focuses only on solicited ratings as the fees collected issuing this kind of ratings generate the vast majority of CRAs revenues. Second, not all solicited ratings are equal. Ratings of structured finance products are markedly different – and way more problematic – than ratings of corporate bonds.

History of the Market for Ratings

In 1909 John Moody published the first publicly available bond ratings. Other firms soon engaged in this practice creating the market for ratings. At the time, these ratings were sold to investors who paid to have an overview of the creditworthiness of a number of issuers. This is a business model currently referred to as investor-pays model. However, a number of reforms and technological innovations completely transformed the landscape of the market for ratings. A detailed analysis lies beyond the scope of this entry, but a quick overview of the most relevant changes is useful to understand what the problems in the market for ratings are.

Starting from the 1930s, regulators began to grant credit rating agencies an increasingly fundamental role in the functioning of financial markets. For example, in 1936 bank regulators issued a decree that was aimed at preventing banks from investing in “junk bonds,” as defined by “recognized ratings manuals.” As the only recognized ratings manuals were Moody’s, Poor’s, Standard, and Fitch, regulators de facto gave to the judgments of these four rating agencies (later to

become three when Standard and Poor's merged into Standard & Poor's) the status of law (White 2010). Insurance and pension regulators adopted very similar rules (White 2010). In the 1970s, Moody's, Standard & Poor's, and Fitch relevance on financial markets was further increased by the combined effect of two decisions of the Security and Exchange Commission (SEC). On the one hand, the SEC imposed that the minimum capital requirements of brokers-dealers had to be tied to the riskiness of their asset portfolio, and ratings were to be used to determine the level of risk. On the other hand, afraid that smaller CRAs not constrained by reputational concerns could issue inflated ratings to attract customers, the SEC dictated that only ratings issued by "nationally recognized statistical rating organization" (NRSRO) could influence the minimum capital requirements. Moody's, Standard & Poor's, and Fitch were the only CRAs that obtained the status of NRSRO. Combined with a number of other regulations, the effect of these reforms was to grant Moody's, Standard & Poor's, and Fitch a quasi-regulatory power.

Another piece of the puzzle is the change in the business model that CRAs undertook in the 1970s. Potential free riding problems associated with the diffusion of fast photocopy machines undermined the investor-pays model. In particular, because it was becoming easy, quick, and cheap to disseminate information, rating agencies feared that the content of their ratings could have circulated also among investors that did not pay the relative fee. Therefore, instead of exacting a payment from investors who wanted to see the ratings of a given issuer, CRAs started to require a fee from the issuer that they had to rate. Nowadays, 95% of CRAs revenues derive from the fees collected from issuers (Partnoy 1999). Last, over the last decades rating agencies started to rate structured finance products that were becoming more and more complex.

Summarizing, three main factors characterized the evolution of the market for ratings:

- An increased regulatory relevance of ratings that granted CRAs a quasi-regulatory power.
- The adoption of an issuer-pays model.

- CRAs started rating complex structured finance products.

Failures in the Market for Ratings

In the wake of the crisis, referring to CRAs the Nobel Prize winner Paul Krugman wrote that:

It was a system that looked dignified and respectable on the surface. Yet it produced huge conflicts of interest. Issuers of debt could choose among several rating agencies. So they could direct their business to whichever agency was most likely to give a favorable verdict, and threaten to pull business from an agency that tried too hard to do its job. (Krugman 2010)

To put it differently, CRAs rate their clients, and hence they could be inclined to cater to the needs of the latter to attract more business (Darcy 2009). This view has been extremely influential in the literature, in the policy debate, and in the media, but it overlooks the insights of reputational capital theory (e.g., Choi 1998). According to this theory, "the only reason that rating agencies are able to charge fees at all is because the public has enough confidence in the integrity of these ratings to find them of value in evaluating the riskiness of investments" (Macey 1998). Therefore, absent other market failures, reputational sanctions would discipline CRAs' behavior preventing them from inflating ratings. In this vein, the literature has looked beyond the conflict of interest and identified three other market failures that altered the functioning of the market for ratings.

First, reputational capital theory is grounded on the idea that investors are sophisticated enough to determine when ratings are inflated. However, if a large enough fraction of investors is Naive and cannot identify inaccurate ratings, reputational sanctions become largely ineffective (Bolton et al. 2012). Second, CRAs collect their fee only when they publish the ratings, and hence issuers could contact multiple rating agencies and request publication only for the most favorable rating received (Dennis 2009). This practice of shopping for the most favorable rating can result in rating inflation, especially for complex assets (Skreta and Veldkamp 2009). Third, as high ratings are

associated with regulatory benefits, issuers might be interested in purchasing good ratings, regardless of whether investors trust the ratings. Thus, when the regulatory benefits attached to high ratings are sufficiently relevant, a rating agency “finds it profitable to stop acquiring any information and merely facilitates regulatory arbitrage through rating inflation” (Opp et al. 2013).

Therefore, there is a combination of market failures that, associated with the issuer-pays model, induces CRAs to inflate their ratings. And indeed, while the effective contribution of rating inflation to the financial crisis is still disputed (Gorton and Ordoñez (2014) argue that the contribution was likely to be limited), a large part of the literature finds that ratings, especially of structured finance products, were inflated (e.g., Calomiris 2009).

Fixing the Market for Rating

The issuer-pays model, combined with the possibility of shopping for the most favorable rating, the regulatory benefits attached to high ratings, and the naivety of some investors, creates incentives for the credit rating agencies to inflate their ratings. As there is such a complex web of market failures, inducing CRAs to issue accurate ratings is no easy task. Moreover, not all ratings are equal, and ratings of complex structured finance products create more concerns than the traditional ratings of corporate bonds. The main reasons are that (i) structured finance products are more complex and rating inflation can be more severe for complex bonds (Skreta and Veldkamp 2009) and (ii) many structured finance products behave as economic catastrophe bonds, that is, these financial assets are less resistant to economic downturns and their defaults are highly correlated (Coval et al. 2009). Any proposal that aims at improving the functioning of the market for ratings should account for these differences.

The market for ratings is an oligopoly with high barriers to entry dominated by Moody's, Standard & Poor's, and Fitch, and hence an obvious solution to ameliorate CRAs incentives could be increasing the competition in the market (Hill

2003). Nevertheless, empirical research shows that under the *status quo*, competition *worsens* the quality of ratings (Becker and Milbourn 2011). In presence of rating shopping more competition negatively affects the quality of ratings because the issuer has more choices when searching for the most favorable rating (Becker and Milbourn 2011). Alternatively, as regulatory benefits attached to high ratings neutralize – or at least reduce the impact of – reputational sanctions, one obvious solution to improve the quality of the ratings is reducing their regulatory relevance (Flannery et al. 2010). The Dodd-Frank Act takes exactly this path, but it seems unlikely that the implemented reforms will suffice to eliminate both direct and indirect regulatory benefits derived from relying on ratings (Hill 2010). And indeed, the European regulator explicitly remarked that there are no perfect substitutes for ratings, and hence ratings are bound to have some regulatory value (Pacces and Romano 2015). As noted by Coffee (2011), regulators decided to assign regulatory value to ratings because they have limited information and cannot develop reliable measures of risk. In other words, ratings are a precious component of regulation, provided that they are accurate, because there is no guarantee that alternative solutions to identify excessive risk will not prove to be even more problematic. Therefore, reforms should attempt to improve CRAs' incentives while preserving – at least partially – the role played by ratings in financial regulation. Another possible reform is forcing CRAs to abandon the current issuer-pays model in favor of different business models (Mathis et al. 2009) or even introducing some sort of public funding for CRAs (Listokin and Taibleson 2010). However, the information contained in ratings has the nature of a public good because it can easily be disseminated among investors, and hence alternatives to the issuer-pays model are generally considered unworkable (Partnoy 1999; Coffee 2011). Another proposal is to pay rating agencies with the debt that they rate (Listokin and Taibleson 2010). In this vein, CRAs would be punished when issuing overoptimistic ratings because they receive debt that is worth less than they claim. Last, a path that has been widely

explored by the law and economics literature is to make the liability threat faced by CRAs issuing inaccurate ratings more credible. In fact, for many years CRAs have been de facto immune to liability claims (Coffee 2006; Partnoy 2006), and in the United States they were even put under the umbrella of the First Amendment on the freedom of speech (Deats 2010). Generally, the literature considers a negligence rule as the most appropriate to induce credit rating agencies to issue accurate ratings. Under this rule, rating agencies are asked to compensate the investor only when they have been negligent in formulating their rating. In Europe, the United States, and Australia the liability of CRAs is largely based on this logic. There are, however, a number of problems with this approach (Pacces and Romano 2015). First, it is extremely hard for courts to identify the optimal level of care (Coffee 2004), because ratings are complex and prospective judgments that necessarily involve at least some subjectivity on the part of the raters. Determining when there was negligence in the formulation of this prospective judgment has been defined a Serbonian Bog by the literature (Coffee 2004). Imprecise and uncertain standards of care are notoriously associated with an increase in transaction costs and more unpredictability of courts' behavior. Second, CRAs are not responsible for all the losses associated to a default. For example, CRAs did not lower Enron rating below investment grade until only a few days before the bankruptcy, thus fueling accusations of negligent behavior on their part (Frost 2007). However, while CRAs can be held liable for not detecting Enron's problems, they are certainly not liable for the fact that Enron went bankrupt. Therefore, they should be asked to compensate only a fraction of the harm associated with Enron's bankruptcy. Identifying which fraction of the harm is attributable to CRAs conduct is extremely challenging, if not impossible (Pacces and Romano 2015). If the liability threat is reinforced via a negligence rule, the combined effect of these two problems might be that rating agencies become exceedingly conservative in their judgments or even refuse to rate risky securities. These problems would be especially severe because, on the one hand, risky assets are exactly

those for which ratings are more needed. On the other hand, CRAs can play a beneficial role only if they issue accurate ratings, not if they issue deflated ratings. Empirical evidence and theoretical studies suggest that this risk is concrete. The Dodd-Frank Act significantly increased CRAs' liability exposure, and Dimitrov et al. (2015) found that as a consequence the informative content of corporate ratings further worsened. At a theoretical level, Goel and Thakor (2011) show that when the liability threat is severe, credit rating agencies have an incentive to issue deflated ratings, because it is unlikely that courts will hold them liable for conservative ratings. To put it differently, the expected liability faced by CRAs issuing inflated ratings is larger than that faced by CRAs issuing deflated ratings.

The problems of a strict liability rule are as severe. On the one hand, if CRAs are asked to cover all the losses whenever an issuer they rated defaults, they would bankrupt almost immediately as the default of a single large issuer might cause losses that exceed the assets of a CRA. Moreover, under a strict liability rule, the injurer (here the CRA) acts as a de facto insurer (Priest 1987). It is common wisdom that only uncorrelated risks can be insured (Priest 1987), whereas defaults – especially of structured finance products – are highly correlated and concentrate during economic crises (Coval et al. 2009). Pacces and Romano (2015) attempt to cope with this problem proposing a less intrusive form of strict liability that relies mainly on market forces. Introducing a damage cap based on objective factors and corrections to shield CRAs from the risk of correlated defaults, Pacces and Romano (2015) argue that a modified regime of strict liability might induce CRAs to issue ratings that are as accurate as the available forecasting techniques allow.

In conclusion, the market for ratings is characterized by multiple market failures, and hence the literature is still struggling to find effective solutions.

Future Research

How to prevent future malfunctioning in the market for ratings is still an open issue. In the coming

years quantitative studies might attempt to disentangle the size of the impact of each market failure and to understand how these market failures interact with each other. For example, regulatory benefits and investors' naivety reduce the effect of reputational sanctions, but how their effects are related is more obscure. The relationship between the effects of these two failures might be additive (e.g., if each reduces the magnitude of reputational sanctions by 10%, then the joint effect is 20%), or the effects of the two market failures could stand in more complex relationships (e.g., the presence of naïve investors magnifies the effect on reputational sanctions of the regulatory benefits, thus producing a joint effect larger than 20%). And indeed, we have seen that the issuer-pays model in itself is not necessarily problematic, but it creates perverse incentives when combined with other market failures. A clearer and more quantitative understating of the interactions among market failures in the market for ratings might help improving the regulatory regime of rating agencies.

Another important question that awaits an answer is to which extent regulatory reliance on rating agencies is motivated. In turn, answering this question implies that alternative solutions and their possible limitations are explored. Flannery et al. (2010) attempt exactly this task, but more studies are warranted.

Cross-References

- [Conflict of Interest](#)
- [Market Failure: Analysis](#)
- [Strict Liability Versus Negligence](#)

References

- Becker B, Milbourn T (2011) How did increased competition affect credit ratings? *J Financ Econ* 101:493
- Bolton P, Freixas X, Shapiro J (2012) The credit ratings game. *J Finance* 67:85
- Calomiris CW (2009) A recipe for ratings reform. *Econ Voice* 6:1
- Choi S (1998) Market lessons for gatekeepers. *Northwest Univ Law Rev* 92:916–919
- Coffee JC Jr (2004) Gatekeeper failure and reform: the challenge of fashioning relevant reforms. *Boston Univ Law Rev* 84:301
- Coffee JC Jr (2006) Gatekeepers: the professions and corporate governance. Oxford University Press, Oxford
- Coffee JC Jr (2011) Ratings reforms: the good, the bad and the ugly. *Harv Bus Law Rev* 1:231
- Coval J et al (2009) The economics of structured finance. *J Econ Perspect* 23:3
- Darcy D (2009) Credit rating agencies and the credit crisis: how the issuer pays conflict contributed and what regulators might do about it. *C Bus Law Rev* 2:605
- Deats C (2010) Note, talk that isn't cheap: does the first amendment protect credit rating agencies' faulty methodologies from regulation? *Columbia Law Rev* 110:1818
- Dennis K (2009) The rating game: explaining rating agency failures in the buildup to the financial crisis. *Univ Miami Law Rev* 63:1111
- Dimitrov V et al (2015) Impact of the Dodd-Frank act on credit ratings. *J Fin Econ* 115:505
- Flannery MJ et al (2010) Credit default swap spreads as viable substitutes for credit ratings. *Univ Pennsylvania Law Rev* 158:2085
- Frost CA (2007) Credit rating agencies in capital markets: a review of research evidence on selected criticisms of the agencies. *J Account Audit Finance* 22:469
- Goel AM, Thakor AV (2011) Credit ratings and litigation risk. Available at <http://ssrn.com/abstract51787206>
- Gorton G, Ordoñez G (2014) Collateral crises. *Am Econ Rev* 104:343
- Hill CA (2003) Rating agencies behaving badly: the case of Enron. *Conn Law Rev* 35:1145
- Hill CA (2010) Justification norms under uncertainty: a preliminary inquiry. *Conn Insur Law J* 17:27
- Krugman P (2010) Berating the raters, *N.Y Times*. http://www.nytimes.com/2010/04/26/opinion/26krugman.html?_r=0
- Listokin Y, Taibleson B (2010) If you Misrate, then you lose: improving credit rating accuracy through incentive compensation. *Yale J Regul* 27:91
- Macey (1998) Wall street versus main street: how ignorance, hyperbole, and fear lead to regulation. *Univ Chicago Law Rev* 65:1487
- Mathis J, McAndrews J, Rochet J-C (2009) Rating the raters: are reputation concerns powerful enough to discipline rating agencies? *J Monet Econ* 56:657
- Opp C et al (2013) Rating agencies in the face of regulation. *J Financ Econ* 108:46
- Paccos A, Romano A (2015) A strict liability regime for rating agencies. *Am Bus Law J* 52:673
- Partnoy F (1999) The Siskel and Ebert of financial markets?: two thumbs down for the credit rating agencies. *Wash Univ Law Q* 77:619
- Partnoy F (2006) How and why credit rating agencies are not like other gatekeepers. In: Fuchita Y, Litan RE (eds) *Financial gatekeepers: can they protect investors?* Nomura Institute of Capital Markets Research/Brooking Institution Press, Tokyo/Washington, DC, p 61

- Priest GL (1987) The current insurance crisis and modern tort law. *Yale Law J* 96:1521
- Skreta V, Veldkamp L (2009) Ratings shopping and asset complexity: a theory of ratings inflation. *J Monet Econ* 56:678
- White LJ (2010) Markets: the credit rating agencies. *J Econ Perspect* 24:211

Crime and Punishment (Becker 1968)

Jean-Baptiste Fleury
THEMA, University of Cergy-Pontoise,
Cergy-Pontoise, France

Definition

Gary Becker's 1968 "Crime and Punishment: An Economic Approach" is one of the first papers using economics to address the questions of crime and law enforcement. To Becker, crime generates costs to society, but fighting crime is also costly. There is, therefore, an optimal amount of crime which minimizes society's total loss and which can be attained by setting the optimal levels of punishment and probability of apprehension and conviction. From that analysis, Becker further claims that the role of criminal law and law enforcement policies should be limited to the minimization of society's loss. Crime is, therefore, framed as an external effect, and criminal law's purpose is redefined as the activity of assessing the harm incurred by crime in order to enforce optimal compensation.

Becker's 1968 "Crime and Punishment: An Economic Approach"

Gary Becker's 1968 "Crime and Punishment: an Economic Approach" is one of the first articles by a modern economist (post-World War II) to address crime (see also Eide et al. 2006). Due to its huge influence not only in creating the whole subfield of economic analyses of criminal behavior and public law enforcement (see the extensive

survey of Polinsky and Shavell 2000) but also in the development of Richard Posner's economic analysis of law (see Posner 1993), the paper is worth studying. The paper is quite typical of Gary Becker's approach to economics. First, it applies standard microeconomic tools to the problem of public law enforcement, which was, up to 1968, a topic traditionally considered to be outside of the domain of economics. Although Cesare Beccaria and Jeremy Bentham interpreted criminal law in a utilitarian and economic framework during the eighteenth and nineteenth centuries, their works were progressively relegated to the fringes of both economics and criminology during the twentieth century and are generally mentioned only for historical references. Second, the scope of the paper is much broader than crime and law enforcement: it is a "generalization of the economist's analysis of external harm or diseconomies" (Becker 1968, p. 201), which offers to redefine both crime and criminal law along the lines of economic efficiency.

The Basic Model

Let's consider the first aspect of Becker's analysis: the application of standard microeconomic analysis to the question of law enforcement. To address the problem, Becker adopts an approach reminiscent of welfare economics and resorts to a social welfare function which computes the total social net loss generated by criminal activities, with such activities being defined – so far – exogenously by the law. To simplify the analysis, Becker reduces the scope of the social loss function to losses in real income only. The function can be decomposed into three different sections: first, the total net social damages due to criminal activity (D). Second, the costs of apprehension and conviction of offenders (C). Finally, the social cost of punishment (S). All these subpart depend on the number of offenses that are perpetrated during a given period of time.

Thus $L = D + C + S$.

Becker takes a utilitarian point of view: a given offense generates damages to the victim and society but generates gains to offenders that have to be

taken into account. Following the general economic assumptions, the additional offense yields positive and increasing harm to society and positive but decreasing gains to offenders. The production of law enforcement is an activity which is costly. Thus, if more resources are invested in increasing the probability of conviction p , the cost increases, and the marginal costs also increase. If the number of offenses increases, the cost of law enforcement increases, as well as its marginal costs. Finally, most forms of punishment (with the notable exception of fines) generate costs to society: prisons need personnel and facilities, while imprisonment also incurs costs to the offender depending on his opportunity cost of time.

Since all the subparts of the equation depend on the number of offenses, Becker's central point is to assess the optimal number of offenses in a given society. Clearly, crime not only generates costs to society: apprehending and punishing offenders also generate costs, so zero crime appears socially inefficient. There must be a total number of offenses that yields a socially desirable outcome. As Becker (1968, p. 170) famously wrote, the main question of the analysis is "how many offenses *should* be permitted, and how many offenders *should* go unpunished?" Public policy exerts an influence on the number of offenses, through two main variables. One is the probability of apprehension and conviction, p . The second is the level of punishment, f . To complete the model, thus, Becker analyzes the supply of offenses by criminals, which mostly depends on those two variables.

Although Becker's analysis of criminal behavior is perhaps the most remembered contribution of his 1968 paper, the actual model of individual behavior is relegated in a footnote, as Becker's central focus is on the market supply of offenses. Becker assumes that individuals are perfectly rational: they maximize the expected utility they derive from the net gains acquired from a criminal activity. It depends, therefore, on p and f . An increase in the probability of conviction p would reduce the expected utility from criminal activity, as would an increase in the level of punishment f . Becker's analysis shows that criminals tend to be risk lovers, which means that a given increase in

f perfectly compensated by a decrease in p would leave the expected gains the same but would make criminals better off. Therefore, when criminals are risk lovers, the elasticity of supply of offenses with respect to p is greater than the elasticity of supply of offenses with respect to f .

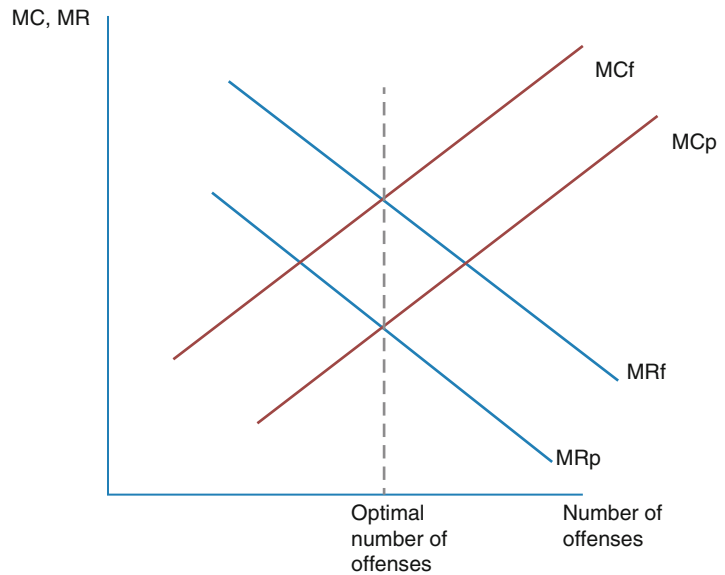
Optimality Conditions

With the model now complete, Becker's formulation of the optimality conditions is organized so as to compare marginal costs and marginal benefits from a given increase in the quantity of offenses, generated by a decrease either in p or f or both. Regarding marginal costs, a small decrease in f would yield additional offenses, incurring therefore additional costs from damages and from conviction and apprehension (curve MC f in Fig. 1). The same could be said for a small decrease in p , but the marginal cost to society would be smaller than in the previous case because, this time, society saves some costs related to law enforcement: reducing the probability of cleared cases means roughly less resources spent on law enforcement (curve MC p). Both curves are upward sloping due to the increasing marginal damages and marginal costs functions associated to an increase in the number of offenses.

Marginal benefits are related to the social costs of punishment: more offenses means less punishment, and, therefore, less social costs due to these punishments. The curve is downward sloping. Therefore, marginal benefits are related essentially to two elements. The first is the coefficient transforming a given punishment into the social costs of such a punishment. The second is the elasticity of supply of offenses with respect to p and f . Indeed, the social "benefits" coming from an increase in offenses depend partly on the sensitivity of offenders to decreases in p or f . Because marginal cost must equal marginal revenue and because the marginal cost of a reduction in p is lower than the one due to a reduction in f , the marginal revenue related to a reduction in p has to be lower than the marginal revenue related to a reduction in f (the MR f curve is higher than the MR p curve). Becker shows that this is

Crime and Punishment (Becker 1968),

Fig. 1 Optimality conditions



only possible if the elasticity of supply of offenses with respect to p is greater than the one with respect to f , in other words, if criminals react more strongly to an increase in p rather than f , that is, if they are risk lovers.

When marginal costs equal marginal revenue, one finds the optimal number of offenses, and the optimal values of p and f leading to such offenses.

From this analysis of optimality conditions, we can firstly conclude that Becker's paper provides a normative analysis of how to allocate resources in order to minimize society's loss, that is, to reach the optimal amount of crime by playing on the values of p and f . But, in the meantime, it also provides an evaluation of current and past policies against crime. Becker generally concludes that what can be observed in terms of actual probability of conviction as well as severity of punishment (and their evolution in time) is overall compatible with his normative statements. Therefore, in a sense, the theory provides also a positive analysis of how society considers crime and has responded to crime over time. Becker's comparative statics are in this respect enlightening, and two examples will be explored.

First, suppose that a given increase in crime yields higher marginal damage to society. The marginal cost curve would shift upward, and the optimal amount of crime would go down. This

would be achieved by an increase in both p and f . Becker's crude empirical investigation showed, indeed, that the more serious the felony is, the more likely the criminal is going to be convicted and the harsher the punishment he will face. Second, suppose that a reduction in the marginal revenue came from an increase in the elasticity of supply (E_f) with regards to f – the social cost of punishment is therefore reduced, which means that the marginal revenue that society gains through additional offenses diminishes – then the optimal level of crime is reduced and is achieved through an increase in f . Such a result leads Becker to show that society would minimize its loss by engaging in price discrimination: different groups of offenders with different elasticities should be charged different levels of punishment. Becker shows that this is consistent with actual practice, where groups being insensitive to punishments, such as impulsive murderers or juveniles, generally face lower punishments and more therapy for similar crimes.

Criminal law and the General Theory of External Effects

Becker's economic analysis supports a view of criminal law and law enforcement practices

grounded on the maximization of social welfare. From that perspective, the main role of punishment is to compensate for the marginal harm done to victims and society. Becker considers other motives such as revenge and deterrence, as secondary. Thus, the amount of punishment does not depend on the specific conditions of the criminal, but on the marginal damage suffered by society, which needs to be compensated. Fines are the punishment that yields almost no social cost: contrary to imprisonment, fines are simply performing wealth transfers. They stand as the perfect compensation mechanism and should be, therefore, used whenever necessary.

From this normative analysis, Becker broadens the scope of his paper. First, he redefines completely the notion of crime, not as an exogenous activity deemed illegal by law, but as any activity which generates harm that was not compensated. Eventually, there are no differences between a person buying a car and a thief stealing one and compensating society afterward through a well-calculated fine. Only when the damage cannot be compensated by fines should the “debtor” repay the involuntary “creditor” as well a society through other forms of punishments, such as prisons. Second, he redefines criminal law: “the primary aim of all legal proceedings would become the same: not punishment or deterrence, but simply the assessment of ‘harm’ done by defendants,” in order to calculate the levels of compensation (Becker 1968, p. 198). Thus, “much of traditional criminal law would become a branch of the law of torts, say, ‘social torts’, in which the public would collectively sue for ‘public harm’” (Becker 1968, p. 198). Note that this aspect of the paper had a tremendous influence on Posner’s subsequent economic analysis of law and his 1985 definition of crime as “market bypassing” (see Posner 1985, 1993). Consequently, Becker concludes that his analysis of crime is a generalization of the theory of external effects, placing criminal law and law enforcement as the central institutional setting to articulate a sort of Pigovian taxation. To Becker, crime becomes nothing more than an external effect which, if negative, has to be taxed and reduced to an optimal level. But Becker even considers the

case of positive external effects, which should be regulated with subsidies, rewards, and other forms of cash prizes, which magnitude and probability to award would be the social variables to be controlled. Law enforcement, in this case, would have to spend resources to find and award inventors and other producers of external benefits.

Cross-References

- [Becker, Gary S.](#)
- [Cost of Crime](#)
- [Criminal Sanctions and Deterrence](#)
- [Economic Analysis of Law](#)
- [Law and Economics](#)
- [Law and Economics, History of](#)
- [Posner, Richard](#)

References

- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Eide E, Rubin PH, Shepherd JM (2006) *Economics of crime*. Now Publishers Inc., Delft
- Polinsky M, Shavell S (2000) The economic theory of public enforcement of law. *J Econ Lit* 38(1):45–76
- Posner RA (1985) An economic theory of the criminal law. *Columbia Law Rev* 85(6):1193–1231
- Posner RA (1993) Gary Becker’s contributions to law and economics. *J Leg Stud* 22(2):211–215

Crime, Attitude Towards

Elena D’Agostino¹, Emiliano Sironi² and Giuseppe Sobbrìo¹

¹Department of Economics, University of Messina, Messina, Italy

²Department of Statistical Sciences, Catholic University of Milan, Milan, Italy

Abstract

Crime negatively affects social welfare and reduces citizens’ trust in public institutions and society. Perhaps the best starting point is trying to look at criminal behavior as human

and social phenomenon in order to understand what pushes people to illegal (and certainly risky) activities. The psychological literature has emphasized the role of attitudes as one of the main determinants of human behavior. We reviewed this literature and its applications to crime.

The Theory of Planned Behavior

Crime negatively affects social welfare and reduces citizens' trust in public institutions and society. Reducing crime is therefore a priority of every public agenda across the world, although what is the most effective program remains an open question. Perhaps, the best starting point is trying to look at criminal behavior as human and social phenomenon in order to understand what pushes people to illegal (and certainly risky) activities.

One approach to crime consists of viewing it simply as a human behavior. As such, the decision to commit a crime is the (more or less) logical consequence of a mental process involving morality, civic awareness, and personal character. According to Ajzen (1991) and his theory of planned behavior, an individual's decision to engage or not engage in a given behavior is anticipated by the formation of positive intentions toward that behavior. Intentions depend primarily on three variables: (1) personal *attitudes* toward a given behavior, (2) people's belief concerning whether he or she is expected by other individuals to perform that behavior (*subjective norms*), and (3) the individual's perceived ease or difficulty of performing the behavior (*perceived behavioral control*).

Applications

The theory of planned behavior has been tested in many fields through cross analysis of attitudes and behaviors. For example, environmental economics suggests that sustainable behaviors are widely promoted by positive attitudes, but the lack of accessible recycling infrastructures and other

constraints work against the original intention. Health economics has used TPB to determine the obesity factors of overweight among Chinese Americans (Liou and Bauer 2007).

Testing the TPB model in a similar way in relation to crime is difficult because people are clearly not keen to admit they have committed crimes, so there is a lack of information concerning the last (and most important) part of the story, that is, intentions or behaviors. Nevertheless, if it is accepted that criminal behavior is a decision-making process, it is possible to understand the main factors that influence people's future illegal actions by focusing on attitudes, used to denote their internal evaluations of the extent to which respecting laws will have positive or negative consequences for their lives and happiness.

There are many attempts in the literature to analyze and understand people's attitudes toward legality and therefore toward crime. Using data from the world value survey, Torgler and Schneider (2007) investigated the determinants of attitudes toward paying more or less in taxes from a cross-country perspective, considering the impact of both sociodemographic and cultural background. Further studies have focused on the relationship between education and crime, commonly arguing that education and the associated higher earnings negatively affect both crime (among others, Buonanno and Leonida 2006) and psychological attitudes toward crime (Arrow 1997). However, these findings are far from conclusive. Groot and Van Den Brink (2010) found evidence of a relationship between high education levels and attitude toward serious crimes in the Netherlands. D'Agostino et al. (2013) confirmed these findings in a cross analysis involving most European countries.

Empirical studies show a strong positive relationship between fear of crime and media consumption (among others, see Barille 1984).

Concluding Remarks

What is common to all these papers is that each of them focuses on specific drivers for determining

attitudes toward crime and punishment (media, economic recession, education). Some of these contributions examines whether attitudes toward crime are influenced more by individual variables, involving personal features (such as gender, race, education, family), or by contextual variables, referring to institutional and economic aspects (such as corruption, GDP, growth, interpersonal safety). The results, although in line with the hypothesis of a combined effect of the context and of individual features, are somewhat counterintuitive: often, more educated individuals seem also to be more tolerant with criminals.

This result is probably due to the high heterogeneity of the class of criminal acts and to the different opinion with respect to the role of punishment and detention. More educated people are also those that more frequently deal with white-collar crimes, a sort of crime that is more widespread and tolerated in the upper class than assaults and physical violence.

This raises important questions about the social dimension of crime as a complex phenomenon involving the individual's life within society. On the other hand, more educated people tend to give the punishment a role that is more rehabilitative than punitive; all these aspects may explain why education moderates attitudes toward crime. However, with respect to crime, the link between attitudes and behavior remains under-investigated in comparison with other social objects. Some questions are still open. In more detail, it would be questioned whether education and wealth work as a consciousness that is an alternative to law to establish what is right and what is not right. On the other hand, does direct experience of corruption and poverty make people desperate to recognize law and legality as the only hope for a better life? And, more importantly, will people behave according or against to their attitudes? What seems to emerge from the literature is that attitudes are probably not enough to understand and to predict future behaviors but may work as an important starting point for any serious analysis of this phenomenon.

References

- Ajzen I (1991) The theory of planned behaviour. *Organ Behav Hum Decis Process* 50:179–211
- Arrow K (1997) The benefit of education and the formation of preferences. In: Behrman J, Stacey N (eds) *The social benefits of education*. University of Michigan Press, Ann Arbor, pp 11–15
- Barille L (1984) Television and attitudes about crime: do heavy views distort criminality and support retributive justice? In: Surette R (ed) *Justice and the media: issues and research*. Charles C. Thomas, Springfield
- Buonanno P, Leonida L (2006) Education and crime: evidence from Italian regions. *Appl Econ Lett* 13:709–713
- D'Agostino E, Sironi E, Sobbrío G (2013) The role of education in determining the attitudes towards crime in Europe. *Appl Econ Lett* 20:724–727
- Groot W, Van Den Brink HM (2010) The effects of education on crime. *Appl Econ* 42:279–89
- Liou D, Bauer KD (2007) Exploratory investigation of obesity risk and prevention in Chinese Americans. *J Nutr Educ Behav* 39(3):134–141
- Torgler B, Schneider F (2007) What shapes attitudes toward paying taxes? Evidence from multicultural European countries. *Soc Sci Q* 88(2):443–470

Crime, Incentive to

Derek Pyne

Thompson Rivers University, Kamloops, BC, Canada

Abstract

The standard economic model of crime assumes criminal incentives depend on (1) the expected payoff from crime, (2) the penalty if apprehended and convicted, (3) the probability of apprehension and conviction, and (4) attitudes toward risk. At times, penalties for failed criminal attempts are also taken into account.

Definition

Criminal incentives either motivate individuals to commit crimes or motivate them to abstain from crime. The key positive incentive to commit crime is the expected benefit a criminal receives from the crime. Key negative incentives involve deterrence

from the probability of detection and the resulting penalty. Increases in the opportunity cost of committing crime rather than engaging in regular employment also act as a negative incentive.

Criminal Incentives

The most common behavioral assumption in economics is that economic agents respond to incentives. Without this assumption, the economics of crime would not exist as an area of study.

The simplest model of a risk-neutral criminal's objective function has the following form:

$$U = B - qD \quad (1)$$

where B represents the benefits of crime, q represents the probability of apprehension and conviction, and D represents the disutility of the penalty if convicted.

B may represent either psychological or financial benefits. The entry by Leroy (2014) deals with a similar distinction between expressive crime and instrumental crime. The entry by Schneider and Meierrieks (2014) discusses the nonfinancial benefits of Terrorism. As those entries focus on expressive crime, this entry will concentrate on crimes with more tangible benefits.

If a criminal has the choice between working and committing crimes, B may be measured relative to his income from employment. If so, B may be a function of employment prospects, the minimum wage, and the criminal's education. Both a theoretical and an empirical finding is that unskilled crimes (e.g., violent and property crime) are negatively related to education and white-collar crimes are positively related to education (Lochner 2004). This is consistent with more educated individuals having greater opportunities (and hence incentives) for white-collar crimes. However, their higher market incomes result in lower benefits, relative to opportunity costs, for other crimes.

B also depends on the opportunities available. These opportunities decrease as electronic

transactions replace cash (Wright et al. 2014). This may be a reason for recent falling crime rates.

The disincentive of punishment (D) can include imprisonment, fines, forfeiture, stigma, poor human capital accumulation, and other penalties. The entry by Prescott (2015) surveys a range of criminal sanctions and their implications for deterrence. The entry by Di Vita (2015) contains a discussion of sanctions and deterrence implications for environmental crime. The entry by Vannini et al. (2015) does the same for kidnapping. The entry by Donohue (2014) focuses on the specific case of capital punishment.

Poor criminals may not have the resources to pay fines. If so, fines are less of a deterrent than imprisonment. The opportunity cost of imprisonment will partly depend on forgone wages. Thus, the disincentive effects of both imprisonment and fines may be positively related to income.

With incarceration, poor prison conditions increase D and discourage crime (Katz et al. 2003; Pyne 2010). However, there is evidence that incarcerated criminals are more likely to reoffend when prison conditions are poor (Drago et al. 2011). This may be because harsher prison conditions lead to a deterioration of human capital and poorer future employment prospects. In other words, a reduction in prison conditions may lead to a contemporaneous increase in D while increasing future values of B (relative to the opportunity cost of working instead) for those incarcerated.

With incarceration, there is also evidence of an intertemporal relationship between B and D through prison increasing criminal capital (Bayer et al. 2009).

The stigma of conviction increases D . Stigma effects may work through social ostracism or poorer future employment prospects. Typically, it is assumed that stigma effects are greatest for the first conviction and then decline. If so, *ceteris paribus*, a criminal's incentives to commit crime increase after a first conviction. This is one explanation for the observation that most formal penalty schedules involve harsher penalties for repeat offenders. The increase may be necessary to counter the decreasing deterrent effects of stigma (Miceli and Bucci 2005).

Somewhat similar to stigma effects are effects on human capital. There is evidence that incarceration decreases the ability of juveniles to acquire human capital. If so, their wages as adults are lower, which lowers their future opportunity cost of incarceration. Pyne (2010) offers this as an explanation for the more lenient treatment of juveniles in most justice systems.

A wide range of other penalties can result in disutility. Examples include impaired drivers being barred from driving, licensed professionals losing their license, problems crossing borders due to convictions, and having to compensate victims.

The probability (q) of being caught and convicted for a criminal act provides a negative incentive to commit crime. This may be a function of a criminal's ability (Miles and Pyne 2015), reporting by victims and witnesses (Allen 2011), law enforcement expenditures (Vollaard and Hamed 2012), prosecutor behavior (Entorf and Spengler 2015), certainty of conviction (Entorf and Spengler 2015), and the standard of proof required for convictions (Pyne 2004).

Empirical evidence suggests that for many crimes, q is small. For example, based on data from a British longitudinal survey, Farrington et al. (2006) find that when assaults are excluded, on average there were 14 self-reported crimes for every conviction. When assaults are included, the ratio of self-reported crimes to convictions is 22.

Higher-ability criminals face a lower probability of being convicted for their crimes. This gives them a greater incentive to commit crimes (This assumes they are not also proportionally better at earning income from noncriminal activities.). Thus, it is possible that those with more convictions may have committed proportionately more crimes than those with fewer convictions (Miles and Pyne 2015).

In some cases, there may be free riding by enforcers which results in a less than optimal q . This, and related enforcement problems, are discussed in the entry by Hallwood and Miceli (2014) on modern piracy.

Whether the probability of being wrongly convicted of a crime changes the disincentive

effects of q is an unsettled question. Many argue that the likelihood of being wrongly convicted of a crime reduces the net increase in the probability of being convicted of a crime when one actually commits the crime (Pyne 2004; Polinsky and Shavell 2007; Rizzolli and Stanca 2012). However, some have argued to the contrary. For example, Lando (2006) argues that one may be wrongly convicted of a crime committed by someone else, independently of whether one has committed a crime of their own.

Attitudes towards risk affect the relative deterrent effects of penalties and conviction probabilities. Equation (1) assumed individuals were risk-neutral. Thus, changes in penalties should have the same deterrent effect as changes in conviction probabilities that have the same effect on qD . If criminals are risk-averse, changes in penalties should have greater deterrent effects than changes in conviction probabilities. In either case, if enforcement and punishment is costly, it would be best to have punishments that do not have to be enforced (Becker 1968). The basic economic model of crime implies that if it were feasible to set penalties high enough, all crime could be deterred.

Not only are extreme punishments not used in practice but empirical evidence finds that changes in the probability of conviction have a greater deterrent effect than changes in penalties (Eide 2000). Several approaches have been offered to reconcile the basic model with the empirical evidence.

One approach involves explanations consistent with standard expected utility theory. For example, criminals may be risk lovers (Becker 1968). However, this is contrary to standard economic assumptions used in other contexts. Alternatively, Pyne (2012) shows that low-ability criminals may learn more about their innate ability from increased enforcement than from increased penalties, which reveal no information about ability. Another approach is to relax the standard assumptions of the expected utility model (Neilson and Winter 1997).

A different approach is to entirely reject expected utility theory. An example is

Al-Nowaihi and Dhami's (2010) use of composite cumulative prospect theory.

Another complication of the basic model involves the possibility that a criminal's attempt at a crime is unsuccessful. This will, in part, depend on resistance and precautions taken by the victim. Allen (2011) surveys literature related to victim behavior and offers an analysis.

The lower the punishment for unsuccessful attempts, the greater incentive criminals have to target victims who take fewer precautions. Ben-Shahar and Harel (1996) argue that this may be efficient if victim precautions are too high. Those who would otherwise exercise a high level of precaution have less incentive to do so as criminals are targeting those taking a lower level of precautions instead. However, if high-income potential victims can afford greater precautions, distributional considerations may also be involved.

The probability of success may depend on competition from other criminals. In some instances, imperfect enforcement may actually benefit some criminals by reducing competition from competitors (Miles and Pyne 2014).

Expected punishment also affects the criminal's choice between crimes. If the punishment for less socially harmful crimes increases, criminals have an incentive to substitute towards more socially harmful crimes (Stigler 1970).

There has been an increasing focus on behavioral economic considerations of criminal incentives. Typically, behavioral considerations have more of a quantitative effect on incentives than a qualitative effect. For example, hyperbolic discounting changes the value a prospective criminal places on a marginal increase in prison sentences, but the increase is still an incentive to not commit crime. Other behavioral considerations include loss aversion, issues involving probabilities, bounded rationality, and emotional considerations. For surveys of the behavioral economics literature on crime, see Garoupa (2003) and van Winden and Ash (2012). For a general introduction to its use in law and economics, see the entry by Ko (2014).

Cross-References

- ▶ [Becker, Gary S.](#)
- ▶ [Behavioral Law and Economics](#)
- ▶ [Cost of Crime](#)
- ▶ [Crime: Expressive Crime and the Law](#)
- ▶ [Crime: Organized Crime and the Law](#)
- ▶ [Crime and Punishment \(Becker 1968\)](#)
- ▶ [Criminal Sanctions and Deterrence](#)
- ▶ [Death Penalty](#)
- ▶ [Environmental Crime](#)
- ▶ [Piracy, Modern Maritime](#)
- ▶ [Ransom Kidnapping](#)
- ▶ [Terrorism](#)

References

- Al-Nowaihi A, Dhami S (2010) The behavioral economics of crime and punishment. Discussion papers in economics, University of Leicester
- Allen DW (2011) Criminals and victims. Stanford University Press, Stanford
- Bayer P, Hjalmarsson R, Pozen D (2009) Building criminal capital behind bars: peer effects in juvenile corrections. *Q J Econ* 124(1):105–147
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Ben-Shahar O, Harel A (1996) The economics of the law of criminal attempts: a victim-centered perspective. *Univ Pa Law Rev* 145(2):299–351
- Di Vita G (2015) Environmental crime. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Donohue JJ (2014) Death penalty. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Drago F, Galbiati R, Vertova P (2011) Prison conditions and recidivism. *Am Law Econ Rev* 13(1):103–130
- Eide E (2000) Economics of criminal behavior. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics*. Edward Elgar, Cheltenham
- Entorf H, Spengler H (2015) Crime, prosecutors, and the certainty of conviction. *Eur J Law Econ* 39(1):1–35
- Farrington DP, Coid JW, Harnett LM, Jolliffe D, Soteriou N, Turner RE, West DJ (2006) Criminal careers up to age 50 and life success up to age 48: new findings from the Cambridge Study in Delinquent Development, vol 299, Home office research study. Home Office, London
- Garoupa N (2003) Behavioral economic analysis of crime: a critical review. *Eur J Law Econ* 15(1):5–15
- Hallwood P, Miceli TJ (2014) Piracy, modern maritime. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York

- Katz L, Levitt SD, Shustorovich E (2003) Prison conditions, capital punishment, and deterrence. *Am Law Econ Rev* 5(2):318–343
- Ko H (2014) Behavioral law and economics. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Lando H (2006) Does wrongful conviction lower deterrence? *J Leg Stud* 35(2):327–337
- Leroch MA (2014) Crime (expressive) and the Law. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Lochner L (2004) Education, work, and crime: a human capital approach. *Int Econ Rev* 45(3):811–843
- Miceli TJ, Bucci C (2005) A simple theory of increasing penalties for repeat offenders. *Rev Law Econ* 1(1):71–80
- Miles S, Pyne D (2014) The economics of scams. Presented at Canadian Economic Association, Conference, Vancouver
- Miles S, Pyne D (2015) Deterring repeat offenders with escalating penalty schedules: a Bayesian approach. *Econ Gov* 16(3):229–250
- Neilson WS, Winter H (1997) On criminals' risk attitudes. *Econ Lett* 55(1):97–102
- Polinsky AM, Shavell S (2007) Public enforcement of law. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*. Elsevier, Amsterdam
- Prescott JJ (2015) Criminal sanctions and deterrence. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Pyne D (2004) Can making it harder to convict criminals ever reduce crime? *Eur J Law Econ* 18(2): 191–201
- Pyne D (2010) When is it efficient to treat juvenile offenders more leniently than adult offenders? *Econ Gov* 11(4):351–371
- Pyne D (2012) Deterrence: increased enforcement versus harsher penalties. *Econ Lett* 117(3):561–562
- Rizzolli M, Stanca L (2012) Judicial errors and crime deterrence: theory and experimental evidence. *J Law Econ* 55(2):311–338
- Schneider F, Meierrieks D (2014) Terrorism. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Stigler GJ (1970) The optimum enforcement of laws. *J Polit Econ* 78(3):526–536
- van Winden F, Ash E (2012) On the behavioral economics of crime. *Rev Law Econ* 8(1):181–213
- Vannini M, Detotto C, McCannon B (2015) Ransom kidnapping. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York
- Vollaard B and J Hamed J (2012) Why the police have an effect on violent crime after all: evidence from the British Crime Survey. *Journal of Law and Economics* 55(4):901–924
- Wright R, Tekin E, Topalli V, McClellan C, Dickinson T, Rosenfeld R (2014) Less cash, less crime: evidence from the electronic benefit transfer program, vol 19996. National Bureau of Economic Research, Cambridge, MA

Crime, Unemployment and

Kangoh Lee

Department of Economics, San Diego State University, San Diego, CA, USA

Definition

Crime is one of the most important social issues citizens are concerned about. As such, it has received a good deal of attention from policymakers and scholars across fields such as criminology, economics, law, psychology, and sociology. Crime is partly motivated by economic conditions, and a large body of research has studied the relationship between economic factors, particularly unemployment, and crime for more than 50 years. Available evidence shows that the relationship is ambiguous, and it is an actively researched topic.

Introduction

Crime was first analyzed by economists in the late 1960s and early 1970s (e.g., Becker 1968; Ehrlich 1973). The analysis is based on the cost and benefit of crime. The cost to an individual who commits property crime is the lost legitimate incomes and the penalties when crime is caught, and the stolen properties and incomes constitute the benefit (see Anderson 2017 for a general discussion of the cost). The individual then compares the cost and the benefit when deciding to commit crime.

Applying the cost-benefit analysis of crime to the relationship between unemployment and crime, an increase in unemployment during economic downturns should increase crime, as the increase in unemployment decreases the cost of crime, namely, the lost legitimate income. However, this seemingly obvious and intuitive prediction has not been empirically supported. Rather, empirical findings have been mixed, ranging from positive effects of unemployment on crime to no stable or statistically significant effect and to negative effects. The relationship between

unemployment and crime is thus intriguing and one of the most controversial topics in law and economics. At the same time, this discrepancy between economic theory and empirical findings indicates that the unemployment-crime nexus is more complicated than the standard economic theory of crime predicts and calls for more research.

Effects of Unemployment on Crime

Overview

Fleisher (1963) is the first study to examine the effect of unemployment on crime. Based on data for property crimes from FBI Uniform Crime Reports over the period 1932–1961 across Boston, Chicago, and Cincinnati, he finds that the elasticity of the arrest rates with respect to the unemployment rate of young males is between 0.10 and 0.25. Since his work, a large number of research papers have empirically studied the relationship between labor market conditions and crime. These studies typically utilize panel data across jurisdictions such as counties or states during a certain period of time. They typically use instrumental-variable approaches to take into account the possible endogeneity of unemployment and to correct for simultaneity between unemployment and crime and controls for other covariates such as ages, incomes, police expenditures, and metropolitan areas (e.g., Raphael and Winter-Ebmer 2001; Gould et al. 2002; Lin 2008).

The literature on the topic has significantly grown, and a good number of papers have reviewed this growing literature (Long and Witte 1981; Freeman 1983, 1995; Chiricos 1987; Levitt 2004; Blumstein and Wallman 2006). These reviews show that the effects of unemployment on crime are positive, and sometimes statistically insignificant, and sometimes even negative. For instance, Freeman (1983) reviews 15 empirical studies and finds that unemployment has positive and significant effects on crime only in four studies. Chiricos (1987), by contrast, finds that unemployment has significant positive effects on crime in a majority of 63 studies he reviews.

These diverse findings show that the relationship between unemployment and crime is complicated. At the same time, these findings reflect the possibility that the relationship also hinges on the types of crimes and the characteristics of those individuals who commit crime. Crimes of passion such as murder and rape would depend less on economic conditions than crimes driven by economic incentives such as burglary and other types of property crimes. The youths have fewer economic opportunities than adults, and their motivation to commit crime may differ from adults'. Social customs and cultures that may govern the incentives to commit crime play a role in shaping the relationship. It is thus useful to examine the relationship between unemployment and crime separately for different types of crimes, for different groups of individuals, and for different countries.

Types of Crimes

There are different ways of categorizing crimes, but for the purpose of this entry, two types of crimes, violent crimes and property crimes, are relevant. According to the FBI Uniform Reporting Program, violent crimes are the offenses involving force or threat of force and consist of murder, rape, robbery, and aggravated assault. Property crimes do not involve force or threat of force against the victims and include burglary, larceny, motor vehicle theft, and arson.

The standard economics argument dictates that an increase in the unemployment rate in general should increase property crime but may not have much effect on violent crime. The reason is that those who commit property crime are interested in the properties of the victims rather than the victims themselves, and the economic conditions of the offenders such as their wages and employment status and those of the properties such as their market values affect the incentives to commit property crime. By contrast, violent crimes are the offenses against the victims, and the economic conditions of the offenders or the victims would not affect the incentives to commit violent crime in an important manner.

Recent empirical studies have on average confirmed the standard economics argument, but have

found that unemployment does not have uniform effects on property crime or violent crime (Raphael and Winter-Ebmer 2001; Gould et al. 2002; Lin 2008). In particular, an increase in unemployment is on average associated with a significant increase in property crime but has little effect on violent crime. However, an increase in the unemployment rates increases assault and robbery, but decreases murder and rape (Raphael and Winter-Ebmer 2001). In addition, the unemployment rate has a positive and significant effect on burglary and larceny, but a negative and significant effect on auto theft (Gould et al. 2002). Thus, unlike the standard economics argument, an increase in unemployment may affect some types of violent crimes in a significant manner and may decrease rather than increase some types of property crimes.

As these studies show, crimes driven by economic incentives, particularly burglary, appear to be strongly affected by unemployment, but the linkage between unemployment and violent crime is weak. If unemployment has any stable and expected effect on violent crime, it would affect robbery in a predictable manner, because robbery is a violent crime but close to property crime in its nature.

Different Countries

Diverse empirical findings may be attributed to different cultures and social norms, as the decision to commit crime may depend in part on cultures. This section discusses empirical findings across different countries to separate the effect of unemployment from the role of cultures and other social influences in crime. The literature mentioned above focuses on US data, and other countries and cross-country studies are considered below.

A number of studies have examined the link between unemployment and crime in Europe. In Germany, the effect of unemployment differs between before and after reunification. In particular, unemployment has little effect on crime in West Germany but has significant and positive effects on robbery and theft in reunified Germany (Entorf and Spengler 2000). In Sweden, unemployment has the expected effects on crime in the sense that unemployment has a positive and

significant effect on property crime, but no significant effect on violent crime (Edmark 2005). In England and Wales, an increase in unemployment decreases fraud and increases drug and other crimes among ten police-recorded crimes (Wu and Wu 2012).

As for countries other than the USA and European countries, a positive and significant relationship between unemployment and crime has been found in New Zealand (Papps and Winkelmann 2000), in Malaysia (Tang 2011), and in some Latin American cities but not in other Latin American cities (Hojman 2004). As for multi-country studies, Wolpin (1980) explores three-country data, Japan, the UK, and the USA, and demonstrates that unemployment has a positive effect on crime, and Altindag (2012) analyzes a set of European cross-country data and shows that unemployment increases crime.

The above studies show that the unemployment-crime connection has been extensively investigated across countries, reflecting global interests in the topic. At the same time, judging from their findings, it appears that the relationship between unemployment and crime does not crucially depend on countries in the sense that unemployment largely tends to increase property crime.

Groups of Individuals

Youth crime has been an important subject of crime research in general, as youths are more likely to commit crime than adults and face different economic opportunities. A number of studies consider the effect of youth unemployment on youth crime, and the effect appears to differ from the standard results in the literature above. In particular, it has a long-run positive effect on fraud, homicide, and motor vehicle theft but not on other types of crimes in Australia (Narayan and Smyth 2004), and it increases burglary, theft, and robbery, but not violent crime in England and Wales (Carmichael and Ward 2001). In France, the unemployment rates of 15–24-year-olds increase almost all types of crimes, including violent crimes, but the unemployment rates of 25–49-year-olds decrease almost all crimes, again including violent crimes, although the effects of the

unemployment rates of the latter group are not statistically significant (Fougère et al. 2009).

A rare study, based on Florida county-level data, investigates the effects of unemployment on reconviction for black males and for white males separately (Wang et al. 2010). They find that black ex-prisoners released to areas with high black-male unemployment rates are more likely to commit violent crime, but no such effect is found for white counterparts. They also find that white ex-prisoners released to areas with high white-male manufacturing employment rates are less likely to commit violent crime, but no such effect is found for black counterparts. It is interesting to observe that unemployment rates or employment rates of the areas have no effect on the likelihood of ex-prisoners committing property crime.

The effects of youth unemployment on youth crime appear to differ in the sense that youth unemployment also affects youth violent crime at least in France and Australia. Many studies include the proportion of blacks in a county or region or state as a covariate in their regressions, but few studies consider blacks and whites separately, and the study by Wang et al. (2010) shows that significant differences exist between blacks and whites in terms of the effects of unemployment on recidivism. Another group of individuals of interest are females. Females are known to be less likely to commit crime than males, but no research appears to examine the relationship between female unemployment and female crime.

Moderating Factors

Many studies on the relationship between unemployment and crime control for other covariates, as noted above, but the relationship itself depends on moderating variables. For instance, the effect of unemployment on crime hinges on the apprehension rate (Lee 2016). In particular, an increase in the unemployment rate decreases the crime rate at low apprehension rates but increases the crime rate at high apprehension rates, as apprehension affects both the cost of crime and the benefit of crime.

In regression analysis, moderating variables are captured by interaction terms. While interaction terms have been rarely studied in the

literature, Baron (2008) includes an interaction term between youth unemployment and monetary dissatisfaction in the youth violent-crime regression and an interaction term between youth unemployment and job search activities in the youth drug-crime regression. The regression results show that an increase in youth unemployment tends to increase youth violent crimes when youths are more dissatisfied with their monetary situations, and it tends to decrease youth drug crimes when they search for jobs more regularly.

Other Labor Market Conditions and Crime

Among labor market opportunities that may affect crime, unemployment has been the most important determinant of crime. However, other labor market conditions also play an important role in crime. An increase in wages of noncollege-educated or college-educated men reduces crimes broadly, including both property crime and violent crime in the USA (Gould et al. 2002). A decrease in the wages of low-wage workers substantially increases all types of property crime in England and Wales (Machin and Meghir 2004).

Income inequality is another factor that affects crime. The way income inequality affects property crime and violent crime differs significantly from the way unemployment does. More inequality significantly increases violent crime but has no effect on property crime in the USA (Kelly 2000). This finding stands in sharp contrast with the result that unemployment has a positive effect on property crime, but little effect on violent crime. One interpretation of the contrast is that property crime is influenced by economic incentives while violent crime is driven by strain and social situations. The income inequality may affect crime differently in a time-series analysis. In particular, an increase in inequality increases crime rates in the cross-section analysis but decreases crime rates in the time-series analysis (Brush 2007). In a study that relates inequality and unemployment to crime in Latin American cities, Hojman (2004) finds that more inequality significantly increases

crime in Greater Buenos Aires. It is difficult to directly compare Kelly (2000) with the last two studies, as they do not distinguish property crime from violent crime, but income inequality appears to influence crime in general although the inequality-crime nexus has not been extensively studied.

Conclusions

Crime causes private costs to victims and criminals and social costs as well, and policymakers have attempted to reduce crime by spending resources. To the extent that there is a connection between crime and economic incentives, any crime policy has to examine such economic incentives. This entry has focused on unemployment as the key part of economic incentives. Despite a large volume of empirical research on the topic, the literature has not reached a consensus on the effects of unemployment on crime even though the standard economics argument dictates that unemployment must have a positive effect on crime. Rather, the effects depend crucially on the types of crimes and on the characteristics of individuals as well. The literature is still growing and expected to attract attention from scholars and policymakers.

While unemployment is important, other labor market conditions such as wage levels and wage inequality appear to influence the decisions to commit crime. Nevertheless, little research on the wage-crime link or the inequality-crime link has been conducted, and more research on the issue is warranted. In addition, many interesting aspects of crime have been omitted in this entry, but other entries discuss them (e.g., Baker 2017; Leroch 2017; Prescott 2017; Pyne 2017).

References

- Altindag DT (2012) Crime and unemployment: evidence from Europe. *Int Rev Law Econ* 32:145–157
- Anderson D (2017) Cost of crime. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York. forthcoming
- Baker M (2017) Crime (organized) and the law. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York. forthcoming
- Baron SW (2008) Street youth, unemployment, and crime: is it that simple? Using general strain theory to untangle the relationship 1. *Can J Criminol Crim Justice* 50:399–434
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 73:169–217
- Blumstein A, Wallman J (2006) The crime drop and beyond. *Ann Rev Law Soc Sci* 2:125–146
- Brush J (2007) Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties. *Econ Lett* 96:264–268
- Carmichael F, Ward R (2001) Male unemployment and crime in England and Wales. *Econ Lett* 73:111–115
- Chiricos T (1987) Rates of crime and unemployment: an analysis of aggregate research evidence. *Soc Probl* 34:187–212
- Edmark K (2005) Unemployment and crime: is there a connection? *Scand J Econ* 107:353–373
- Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. *J Polit Econ* 81:521–565
- Entorf H, Spengler H (2000) Socioeconomic and economic factors of crime in Germany: evidence from panel data of the Germany states. *Int Rev Law Econ* 20:75–106
- Fleisher BM (1963) The effect of unemployment on juvenile delinquency. *J Polit Econ* 71:543–555
- Fougère D, Kramarz F, Pouget J (2009) Youth unemployment and crime in France. *J Eur Econ Assoc* 7:909–938
- Freeman RB (1983) Crime and unemployment. In: Wilson JQ (ed) *Crime and public policy*. Institute for Contemporary Studies Press, San Francisco, pp 89–106
- Freeman RB (1995) The labor market. In: Wilson J, Petersilia J (eds) *Crime: public policies for crime control*. Institute for Contemporary Studies Press, San Francisco, pp 171–192
- Gould E, Weinberg B, Mustard D (2002) Crime rates and local labor market opportunities in the United States: 1979–1997. *Rev Econ Stat* 84:45–61
- Hojman DE (2004) Inequality, unemployment and crime in Latin American cities. *Crime Law Soc Chang* 41:33–51
- Kelly M (2000) Inequality and crime. *Rev Econ Stat* 82:530–539
- Lee K (2016) Unemployment and crime: the role of apprehension. *Eur J Law Econ*. Forthcoming. <https://doi.org/10.1007/s10657-016-9526-3>
- Leroch M (2017) Crime (expressive) and the law. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York. forthcoming
- Levitt SD (2004) Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. *J Econ Perspect* 18:163–190
- Lin M-J (2008) Does unemployment increase crime? Evidence from U.S. data 1974–2000. *J Hum Resour* 43:413–436
- Long SK, Witte AD (1981) Current economic trends: implications for crime and criminal justice. In: Wright KD (ed) *Crime and criminal justice in a declining*

- economy. Oelgeschlager, Gunn and Hain, Cambridge, MA, pp 69–143
- Machin S, Meghir C (2004) Crime and economic incentives. *J Hum Resour* 39:958–979
- Narayan PK, Smyth R (2004) Crime rates, male youth unemployment and real income in Australia: evidence from Granger causality tests. *Appl Econ* 36:2079–2095
- Papps K, Winkelmann R (2000) Unemployment and crime: new evidence for an old question. *N Z Econ Pap* 34:53–71
- Prescott JJ (2017) Criminal sanction and deterrence. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York. forthcoming
- Pyne D (2017) Crime (incentive to). In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer, New York. forthcoming
- Raphael S, Winter-Ebmer R (2001) Identifying the effect of unemployment on crime. *J Law Econ* 44:259–283
- Tang CF (2011) An exploration of dynamic relationship between tourist arrivals, inflation, unemployment and crime rates in Malaysia. *Int J Soc Econ* 38:50–69
- Wang X, Mears DP, Bales WD (2010) Race-specific employment contexts and recidivism. *Criminology* 48:1171–1211
- Wolpin KI (1980) A time series-cross section analysis of international variation in crime and punishment. *Rev Econ Stat* 62:417–423
- Wu D, Wu Z (2012) Crime, inequality and unemployment in England and Wales. *Appl Econ* 44:3765–3775

Crime: Economics of, Different Paradigms

Sven Grüner and Norbert Hirschauer
Agribusiness Management,
Institute of Agricultural and Nutritional Sciences,
Martin Luther University Halle-Wittenberg,
Halle (Saale), Germany

Definition

Economics of crime describes the attempt to explain rule-breaking behaviors based on the assumption that people make purposive choices under conditions of scarcity.

Economics of Crime: Different Paradigms

Economics of crime can be broadly defined as the attempt to explain rule-breaking behaviors based

on the assumption that people make purposive choices under conditions of scarcity. People's choice sets regularly contain both permissible and illicit choices. Illicit choices do not always constitute criminal acts in terms of violations of the criminal law. However, throughout this entry we use crime, illegal behavior, illicit choice, non-compliance, offense, and rule breaking as interchangeable terms for the infringement of formal (codified) rules. **Economics of crime** includes the analysis of **economic crime** and white-collar crime such as corruption, patent infringements, or the violation of industrial safety laws, but it explicitly applies economic approaches to the study of illicit behaviors such as traffic offenses and vandalism, which are often considered being beyond the realm of economics.

Economic approaches share the core understanding that human behaviors are the result of purposive choices – either fully rational (e.g., von Neumann and Morgenstern 1944) or 'intendedly rational' (Simon 1955: 114). However, there is no uniform conception of economic man. Instead, differing models – from purely materialistic and self-interested **rational choice** to multigoal and **bounded-rational choice** – are used to study behaviors in different contexts. This entry systematically discusses competing conceptions (paradigms) of illicit choice.

Beccaria, Bentham, Becker

Gary S. Becker (1968) is one of the most prominent, but by far not the first one to study both crime (self-interested private choice) and crime prevention (welfare-oriented public choice) from an essentially economic (utilitarian) perspective. There are many precursors of Becker in modern history who applied the logic of utilitarian calculus to the study of crime and its **prevention through deterrence**. The most noteworthy one is Beccaria whose famous treatise *Dei delitti e delle pene* (*On Crime and punishment*; 1764/1995) combined utilitarian reasoning with the Enlightenment call for the rule of law and the plea against excessive and disproportionate punishment. Drawing on Hobbes' social contract

theory (1651/2010), Beccaria justified state-imposed punishment as the necessary means to enforce the social contract as codified in the law. Beccaria's ideas on punishment heavily influenced Bentham (1789). Beccaria and Bentham agreed on three fundamentals: (1) the need for legal formalism to justify the justice system, (2) the rejection of capital punishment due to its brutalizing effects on society, and (3) the endorsement of a preventive instead of a retributive rationale in sentencing that limits punishment (Harcourt 2014). Antedating modern-day understanding of deterrence as being both specific and general, Beccaria (1764/1995: 31) detailed that the only purpose of punishment is to "prevent the offender from doing fresh harm to his fellows and to deter others from doing likewise." He also noted that, from a utilitarian point of view, there is a moral-philosophical limitation to punishment: "If a punishment is to serve its purpose, it is enough that the harm of punishment should outweigh the good which the criminal can derive from the crime, [. . .]. Anything more than this is superfluous and, therefore, tyrannous" (Beccaria 1764/1995: 64).

The general idea of **Beccarian deterrence** is that presumptive criminals will respond to the **severity** ("size") of punishment, its **certainty** (probability), and its **celerity** (swiftness). Beccaria and Bentham strongly advocated mild but highly probable and swift punishment. They argued that certainty and swiftness are necessary to make potential criminals realize that crime is closely associated with punishment. In their understanding, the fear of a more severe but less likely and delayed punishment, which inevitably goes hand in hand with the hope of not being captured at all, will deter crime less effectively. They furthermore noted that overly severe penalization might harden criminals and provoke follow-up crimes to avoid capture and hard punishment.

Neoclassical accounts of crime based on twentieth-century microeconomic theory, such as those of the Chicago School of economics and notably Becker (1968), took up the Beccarian idea of utilitarian calculus and deterrence. Becker modeled individuals as materialistic,

self-interested, and rational decision makers who maximize their utility. In line with the **neoclassical paradigm**, Becker supposed people's preferences to be context-independent ("exogenous") and stable. He assumed that individuals calculate and weigh the material benefits (utility) and costs (disutility) of rule breaking and rule abidance without moral considerations. Based on this assumption, Becker arrived at normative regulatory conclusions through the **externality** argument. He contended that the **expectation value of the sanction** ("expected sanction") – as resulting from its size and probability – should be set at a level that makes would-be offenders internalize the negative externalities (harms) they would inflict on the victims. Using the language of price theory, Becker claimed that, to decrease the "supply" of crime, regulators need to engage in monitoring and sanctioning activities (enforcement) to reduce the excessive "price" that criminals could otherwise obtain from illegal activity. Regulators should undertake these law enforcement activities up to the level where their marginal social costs equal the marginal social costs of criminal behavior. This implies tolerating a certain amount of crime because, at some point, the costs of more law enforcement would exceed its benefits.

Becker's normative claim that sanctioning should make offenders internalize the harms inflicted on others was challenged from within the neoclassical camp. Posner (2014) claimed that the expected sanction should prevent crime altogether, by fixing it at a level that (marginally) surpasses the offender's illicit profit. According to Becker's *internalization approach*, one should tolerate "**efficient**" crimes (the breaking of inefficient rules) that lead to a more efficient allocation of resources (Harel 2014). A polluter whose illicit profit exceeds the environmental damage should not be deterred, for example. In contrast to Becker but in line with Beccaria, Posner's *prevention approach* implies that any crime (injury of a legally protected interest) should be prevented by turning it into an inferior option for the agent deliberating it.

Becker focused on "**taxing**" rule breaking as the means to reduce its "price" relative to the one

attached to legal activity. However, as noted by himself, the attractiveness of rule breaking can also be reduced by “**subsidizing**” rule abidance and increasing the opportunity costs of crime – for example, by providing education and legal income opportunities to destitute people who are at risk of becoming offenders. The deterrence argument leaves open the empirical question whether money spent on monitoring and sanctioning (taxing illegal behavior) deters crime more efficiently than money spent for subsidizing legal behavior. Nonetheless, Becker and many other advocates of “negative” deterrence seem to focus from the outset on harsh penalization.

While Becker (1968) praised Beccaria as the first one to have used economic calculus in the study of crime, their ideas differ markedly: first, Becker favored capital punishment. Second, he did not elaborate on the effects of the celerity of capture and punishment. Third, he did not believe that *low-size-high-probability sanctioning* (**Beccarian deterrence**) deters crime more effectively than *high-size-low-probability sanctioning*, as proposed by himself (**Beckerian deterrence**). Based on expected utility theory, he argued that Beccarian deterrence would only work better if people were risk preferring (which is not plausible). In contrast, increasing the size of punishment and decreasing its probability, while maintaining its expectation value, will reduce the utility of illegal behavior and thus deter crime more effectively if people are risk averse (which is plausible). Becker claimed that his sanctioning regime would be superior even if people were risk neutral because it would deter crime as effectively as Beccarian deterrence but allow regulators to economize on enforcement costs. Kolm (1973: 266) pointedly highlighted the weak spot of the Beckerian conclusion as “*hang offenders with zero probability.*”

The legal doctrine regarding punishment:

Despite its label, *economics of crime* is commonly seen as encompassing the study of criminal and noncriminal offenses. This loose terminological usage would not be acceptable for legal scholars.

Criminal lawyers will note that the penal code differs substantially from other codes of laws and that its provisions must be limited to reflect society’s most fundamental moral sentiments of right and wrong. Inversely, it is to provide a stigma capable of causing social costs to criminals and shaping the norms of people. That is, less forceful legal institutions such as administrative and tort law should be used to outlaw less serious misconducts. Otherwise, the reciprocal links between the criminal law and society may erode which, in turn, may cause perverse behavioral effects and the degradation of the criminal law system.

Legal doctrine (cf., Radbruch 2011) must reject Becker’s normative conclusions for three reasons. *First*, Becker’s internalization approach transfers the liability rationale of tort law to criminal law. Like administrative law, tort law does not provide a strong stigma. Hence, many ordinary (noncriminal) people find it acceptable to occasionally break civil or administrative law provisions (minor torts or violations of building law, traffic law, etc.). Becker suggested that even crime be accepted if it facilitates an efficient allocation of resources. The criminal law, however, must stigmatize crime – be it “efficient” or not – more than other transgressions because crime not only reallocates resources but also violates the inalienable individual rights of a person and the universal rights of society as a whole. *Second*, legal scholars will not endorse the all-dominant role of the expectation value of the sanction, and they will reject the suggestion to adjust the size of a sanction contingent on its probability of being imposed. The justice system must not only produce deterrence but also be proportionate and fair to the individual person (proportionality principle). The rule of law requires a transparent uniformity of the law aimed at meeting equal crimes with equal punishment. This precludes subjecting individuals

committing minor transgressions such as parking offenses to extreme punishment (“hanging”) on the sole ground that detection probability is low. *Third*, practically minded legal scholars will add that it is not possible to differentiate sanctions contingent on specific enforcement contexts. Imagine a city council that reduces the number of police officers for budgetary reasons. In Becker’s view, this would require adjusting the sanction catalog in this city and punishing offenders, who are caught despite slack enforcement, more severely. How should the lawmaker learn about each city’s specific circumstances? What about law adjustment costs? Moreover, how should presumptive offenders learn about the sanctions in place?

It seems that the normative recommendations of conventional *economics of crime* are often not feasible from a legal doctrine point of view. This is because conventional *economics of crime* pays notoriously little attention to legal requirements that limit the set of deterrence strategies available to regulators. Following its recommendations would more often than not violate fundamental legal principles and compromise the rule of law.

Neoclassical approaches to deterrence are not naïve. They do not deny that individual behaviors are frequently incompatible with the axioms and predictions of rational choice. They contend, however, that deviations from the standard model are minor or at least symmetrical (Korobkin and Ulen 2000). In their view, it is therefore adequate to model people as rational and self-interested utility maximizers even if there is not a single individual in the world who is a fully rational egoist. For illustration sake, let us look at the “supply” of a criminal act such as theft. If people’s evaluations were randomly distributed around a mean defined by a risk averse rational egoist, the utility obtained from theft (and thus its “supply”) could be predicted to decrease

if regulators, while maintaining the expectation value of the sanction, compensated low risks of punishment with high sanction levels. However, this will not hold if the regulatees’ *perception* of the detection risk is *asymmetric* because they underrate or even neglect small probabilities. In this case, Beckerian (high-size-low-probability) sanctioning will prevent less crime than Beccarian (low-size-high-probability) sanctioning. If people’s judgments are affected by such **systematic biases** (either generally or in certain social groups), understanding these biases will enable regulators to better predict and eventually influence people’s behavior.

In Beccaria and Bentham’s view, highly certain and swift punishment is needed to make people fully *perceive* the association of crime with punishment. These early pioneers seem to have anticipated cognitive features that behavioral economists today would classify as forms of bounded rationality, namely, “**nonlinear weighting of probabilities**” and “**hyperbolic discounting**.” Beccaria and Bentham apparently realized that people might underrate or neglect small probabilities and be biased towards the present. When both effects combine, people will attribute a high value to the immediate benefits of crime and a very low or no value to its high but uncertain future costs. This seems a plausible explanation of many rule breaking instances including illicit drug use with its mainly self-inflicted future punishment of bad health or even premature death. Nonetheless, Becker disregarded bounded-rational judgments and was convinced that one can dispense with all other theories including those of “psychological inadequacies.” He believed *his* theory of crime and emphasis on harsh punishment to be an improvement on not only Beccaria but all crime scholars who do not exclusively rely on the neoclassical paradigm. Due to frustrating experiences with harsh punishment, many scholars today think that Beccaria was right to advocate mild but certain and swift punishment. Nevertheless, Becker and his followers, backed up by the electorate’s liking for harsh punishment, seem to have had an influence that partly reverted crime policies, especially in the USA, back to premodern penalization

regimes with capital punishment, long prison sentences, and mass incarceration. Even though heavily criticized on methodical grounds, a time-series analysis by Ehrlich (1975), claiming that capital punishment substantially reduced capital crimes in the USA, fueled this development. Using more adequate econometric approaches, more recent studies did not find that capital punishment deters capital crime (e.g., Zimmerman 2004; Kovandzic et al. 2009).

Behavioral Economics of Crime

There is considerable evidence from experimental and observational studies that, for various reasons, people’s behaviors deviate substantially *and* systematically from rational choice predictions (Thaler 2016). This is why regulatory strategies exclusively based on the neoclassical paradigm, which is still mainstream in law and economics, are a socially inefficient way to prevent illicit behaviors in many contexts. An educated guess would be that they could work for tax offenses and securities fraud but less so for street and property crimes or drug offenses.

A **behavioral-economics-of-crime** analysis that provides a reconstructing understanding of perpetrators’ judgments is needed to close the gap between rational choice predictions and actual behavior. Instead of relying on rational choice and objective facts, **reconstructing understanding** implies trying to understand people’s options of choice and calculi according to their subjective perceptions and evaluations. As Rubinstein pointed out (1991: 910), “these

[perceptions] need not necessarily represent the physical rules of the world.” As people are heterogeneous in both their goals and judgments, making reliable predictions and identifying adequate enforcements strategies is a challenging task. It is also a promising task because many laws address specific groups whose members share behavioral regularities. Such regularities can be identified and considered when specifying contextual enforcement strategies. Adequate strategies will substantially differ, for example, between fields such as securities legislation, traffic law, or drug law. A promising regulatory potential termed “**nudge**” by Thaler and Sunstein (2008) arises from the fact that people’s behaviors often depend on how contexts are described or “framed” (Tversky and Kahneman 1981). Finally, people’s goals and judgments change and can be shaped (to a certain degree) over time.

Table 1 itemizes a selection of deviations from rational choice that seem especially relevant for the study of illegal behaviors. The large number of “anomalies” – as they would be labeled from the neoclassical point of view – forces us to limit this entry to a brief and focused description of those deviations that seem to be most relevant for understanding compliance decisions. Further descriptions regarding behavioral economics applications with respect to noncompliance can be found in Dhami (2016), Thaler (2016), or Korobkin and Ulen (2000), for example.

Multiple Goals, Social Norms, and Bounded Self-interest

People’s utility does not exclusively hinge on material outcomes including leisure and

Crime: Economics of, Different Paradigms, Table 1 Deviations from rational choice relevant for the study of crime

Multiple goals, social norms, and bounded self-interest	B o u n d e d - r a t i o n a l j u d g m e n t s		
	Cognitive biases	Heuristics	Bounded self-control
Striving for conformity with external social expectations	Non-linear weighting of probabilities	Habits	Hyperbolic discounting
Striving for consistency with internalized values and norms	Overconfidence bias	Adhering to defaults	Emotions
	Loss aversion	Mental accounting	Limited attention

conveniences. In contrast, people have complex **multidimensional goal systems** that are influenced by the apparently innate human urge to repay both good and bad experiences in kind (**reciprocity**). Besides the self-interested pursuit of material benefits, people particularly strive for two types of intangible outcomes: (1) conformity with external social expectations and (2) consistency with internalized values and norms. These intangible goals can limit people's self-interest (**bounded self-interest**). Different labels have been attached to these two behavioral drivers. Psychologists often speak of an extrinsic as opposed to an intrinsic motivation to comply. Coming from an applied management research background, Nielsen and Parker (2012) use the terms "social preferences" as opposed to "normative preferences." Adopting an institutional economics perspective on rule-governed social life in general, Ostrom (2005) speaks of external as opposed to internal "*delta* parameters" to indicate that social norms have to be considered as behavioral drivers in *addition* to material incentives.

Criminology with its primary focus on the law and law enforcement provides still another terminology and talks of "protective factors" and "risk factors" (Hirschauer and Scheerer 2014). **Protective factors** are bonds to social norms that discourage illicit choices by causing non-material costs for rule-breaking and nonmaterial benefits for rule abidance. Protective factors can be seen as mechanisms that impose "intangible taxes" on illegal acts and provides "intangible subsidies" for legal acts. These taxes and subsidies arise from external social control (social disapproval/ostracism vs. social approval) as well as from internalized values (self-disrespect/guilt vs. self-esteem). Protective factors can be related to the concept of **resilience**. Instead of seeing people as being either law-abiding or criminal, this implies understanding people as being more or less resilient to moral hazards in that they have a positive but limited willingness-to-pay for rule abidance in exchange for social approval and a feeling of integrity.

In contrast, **risk factors** arise in deviant subcultures in which social norms are not in line with those of the law-approving "conventional"

society. Examples are youth gangs, organized crime communities, or groups hostile to state authority due to extreme religious or ideological beliefs. Like protective factors, risk factors can arise from external and internal sources. They act as "intangible taxes" for legal behaviors (disrespect by deviant peers and conflict with the deviant internal self) and as "intangible subsidies" for illegal behaviors (respect by deviant peers and affirmation of the deviant internal self). Altruistic rule-breaking and **reactance** (Miron and Brehm 2006), for example, are internal risk factors resulting from a deviant self. Even though frequently neglected, risk factors and protective factors are often decisive for compliance decisions. Group-specific social pressures and rewards to break the law, in conjunction with weak conventional norms against illegal behavior or outright deviant selves, for example, are likely to be more relevant causes of juvenile delinquency, such as illegal road races or vandalism, than illicit profits. The interplay of protective factors and risk factors is also crucial for understanding how individual employees behave within deviant corporate cultures (multiple-selves problem).

Crime prevention is an intentional manipulation of factors that determine people's behavior with regard to the law. Three ideal-type strategies are distinguished (Picciotto 2002): **incapacitation** (e.g., through license withdrawal or jailing) is aimed at reducing would-be perpetrators' opportunities to commit offenses. **Deterrence** (e.g., through monitoring and sanctioning) is to reduce the economic temptations (incentives) for rule breaking. **Accommodation** (e.g., through persuasion and advice) is to bolster people's propensity to comply by strengthening the social norms that back up the rules (protective factors). For example, informing defaulting taxpayers that the majority of people in the same town have already paid their tax debts was found to have a positive effect on tax compliance (The Behavioural Insights Team 2011).

Identifying adequate prevention strategies requires a normative analysis based on conditional forecasting. In this forecasting exercise, the behavioral effects of economic incentives and norm-based "taxes" and "subsidies" cannot

be considered in isolation. Reducing the economic attractiveness of rule breaking through tight controls and harsh sanctions may reinforce social norms in some instances but weaken them in others. The label “**crowding in**” describes interventions that simultaneously provide the “right” incentives and strengthen desirable social norms. Evidence from fields as far apart as industrial safety and tax legislation indicates that successful strategies manage to crowd in (Braithwaite 2009). They avoid the dysfunctional effects of oppressive control and harsh sanctions (e.g., reactance) as well as the negative effects of overly lenient approaches (blurring of standards). In contrast, “**crowding out**” describes interventions that involuntarily weaken desirable social norms in the attempt to reduce the economic temptations for rule breaking (Frey 1997). An illustrative example comes from a field experiment in daycare centers (Gneezy and Rustichini 2000). After a small fine was introduced for collecting children late, the frequency of late pickups increased. The interpretation by Sandel (2012) is that under the old regime parents tried hard to be in time because they were ashamed to cause inconvenience for the staff. After the introduction of the fine, which doubtlessly increased the economic costs of coming late, parents felt no shame anymore but came to see late pickups simply as an extra service for which they paid a “fee.”

The intended primary effect of criminal sanctioning is often not economic deterrence. Instead, one hopes that the social stigma provided by the penal law, such as banning the beating of children, will reinforce desirable social norms. Including offenses into the penal code may backfire if the values of the lawmaker and of those subjected to the law do not correspond. Criminalizing activities that many people do not consider “real” wrongs may weaken the criminal law’s general capacity of cultivating social norms. What is more, it might provoke **perverse effects** beyond those caused by the weakening of crime-inhibiting social norms. For example, criminal sanctions on the use of soft drugs, while increasing its cost, might actively encourage substance abuse if users develop individual reactance or even group norms that favor defying what they

consider illegitimate and oppressive government intervention.

With a focus on corporate governance and recidivism, Braithwaite (2002) emphasized that a transparent progression of regulatory responses – accommodation first, deterrence second, incapacitation third – is needed to fully exploit the steering potential of regulatory enforcement. Using the terms “**enforcement pyramid**” and “**responsive regulation**,” he claimed that regulators should meet noncompliance always with a clear disapproval and the request to remedy the wrong. They should use increasingly severe deterrent measures (warning letters, tightening controls, increasing fines) if offenders continue to break the rules. As a last resort, recidivists should be subjected to increasing levels of incapacitation (license suspension, license revocation, jailing). Braithwaite’s key message is that regulators should aim to reintegrate offenders into rule-abiding society by using *transparent* and *graduated* response measures contingent on the degree of bad/good conduct. While regulators should always start softly, the concept of responsive regulation, which is inherently aimed at crowding in, is not leniency. On the contrary, it assumes that the harsher the available ultimate sanctions, the more likely regulators will achieve compliance through persuasion. The recommendation to regulators would therefore be to “speak softly, while carrying very big sticks” (Ayres and Braithwaite 1992: 40).

Bounded Rational Judgments

Besides social norms and bounded self-interest, bounded-rational choice needs to be considered when trying to understand and steer people’s behavior. The term “**bounded rationality**” was coined by Simon (1957) to describe that individuals have limited access to decision-relevant information and that they have limited abilities to process that information. Consequently, choices are often biased and may depend on the presentation of the decision problem (framings) even if the framings have no effect whatsoever on outcomes. To reduce complexity and cope with their bounded-rationality, people often use simplifying decision rules or “heuristics”

(Gigerenzer and Todd 1999). **Cognitive biases, heuristics, and bounded self-control** can lead to choices that are contrary to rational choice expectations according to which fully informed people judiciously weigh the costs and benefits associated with their choices.

Nonlinear weighting of probabilities: Expected utility theory assumes linearity in the weighting of probabilities. In contrast, Kahneman and Tversky (1979) claimed, on the one hand, that people frequently underweight high probabilities and overweight low probabilities. On the other, they noted that people's perception is reversed near the end points in that very high probabilities are perceived as certain, whereas very low probabilities are underrated or even ignored. In the context of crime, the latter misconception is sometimes fueled by "optimism bias" (Weinstein and Klein 1996), which describes a person's belief to be less prone to risks and negative events than others. An important source of such misconceptions are premature generalizations from personal experiences that are easily mentally available and thus salient ("availability bias"). Illicit behaviors associated with low sanctioning probabilities, such as traffic offenses or free-riding public transport without a ticket, are illustrative examples. They may become business as usual for offenders who repeatedly experience no controls. The ostensible remedy of making the risk of being caught more salient by disclosing the number of caught offenders may produce perverse effects, however. For example, while such a piece of information clearly points out to free-riders that the probability of being caught is *not* zero, nondelinquent individuals might adjust a previously overrated detection probability downwards which, in turn, might produce some new free-riders. What is more, obtaining the knowledge that free-riding is a common phenomenon might erode the social norm that backs up the formal obligation to purchase a ticket. In this context, Popitz (1968) spoke of the "preventive effect of ignorance."

Overconfidence bias: Offenders are not average representatives of the population. At least a certain subset of wrongdoers may overestimate their capabilities and believe that they are more apt and clever to commit crimes and avoid

punishment than others. What is more, they may also have the disposition to interpret even quite unambiguous information in ways that fall into place with their preconceived notions ("self-serving bias"). Overconfidence and self-serving bias can explain "silly" crimes for which no explanation in terms of the criminal's self-interest can be found. Deterrence strategies that target presumptive offenders who exhibit overconfidence and/or self-serving bias need to strengthen the salience of controls and sanctions. This may require increasing controls and sanctions beyond the level that would be needed to deter unbiased offenders.

Loss aversion: People often do not think in terms of final states but rather in terms of changes that they perceive as gains or losses with respect to a reference point (Markowitz 1952). Kahneman and Tversky (1979) claimed that people experience losses about twice as strongly as gains. The perception of a change as either a gain or a loss depends on contexts/framings and can affect compliance. Imagine two tax regimes. In regime 1, a taxpayer has to pay 1,000 in income tax per month. In regime 2, the total tax debt of 12,000 is payable after the end of the tax year. In both regimes, tax dodgers run a 25% risk of being detected and fined to pay a penalty of 48,000 in addition to the 12,000 in due taxes. Rational choice theory would predict that taxes are paid lawfully in both regimes because the *expected* reduction in income amounts to 15,000 ($= 0.25 \cdot 60,000$) in the case of tax evasion, but only to 12,000 in the case of lawful tax payment. In addition, tax evasion would increase income *risk*. Things are different if the taxable person exhibits loss aversion. In regime 1, such a taxpayer is likely to perceive the income level *after* regular monthly tax payments as reference point and experience the taxes as nonrealized income gains (opportunity costs). In regime 2, the taxable person is likely to get used to the income level *before* paying taxes and correspondingly adjust the reference point upwards. From the increased reference point, paying the tax debt of 12,000 will be perceived as a certain loss (out-of-pocket costs) that, despite being identical, is experienced as a heavier burden than the taxes in regime 1.

Depending on the strength of this effect, the taxable person in regime 2 will evade taxes and exchange the certain loss for the uncertain chance to escape tax payments. From a rational choice point of view, this implies behaving like a risk seeker. Loss aversion can be related to status quo bias and, more particularly, the endowment effect. The latter describes the fact that people find it generally hard to part with goods and wealth they already enjoy and (believe to) legitimately own.

Habits: Depending on the familiarity with and the relevance of choices, individuals either rely on habitual behaviors or consciously resort to problem-solving procedures (Katona 1953). Each day, people need to make so many choices that they must be made fast. This is why most choices are not the result of analytical procedures. Instead, people cannot help but make most choices habitually and spontaneously. This has been labeled “thinking fast” by Kahneman (2011). Having a habit means making the same choices as in the past when dealing with recurring or similar decision problems. Habits are arguably the most important heuristic without which we would not be able to survive. Habitual behaviors can be important for understanding some types of illegal behaviors. Depending on the individual’s life course, offenses such as speeding or even evading taxes may have become subconscious habits irrespective of whether they provide noteworthy benefits for the offender or not. Habits in conjunction with self-serving bias may pose serious obstacles to effective deterrence, which is inherently based on informing people about controls and sanctions. This is because people are prone to dissolve cognitive dissonances by aligning their interpretation of information to their deep-set habitual behaviors (self-manipulation of beliefs) instead of aligning their behaviors to the information.

Adhering to defaults: Default options seem to exercise a stronger influence on people’s behavior than what we would expect when considering the usual small effort needed to opt against the default. An illustrative example is people’s acceptance to become organ donors (Gigerenzer 2010). The willingness to donate organs is considerably lower in countries that use “no organ donation” as default compared to countries that apply the

nudge device of making “organ donation” the default option. A plausible explanation is that individuals orient themselves to what they perceive as the socially desired standard. People’s preference for the as-is state (the default) can also be understood as a habit or a special form of the “status quo bias.”

Mental accounting: People occasionally classify their choices into categories or “accounts” (health account, environment account, etc.) within which they offset their choices against each other (Thaler 1999). Imagine dog owners who mentally enter “paying dog tax” and “cleaning up when walking the dog” to the same account. After the introduction of a dog tax, they might feel that they can offset their cleaning up against their paying the dog tax. The important message is that introducing new regulations can have negative side effects that are easy to overlook. The effect produced by mental accounting looks like crowding out at first view. The difference is that mental accounting describes how the levels of *different* activities, which people see as belonging to the same “account,” affect each other. Crowding out, in contrast, looks at *one* activity and describes that incentives might involuntarily weaken favorable social norms.

Hyperbolic discounting: The conception of rational choice includes the concept of time preferences that assumes that people discount future outcomes by using their individual rate of time preference. They presumably do so, however, by using constant discount rates per period. In contrast, behavioral economists note that people often exhibit a “*bias* towards the present” in that they prefer the present more than what would be expected from a rational choice perspective. To be more precise, instead of discounting consistently over time, they apply very high discount rates per period for outcomes in the near future but falling discount rates the further in the future the outcomes arise. This has been termed “hyperbolic discounting” (Phelps and Pollak 1968). It constitutes myopia in that people who are prone to hyperbolic discounting consume more in the present than what they had previously planned. Thaler (1991) illustrated the phenomenon by noting that most individuals would inconsistently prefer one apple today to two apples tomorrow but

two apples in 51 days to one apple in 50 days. Crimes associated with drug addiction provide illustrative examples for hyperbolic discounting. If punishment does not follow swiftly, it will not work as a deterrent because addicts attribute a very high value to the fulfillment of their immediate cravings but a very low value to future costs. In the case of strong addictions, juvenile delinquency, and extreme social deprivation, hyperbolic discounting may be so strong that deterrence based on punishment, which is inevitably separated from the offense by a certain time lag, will not work at all. The stronger the presumptive offender's hyperbolic discounting, the swifter the punishment must follow to work as a deterrent.

Emotions: Many emotions arise from people's social interactions and the apparent human need to reciprocate good as well as bad experiences. Emotional behavior can be related to myopia and hyperbolic discounting in that people succumb to immediate behavioral urges without being able to judiciously consider the future consequences and the legality of their actions. Examples are crimes of passion – directed against individuals by whom the offender believes to have been wronged – such as violent acts of rage or revenge committed at the height of an emotional crisis. The analytical potential of economics of crime, which is inherently based on the assumption of *purposive* action, is limited in the case of spontaneous acts committed in the heat of the moment. But not all emotional crimes are undeliberated acts caused by failing self-control. Instead, some crimes of passion are planned well in advance. The same holds for hate crimes that are directed against people for no other reason than their belonging to certain groups (homosexuals, Jews, people of color, refugees, etc.). People in democratic countries with a high level of economic prosperity are not immune to hate crimes. For example, refugees are often met with anxieties and fears that are amplified by social media echo chambers and exploited by populist politicians. In a hostile public atmosphere, members of deviant political subcultures may feel invited to even physically attack refugees and volunteers without prior reason.

Limited attention: Some noncompliance instances are due to the fact that people make

mistakes because they are forgetful and because they can only pay attention to a limited number of issues at a time. If someone focuses strongly on one matter, other issues may receive too little or no attention despite being relevant (Simons and Chabris 1999). Not paying taxes in time or not complying with the complex legal codes of practice in food production, for example, is often not the result of purposive action but the result of negligence caused by the individual's limited attentiveness. This does not rule out an economic approach. Negligence can often be economically deterred because the rule addressees can and will use additional resources and manpower to avoid negligent mistakes if they are more tightly controlled and sanctioned or made liable for damages.

Cross-References

- Behavioral Law and Economics
- Bounded Rationality
- Crime and Punishment (Becker 1968)
- Crime: Economics of, the Standard Approach
- Criminal Sanctions and Deterrence
- Economic Analysis of Law
- Endowment Effect
- Experimental Law and Economics
- Externalities
- Harmonization: Legal Enforcement
- Hate Groups and Hate Crime
- Intrinsic and Extrinsic Motivation
- Law and Economics, History of
- Nudge
- Posner, Richard
- Prosocial Behaviors
- Protective Factors
- Public Enforcement
- Rationality
- Retributivism

References

- Ayres I, Braithwaite J (1992) Responsive regulation: transcending the deregulation debate. Oxford University Press, New York
- Beccaria C (1764) Dei delitti e delle pene. English edition: Bellamy R (ed) (1995) On crimes and punishments and

- other writings (trans: Davies R et al.). Cambridge University Press, Cambridge
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Bentham J (1789) An introduction to the principles of morals and legislation. Clarendon Press, Oxford
- Braithwaite J (2002) Rewards and regulation. *J Law Soc* 29(1):12–26
- Braithwaite V (2009) Defiance in taxation and governance: resisting and dismissing authority in a democracy. Edward Elgar, Northampton
- Dhami S (2016) The foundations of behavioral economic analysis. Oxford University Press, Oxford
- Ehrlich I (1975) The deterrent effect of capital punishment: a question of life and death. *Am Econ Rev* 65(3): 397–417
- Frey BS (1997) Not just for the money. An economic theory of personal motivation. Edward Elgar Publishing, Cheltenham
- Gigerenzer G (2010) Moral satisficing: rethinking moral behavior as bounded rationality. *Top Cogn Sci* 2(3):528–554
- Gigerenzer G, Todd PM (1999) Simple heuristics that make US smart. Oxford University Press, New York
- Gneezy U, Rustichini A (2000) A fine is a price. *J Leg Stud* 29(1):1–17
- Harcourt BE (2014) Beccaria's on crimes and punishments: a mirror on the history of the foundations of modern criminal law. In: Dubber MD (ed) *Foundational texts in modern criminal law*. Oxford University Press, Oxford, pp 39–59
- Harel A (2014) Criminal law as an efficiency-enhancing device: the contribution of Gary Becker. In: Dubber MD (ed) *Foundational texts in modern criminal law*. Oxford University Press, Oxford, pp 297–316
- Hirschauer N, Scheerer S (2014) Protective factors. In: Marciano G, Ramello GB (ed) *Encyclopedia of law and economics*. Springer, New York
- Hobbes T (1651) *Leviathan: or the matter, forme, & power of a common-wealth ecclesiasticall and civill* (ed: Shapiro 2010). Yale University Press, New Haven
- Kahneman D (2011) *Thinking, fast and slow*. Farrar, Strauss and Giroux, New York
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47(2):263–291
- Katona G (1953) Rational behavior and economic behavior. *Psychol Rev* 60(5):307–318
- Kolm SC (1973) A note on optimum tax evasion. *J Public Econ* 2:265–270
- Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. *Calif Law Rev* 88(4): 1051–1144
- Kovandzic TV, Vieraitis LM, Boots DP (2009) Does the death penalty save lives? New evidence from state panel data, 1977 to 2006. *Criminol Public Policy* 8(4):803–843
- Markowitz H (1952) The utility of wealth. *J Polit Econ* 60(2):151–158
- Miron AM, Brehm JW (2006) Reactance theory – 40 years later. *Z Sozialpsychol* 37:9–18
- Nielsen VL, Parker C (2012) Mixed motives: economic, social, and normative motivations in business compliance. *Law Policy* 34(4):428–462
- Ostrom E (2005) *Understanding institutional diversity*. Princeton University Press, Princeton
- Phelps ES, Pollak RA (1968) On second-best national saving and game-equilibrium growth. *Rev Econ Stud* 35:185–199
- Picciotto S (2002) Introduction: Reconceptualizing regulation in the era of globalization. *J Law Soc* 29(1): 1–11
- Popitz H (1968) *Über die Präventivwirkung des Nichtwissens*. Dunkelziffer, Norm und Strafe. Mohr, Tübingen
- Posner RA (2014) *Economic analysis of law*. Wolters Kluwer Law & Business, New York
- Radbruch G (2011) *Rechtsphilosophie*. In: Dreier R, Paulson SL (eds) *Studienausgabe*, 2nd edn. C.F. Müller, Heidelberg
- Rubinstein A (1991) Comments on the interpretation of game theory. *Econometrica* 59(4):909–924
- Sandel MJ (2012) *What money can't buy – the moral limits of markets*. Penguin, London
- Simon HA (1955) A behavioral model of rational choice. *Q J Econ* 69(1):99–118
- Simon HA (1957) *Models of man: social and rational*. Wiley, New York
- Simons DJ, Chabris CF (1999) Gorillas in our midst: sustained inattention blindness for dynamic events. *Perception* 28(9):1059–1074
- Thaler RH (1991) The psychology of choice and the assumptions of economics. In: Thaler RH (ed) *Quasi-rational economics*. Russell Sage Foundation, New York, pp 137–166
- Thaler RH (1999) Mental accounting matters. *J Behav Decis Mak* 12(3):183–206
- Thaler RH (2016) Behavioral economics: past, present, and future. *Am Econ Rev* 106(7):1577–1600
- Thaler RH, Sunstein CR (2008) *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven
- The Behavioural Insights Team (2011) *Behavioural insights team annual update 2010–11*. Cabinet Office, London
- Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453–458
- Von Neumann J, Morgenstern O (1944) *Theory of games and economic behavior*. Princeton University Press, Princeton
- Weinstein ND, Klein WM (1996) Unrealistic optimism: present and future. *J Soc Clin Psychol* 15(1):1–8
- Zimmerman PR (2004) State executions, deterrence, and the incidence of murder. *J Appl Econ* 7(1): 163–193

Crime: Economics of, the Standard Approach

Paolo Buonanno¹ and Juan F. Vargas²

¹Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy

²Department of Economics, Universidad del Rosario, Bogota, Colombia

Definition

Economics of crime aims at studying, theoretically and empirically, which are the determinants of criminal behavior and how it is affected by incentives and punishment. In 1968, Becker presents a paper that radically changes the way of thinking about criminal behavior. Since the beginning of 1980s, Becker's paper opens the door to a new field of empirical research whose main purpose is to verify and study the economic variables that determine criminal choices and behaviors of agents.

Introduction

In 1968 Gary Becker published "Crime and Punishment: An Economic Approach," a paper that radically changed the economic approach to analyzing criminal (as well as all types of illegal) behavior. Criminal choice conduct is not determined by mental illness or bad attitudes. Rather, it is an individual's *choice*, based on a maximization problem in which agents compare the costs and benefits of misbehavior. The costs are given by the probability of being arrested, the likely punishment, and the gain that is forgone by engaging in a crime instead of in legitimate market opportunities. The benefit is the expected return from committing the crime. Thus, in the Beckerian framework criminal decision-making responds to an economic analysis of agents.

The development of this "rational" approach to criminal behavior, in economics and other social

sciences, was favored by "redistributive" and "utilitarian" theories of crime, proposed respectively by Cesare Beccaria (Beccaria 1819) and Jeremy Bentham (1789). In "On Crimes and Punishments," Beccaria argued that crime represents a violation of the social contract, and thus punishment is justified only to defend the social contract and to ensure that everyone will abide by it, deterred from engaging in criminal activities. In turn, in "An Introduction to the Principles of Morals and Legislation," Bentham – largely influenced by Beccaria, first introduced the idea that crime is the consequence of a cost-benefit analysis. He writes: "The profit of the crime is the force which urges man to delinquency: the pain of the punishment is the force employed to restrain him from it. If the first of these forces be the greater, the crime will be committed; if the second, the crime will not be committed" (p. 33). The imprint of the two philosophers on Becker is conspicuous.

The first economics studies of crime came, however, over 200 years after the seminal writings of Beccaria and Bentham. In the 1960s, before Becker's "Crime and Punishment," Belton Fleisher explored the relationship between unemployment and youth crime (Fleisher 1963), and the role of income on the decision to engage in criminal activities (Fleisher 1966). Fleisher was the first economist to analyze empirically the relationship between crime and economic and social variables. However, it was Becker (1968)'s general theoretical framework of crime as an individual's rational choice, what constituted the starting point for analyzing criminal behavior – as well as crime control policies – from an economics perspective.

In his 1993 Economics Nobel Prize acceptance lecture, Becker underlies that while "[i]n the 1950s and '60s, intellectual discussions of crime were dominated by the opinion that criminal behavior was caused by mental illness and social oppressions, and that criminals were helpless victims," the economics approach "[i]mplic(s) that some individuals become criminals because of the financial and other rewards from crime compared to legal work, taking account of the

likelihood of apprehension and conviction, and the severity of punishment." (p. 5).

We now briefly review Becker's basic model of how individuals choose between legitimate and illegitimate activities, basing their decision on a cost-benefit analysis. In Becker's own words "[t]he approach (...) follows the economists' usual analysis of choice and assumes that a person commits an offence if the expected utility to him exceeds the utility he could get by using his time and other resources at other activities. Some persons become 'criminals', therefore, not because their basic motivation differs from that of other persons, but because their benefits and costs differ" (p. 176).

Specifically, Becker defines a supply of offences (O), which relates "the number of offences by any person to his probability of conviction (p), his punishment if convicted (f) and a portmanteau variable (u), such as the income available to him in legal and other illegal activities":

$$O_j = O_j(p_j, f_j, u_j)$$

An agent's choice is made under uncertainty, then the *expected* utility from committing a crime is defined as:

$$EU_j = p_j U_j(Y_j - f_j) + (1 - p_j) U_j(Y_j)$$

where Y_j the income ("monetary plus psychic") of committing a crime, and f_j "the monetary equivalent of the punishment." The expected utility is also determined by the probability of getting away with the crime: $1 - p_j$. The supply of crime is thus decreasing p and f .

To Becker, optimal policies to combat illegal behavior are those that minimize the "social loss" that crime entails. He defines the social loss function from offences as:

$$L = D(O) + C(p, O) + bfpO$$

where D is damage from crime, C the cost of apprehension and conviction, and $bfpO$ is the total

social loss from punishments. Here, the social policy variables are represented by p (the probability of arrest) and f (the punishment). The parameter b aggregates individual offender costs of punishment into a social cost. Solving the model with respect to the policy instruments p and f , while taking into account the equilibrium supply of offences, Becker obtains important implications on agents' propensity toward risk. Crime reduction can occur through reducing the benefits of crime, or else from raising the probability of being caught or the costs of punishment conditional upon being caught. Also "a rise in the income available in legal activities or an increase in law-abidingness due, say, to 'education' would reduce the incentive to enter illegal activities and thus reduce the number of offences" (p. 177).

In 1973, Isaac Ehrlich extended Becker's model, allowing individuals to allocate time between legal and illegal market activities, as well as in nonmarket activities. While Becker's agents are either criminals or not at all (given their cost-benefit calculation), Ehrlich models more sophisticated individuals, in the sense that they can spend time in different activities, both legal and illegal.

In spite of the similarities between Becker's basic model and Ehrlich pioneer (Ehrlich 1973), and subsequent (Ehrlich 1975, 1996) refinements, an important contrast is Ehrlich (1981)'s distinction between the "deterrence" effect of criminal sanction (considered by Becker the sole role of policing and incarceration) and the "incapacitation" effect that locking-up criminals has on crime reductions, by impeding felons to rejoin the crime industry once they are released. In Ehrlich's own words, "deterrence essentially aims at modifying the 'price of crime' for all offenders, potential and actual. (...) [I]ncapacitation, in contrast seek(s) to remove a subset of convicted offenders from the market for offences either by relocating them in legitimate labor markets, or by excluding them from the social scene for prescribed periods of time" (p. 311).

Theoretical refinements of this sort can only improve the design of policies for crime

prevention and crime reduction. After all, the primary objective of Becker himself was to "...use economic analysis to develop optimal public and private policies to combat illegal behavior" (p. 207). Today, the "economics of crime" literature pays constant tribute to the father of the field, with a growing and vibrant set of applications of the basic cost-benefit, rational approach to individual's illegal behavior.

Cross-References

- [Becker, Gary S.](#)
- [Crime and Punishment \(Becker 1968\)](#)
- [Crime, Attitude Towards](#)
- [Crime: Economics of, Different Paradigms](#)
- [Crime, Incentive to](#)
- [Criminal Sanctions and Deterrence](#)
- [Economic Analysis of Law](#)

References

- Beccaria C (1819) *An Essay on Crime and Punishments*, second American edition published by Philip H Nicklin. Translated from the Italian: "Dei delitti e delle pene" (published in 1764 by Marco Coltellini)
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Becker GS (1993) Nobel lecture: the economic way at looking at behavior. *J Polit Econ* 101(3):385–409
- Bentham J (1789) *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press, 1907 (first published in 1789)
- Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. *J Polit Econ* 81(3):521–565
- Ehrlich I (1975) On the relation between education and crime. In: Juster FT (ed) *Education, income and human behavior*. McGraw-Hill, New York, pp 313–337
- Ehrlich I (1981) On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence. *Am Econ Rev* 71(3): 307–22
- Ehrlich I (1996) Crime, punishment, and the market for offenses. *J Econ Perspect* 10(1):43–67
- Fleisher B (1963) The effect of unemployment on juvenile delinquency. *J Polit Econ* 71(6):543–555
- Fleisher B (1966) The effects of income on delinquency. *Am Econ Rev* 56(1/2):118–137

Crime: Expressive Crime and the Law

Martin A. Leroch

Politics and Economy, Johannes Gutenberg
Universität Mainz, Mainz, Germany

Abstract

Expressive crime contrasts with instrumental crime in that delinquents aim to "make a statement", and not to "make a living". This difference in motivation has important consequences for the deterrent effect of policies. Whereas policies aiming at expected material costs of perpetrators have often proven successful in deterring instrumental crime, they may fail to deter delinquents with expressive motivations. In some cases, perpetrators may even feel defied to increase their activity.

Definition

Expressive crime is a form of illegal behavior which is motivated by the desire to communicate personal attitudes to others. It contrasts with instrumental crime, which is motivated by the desire to gain material objects.

Crime (Expressive) and the Law

Crime may be distinguished along several dimensions. One such dimension focuses on the motivation underlying illicit actions. Accordingly, while instrumental crime is motivated by the desire to "make a living," expressive crime is motivated by the will to "make a statement." Examples of instrumental crime include robbery, stealing, and fraud, while suicide terrorism, illegal political protests, or the spraying of graffiti constitute examples of expressive crime. Some incidents of crime share both components. For instance, severely beating up a debtor who owes money to a criminal organization may on the one hand be regarded a signal or statement to others

that they better repay their debt. On the other hand, this message per se is instrumental in acquiring more riches.

Distinguishing crime according to its motivation is important for the design of policies to deter criminals, one of the major foci of economists, as has been argued, for instance, by Cameron (1988), Kirchgässner (2011), or Robinson and Darley (2004). In theory, deterrence can be achieved either by increasing the probability of detection or by reducing the utility levels of detected delinquents. For the case of instrumental crime, this theory appears valid, at least in parts. While there is evidence that policies aiming at increasing detection probabilities may be successful (e.g., in Klick and Tabarrok 2005), most studies call into question the deterrent effect of increases in available punishment. In cases of expressive crime, however, an increase in detection probabilities or punishment may even lead to perverse effects of increases in both number and intensity of illicit behaviors. The reason for these effects is that punishment may promote crime by affecting “the values of certain other variables ... which in turn have a direct effect on deviant behaviour” (Opp 1989, p. 421). Crucial for this crime-promoting effect are indirect effects of punishment, which change the (social) incentive structures for perpetrators. Opp finds support for the defiance of perpetrators in examples of legal and illegal political protest, two prominent forms of expressive behavior.

What makes expressive crime so susceptible for these indirect effects of punishment is its signaling property. By making their illegal statements, perpetrators want to send signals to others, for instance, their peers or their opponents. For instance, perpetrators may wish to signal their opponents that they are dissatisfied with the current organization of society or with certain decisions of the current government. Likewise, they signal their active peers that they do not “stand alone” with their attitude. Because the costs associated with transmitting the signal affect its “quality” positively, delinquents may choose to send more (and/or stronger) signals when harsher means of deterrence are employed. Deliberately facing armed forces on the street, for instance,

may transmit the signal that I am willing to incur physical pain in order to show others that I disagree with certain policies of the current government. The more likely physical pain is, or the more pain I am to expect, the better I can signal my disagreement. The willingness to incur this pain may also increase my standing among peers – assuming that all peers share opposition toward the government and agree that illegal protest is an adequate form of expressing this opposition.

If standard means of deterrence prove counter-effective, the question of how to counter expressive crime evolves. Addressing the signal value of illicit behaviors or the signal as such appears the most promising strategies. The fight against graffiti in New York City’s underground provides a vivid example for such a strategy. After being considered a serious problem in the 1970s, the city managed to almost entirely free its subways from graffiti by 1989. Apparently, the crucial change in strategy was to prohibit sprayed trains from running. Any train leaving the depot was cleaned from graffiti entirely before being brought to service. Sprayers thus knew that spraying was no effective way of transmitting their signals any longer (sprayers typically want to signal disagreement with capitalist ways of organizing society, express artistic desires, or gain attention from others. See Leroch (2014) for a more detailed analysis of graffiti spraying), and they stopped spraying on subways. Cities around the globe copied this strategy successfully in the fight against graffiti.

References

- Cameron S (1988) The economics of crime deterrence: a survey of theory and evidence. *Kyklos* 41: 301–323
- Kirchgässner G (2011) Econometric estimates of deterrence of the death penalty: facts or ideology? *Kyklos* 64(3):448–478
- Klick J, Tabarrok A (2005) Using terror alert levels to estimate the effect of police on crime. *J Law Econ* 48(2):267–280
- Leroch M (2014) Punishment as defiance: deterrence and perverse effects in the case of expressive crime. *CESifo Econ Stud.* <https://doi.org/10.1093/cesifo/ift009>

- Opp K-D (1989) The economics of crime and the sociology of deviant behaviour: a theoretical confrontation of basic propositions. *Kyklos* 42(3):405–430
- Robinson PH, Darley JM (2004) Does criminal law deter? A behavioural science investigation. *Oxf J Leg Stud* 24(2):173–205

Crime: Organized Crime and the Law

Matthew J. Baker

Hunter College and the Graduate Center, CUNY,
New York, NY, USA

Abstract

This entry reviews the literature on the economics of organized crime. While the economics of organized crime is a small subfield of economics, it can offer insights in unexpected areas of economics. The field began some 40 years ago with the study of organized crime and its participation in illicit activities such as prostitution and gambling. Over time, the economic analysis of organized crime has expanded to address broader questions of governance in the absence of formal institutions.

Introduction and Background

The economic literature on organized crime constitutes a rather small portion of the field of law and economics. That having been said, it offers some interesting and unexpected insights into questions of interest to economists. Since its beginnings 40 or so years ago, economic research on organized crime has expanded from a somewhat narrow focus on illicit markets into a broad study of patterns in organizational formation and collective decision-making. The result is that economists' study of organized crime has contributed to the understanding of things of fundamental importance, such as the economics of governance and the formation and organization of property rights.

The phenomenon of organized crime first attracted the interest of economists in the late

1960s and early 1970s. The interest can be attributed to enthusiasm for Becker's (1968) development of an economic theory of criminal behavior and decision-making and also to policy interests deriving from the 1967 report of the President's Racketeer Influenced and Corrupt Organizations (RICO) task force and ensuing adoption of the Organized Crime Control Act of 1970. Initial work focused mainly on description of industries often associated with organized crime, such as loan-sharking, gambling, prostitution, and narcotics, to name a few (see Kaplan and Kessler (1976) for several illuminating examples). Schelling (1967, 1971) provided the first theoretical apparatus for thinking about organized crime. From his initial theoretical steps, the definition and study of organized crime has evolved toward the view that the criminal organization is in fact a form of nascent government, which often arises to perform the usual government functions of enforcement and policing in sectors of the economy where the government is either unable or unwilling to do so.

Defining and Theorizing About Organized Crime

Activity-Based Analysis

Defining organized crime is a difficult task, and, as noted by Fiorentini and Peltzman (1995b, Chapter one) in their review of the literature, research on organized crime has to some degree been driven for a search for an adequate definition of the subject (incidentally, the publication of Fiorentini and Peltzman (1995a) was something of a watershed moment in the economics of organized crime. It is required reading for anyone interested in the economics of organized crime, as is Fiorentini (2000). In fact, this entry has been written with the twin objective of covering necessary ground yet complementing these earlier works in mind). Accordingly, their characterization serves as a useful starting point for characterizing the economic literature on organized crime (defining organized crime is also a difficult multidisciplinary puzzle. See Finckenauer (2005)). As alluded to in the "Introduction" and as Fiorentini

and Peltzman (1995b, Chapter One) emphasize, the first attempts at the definition of organized crime were activity based. A criminal organization was deemed to be one that operates either in full or in part in illegal or illicit markets such as gambling, loan-sharking, and prostitution. Analysis focused on description of the growth and evolution of organized crime in particular industries, operational details, and estimation of the overall size of the industry. One might include in this category even earlier descriptions of the growth and expansion of organized crime, such as analyses of the rise of the American Mafia during Prohibition or analyses of the growth and organization of the Sicilian Mafia (see, e.g., Cressey (1969), who provides a thorough description of the history and nature of organized crime in the United States). One illustrative example of a lucid industry description is given by Kaplan and Matteis (1976) who discuss the day-to-day workings of a loan-sharking operation. Much of this early literature emphasizes that the criminal organization tends to provide goods and services which are illegal, but for which there is nonetheless consumer demand, such as narcotics or prostitution (one might maintain that the development and growth of the Mafia has less to do with serving illegal markets and more to do with filling a vacuum of power. In this way, early literature on the Mafia might have more in common with the theories based upon the development of property rights under anarchic circumstances, which are discussed in section “[Governance and Crime](#)”). Kaplan and Matteis (1976), to continue the example, note that loan sharks are specialists in offering credit to an unserved segment of the market: high-risk, short-term borrowers. Of course, description of industry is essential to its study, and this tradition is, thankfully, still very much alive in the economic literature (see, e.g., Bouchard and Wilkins (2010)).

Extortion and Crime

Schelling (1967, 1971) represents the initial break with the activity-based approach to the study of organized crime and also is the first to present a cohesive theoretical structure for thinking about the economics of criminal organizations. His

point of departure is that a criminal organization is in essence an extortative body. As such, its primary goal is to maintain monopoly control of illicit activities (and maybe even legitimate ones) for the purposes of rent extraction – protection money. There is no real reason, according to Schelling, that prostitution or gambling or other illicit activities have to be monopolized, but they are obvious targets for extortion because they need some degree of visibility to operate and benefit from regular customer bases, but could not operate as effectively without some shield from the law. Moreover, industries like prostitution and gambling are easily monitored for purposes of rent extraction, as things like income and provider ability are easily monitored.

Schelling’s view is attractive in that it has some ability to explain other aspects of organized crime, such as why organized criminal organizations so often engage in turf wars. Buchanan (1973) parlayed Schelling’s position into a supplementary theoretical argument offering some good news for society: since illicit activities are by and large those that society deems unacceptable (social “bads”), monopolization of these activities by criminal organizations actually produces a social benefit, as these goods will tend to be underprovided. Backhaus (1979), however, takes exception with this position, noting that, among other things, there are likely to be scale economies in organizing illicit activities. Thus, a monopolist might provide more illicit goods and services to the market.

The first big challenge to the theory evinced by Schelling came from Reuter (1983), who analyzed a broad array of empirical and historical evidence on organized crime and found that Schelling’s approach did not always match the facts. In particular, Reuter (1983) emphasized that much organized criminal activity seems to be more competitive than monopolistic (or perhaps at least monopolistically competitive), in that many small, independent entities often provide illicit goods or services in the same market area. Some years later, technical aspects of this observation were taken up by Fiorentini (1995), who developed a former mathematical model of oligopolistic competition in illegal markets. He finds that it

is difficult to draw exact conclusions about how resources invested in violence, corruption, or provision of the goods depend upon the degree of competition and that the answer depends in a critical way on what sort of response criminal organizations expect from the government. This result in fact foreshadows results emanating from recent research on organized crime, governance structures, and imperfect enforcement of property rights. A model in a similar vein is presented by Mansour et al. (2006). They ask how it could be the case that, while deterrence expenditures increased throughout the 1980s and 1990s in the United States, criminal output (e.g., output of drugs) increased and prices fell. The essential idea derives from allowing market structure to respond to the pressures of deterrence; if deterrence creates more competition, output may well expand.

An alternative tradition that also departs from the theory proposed by Schelling, owing originally to Anderson (1979) (see also Anderson (1995)), is a transaction cost approach to studying organized crime. Anderson, in her description of the development and organization of the Sicilian Mafia, argues that transaction costs are important drivers of organized criminal activity. Dick (1995) provides a more expansive transaction cost analysis, which emphasizes what even casual observation suggests are universally important practical problems presented by illicit exchange, such as commitment, credibility, and maintenance of contracting arrangements and long-term relationships. Polo (1995) presents a formal theoretical model capturing credibility and commitment in a long-term setting that bears resemblance to models of contracting and commitment in settings in which agents cannot rely on formal institutions, as in Clay (1995) and Greif (1993). Turvani (1997) also attacks the problem of organized crime from a transaction cost economics perspective.

Governance and Crime

In the early 1990s, an alternative way of thinking about the organized criminal organization as a kind of nascent state emerged. This theoretical approach grew out of the literature on endogenous

development and enforcement of property rights in conditions of anarchy, usually associated with the initial efforts of Skaperdas (1992), Hirshleifer (1988), and Grossman and Kim (1995). One essential idea of this work is that security is, at the end of the day, a good like any other, and while the government typically provides it, when it does not, some other actor will invest in the provision of security. Milhaupt and West (2000) provide some data-based empirical support for the idea that illicit activity expands due to inefficiencies in state provision of property rights. Skaperdas and Syropoulos (1995) present a model in which rival gangs allocate resources to both production and security of what is produced. Grossman (1995) models a situation in which a criminal organization arises as a “rival kleptocrat” to the government; that is, the Mafia comes about as a rivalrous organization defending property rights and extracting resources from producers, just as the government imposes taxes. One interesting aspect of Grossman’s argument is that the populace might actually benefit from the competition in services provided by the criminal organization. It both expands the amount of protective services available and also limits rent extraction by the licit government.

This view of the criminal organization was subsequently further developed and expanded upon. Skaperdas (2001) provides an expansive view of historical evidence through this lens. At the same time, more sophisticated views of markets and interactions have been applied to the study of organized crime as well as the basic idea that the criminal organization operates in conjunction with the government and is to some degree fueled by the inadequacy of bureaucrats and provision of basic government services. For example, Kugler et al. (2005) present a model in which criminal organizations compete in a “global” market in providing products, but intercede only locally in bribery of bureaucrats. They find increased deterrence could, in some circumstances, lead to higher crime.

Indeed, this branch of literature has evolved toward the position that the governance and security problems which criminal organizations arise to deal with are not all that different from the

problems of social organization faced by human populations throughout history. Skaperdas and Konrad (2012) provide a model in which self-governance is possible and indeed a best-case scenario for the populace. It is, however, less stable and susceptible to more predatory styles of government, and unfortunately, the more competition in the market for protection, the worse things become. Similarly, Baker and Bulte (2010) focus on the Viking epoch beginning around 800 AD and describe how Viking raids – which might be thought of as organized criminal activity – became more and more sophisticated, even as measures to prevent raiding grew more sophisticated.

Internal Structure, Participation, and Deterrence Policy

The basic facts underlying the internal structure of criminal organizations are ingrained in popular myth, yet deeper details and analysis have been a bit harder to come by (of course, much of the work cited in this entry contain institutional detail. See, e.g., Skaperdas (2001) and Anderson (1979)). There is a large empirical and theoretical literature on the decision to participate in crime dating from Becker's (1968) paper, but the decision to participate specifically in organized crime has received considerably less attention.

The notable exception is Levitt and Venkatesh (2000), who analyze in detail the finances and internal structure of a Chicago drug-selling gang. The gang analyzed by Levitt and Venkatesh has a hierarchical structure with a highly skewed wage scale. Those at the bottom of the hierarchy are paid minimal fixed wages for what is dangerous and risky work. Those at the top of the hierarchy do considerably better, motivating Levitt and Venkatesh to suggest promotion and wages are to some degree governed by a rank-order tournament. Still, they find it is difficult to reconcile gang participation with optimizing behavior if one only considers pecuniary rewards from participation. A further finding of Levitt and Venkatesh is that those in the lower echelons of the gang often simultaneously participate in

legitimate markets; this suggests that members of the lower rungs might be peeled off by targeted policy interventions.

Garoupa (2000, 2007) explicitly focuses on the internal, vertical structure of the criminal organization using principal-agent theory. Garoupa (2000) follows a more or less classic model of enforcement and deterrence, but allows that offenders must pay a fee to a monopolist agent (i.e., the Mafia) to commit a crime. A key result is that less severe enforcement policy may be optimal than it would be in the absence of a Mafia. Garoupa (2007) works in the same framework, but focuses on a wider assortment of potential punishments and different dimensions of the organization, such as how many agents the organization will seek to hire and how mistake-prone the organization will be in light of its size and, ultimately, the sort of punishments it faces. One result of interest is that severe punishment may reduce organizational scope, but increase organizational effectiveness.

Conclusions

While the economic literature on organized crime remains small, within this literature, a considerable measure of historical detail and methodological diversity has been used in its study. Perhaps the most striking thing about the literature on organized crime is that the very different methodologies employed in its study all recommend a degree of care in thinking about its prevention and deterrence. The recurring theme is that heavier investment in policing and punishing organized criminal activity may have unexpected and unpleasant consequences. The sophistication, competence, and extent of the organization may respond in ways contrary to intuition. In the current domestic and international atmosphere, illicit market activity in the form of international drug and human trafficking remains a major concern, as do the terrorist and other non-state organizations which arise whenever and wherever in the world there is a vacuum of power. In light of this state of affairs, the key insights offered by the literature on organized crime seem to be well worth emphasizing.

References

- Anderson AG (1979) The business of organized crime: a Cosa Nostra family. The Hoover Institution, Stanford
- Anderson AG (1995) Organized crime, mafia, and governments. In: Fiorentini G, Peltzman S (eds) *The economics of organized crime*. Cambridge University Press, Cambridge, UK, pp 33–54
- Backhaus J (1979) Defending organized crime? A note. *J Leg Stud* 8:623–631
- Baker MJ, Bulte EH (2010) Kings and Vikings: on the dynamics of competitive agglomeration. *Econ Gov* 11:207–227
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Bouchard M, Wilkins C (eds) (2010) *Illegal markets and the economics of organized crime*. Routledge, London/ New York
- Buchanan J (1973) A defense of organized crime. In: Rottenberg S (ed) *Economics of crime and punishment*. American Enterprise Institute, Washington, DC, pp 119–132
- Clay KB (1995) Trade without law: private-order institutions in Mexican California. *J Law Econ Org* 13:202–231
- Cressey DR (1969) *Theft of the nation: the structure and operations of organized crime in America*. Harper and Row, New York
- Dick AR (1995) When does organized crime pay? A transactions cost analysis. *Int Rev Law Econ* 15:25–45
- Finckenauer JO (2005) Problems of definition: what is organized crime? *Trends Organ Crime* 8:63–83
- Fiorentini G (1995) Oligopolistic competition in illegal markets. In: Fiorentini G, Peltzman S (eds) *The economics of organized crime*. Cambridge University Press, Cambridge, UK, pp 274–288
- Fiorentini G (2000) The economics of organized crime. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics, volume V: the economics of crime and litigation*. Edward Elgar, Cheltenham, pp 434–459
- Fiorentini G, Peltzman S (eds) (1995a) *The economics of organized crime*. Cambridge University Press, Cambridge, UK
- Fiorentini G, Peltzman S (1995b) Introduction. In: Fiorentini G, Peltzman S (eds) *The economics of organized crime*. Cambridge University Press, Cambridge, UK, pp 1–30
- Garoupa N (2000) The economics of organized crime and optimal law enforcement. *Econ Inq* 38:278–288
- Garoupa N (2007) Optimal law enforcement and criminal organization. *J Econ Behav Organ* 63:461–474
- Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. *Am Econ Rev* 83:524–548
- Grossman HI (1995) Rival kleptocrats: the mafia versus the state. In: Fiorentini G, Peltzman S (eds) *The economics of organized crime*. Cambridge University Press, Cambridge UK, pp 143–155
- Grossman HI, Kim M (1995) Swords or plowshares? A theory of the security of claims to property. *J Polit Econ* 103:1275–1288
- Hirshleifer J (1988) The analytics of continuing conflict. *Synthese* 76:201–233
- Kaplan LJ, Kessler D (1976) *An economic analysis of crime: selected readings*. Charles C. Thomas, Springfield
- Kaplan LJ, Matteis S (1976) The economics of loansharking. In: Kaplan LJ, Kessler D (eds) *An economic analysis of crime: selected readings*. Charles C. Thomas, Springfield, pp 178–192
- Kugler M, Verdier T, Zenou Y (2005) Organized crime, corruption, and punishment. *J Public Econ* 89: 1639–1663
- Levitt SD, Venkatesh SA (2000) An economic analysis of a drug-selling gang's finances. *Q J Econ* 115:755–89
- Mansour A, Marceau N, Mongrain S (2006) Gangs and crime deterrence. *J Law Econ Organ* 22:315–339
- Milhaupt CJ, West MD (2000) The dark side of private ordering: an institutional and empirical analysis of organized crime. *Univ Chic Law Rev* 67:41–98
- Polo M (1995) Internal cohesion and competition among criminal organizations. In: Fiorentini G, Peltzman S (eds) *The economics of organized crime*. Cambridge University Press, Cambridge, UK, pp 87–103
- Reuter P (1983) *Disorganized crime: the economics of the visible hand*. MIT Press, Cambridge, MA
- Schelling TC (1967) Economics and the criminal enterprise. *Public Interest* 7:61–78
- Schelling TC (1971) What is the business of organized crime? *Am Sch* 40:643–652
- Skaperdas S (1992) Cooperation, conflict, and power in the absence of property rights. *Am Econ Rev* 82: 720–739
- Skaperdas S (2001) The political economy of organized crime: providing protection when the state does not. *Econ Gov* 2:173–202
- Skaperdas S, Konrad K (2012) The market for protection and the origin of the state. *Econ Theory* 50: 417–443
- Skaperdas S, Syropoulos C (1995) Gangs and primitive states. In: Fiorentini G, Peltzman S (eds) *The economics of organized crime*. Cambridge University Press, Cambridge, UK, pp 61–81
- Task Force Report: Organized Crime (1967) *The President's commission on law enforcement and administration of justice*. U. S. Government Printing Office, Washington, DC
- Turvani M (1997) Illegal markets and the new institutional economics. In: Menard C (ed) *Transactions cost economics*. Edward Elgar, Cheltenham

Crimes Against Nature

► Sex Offenses

Criminal Constitutions

Nicholas A. Snow

Department of Economics, Indiana University,
Bloomington, IN, USA

Definition

Criminal constitutions exist in a number of different criminal enterprises, from the Golden Age of Pirates to modern prison and street gangs. The purpose of these constitutions is to better facilitate cooperation among organizational members in order to help better achieve profit maximization.

Criminal Constitutions

It is easy to imagine criminal organizations being composed of lawless, all out for their own chaotic individuals but this is far from the truth. From an economic perspective, especially since Gary Becker's famous 1968 publication, *Crime and Punishment: An Economic Approach*, the starting point for analyzing criminal behavior is to assume that criminals are rational agents. From this perspective, it is not surprising to find that many black markets and other criminal organizations are extremely orderly and rational. This is, of course, not to say that they are desirable or morally good but if we are to truly understand criminal organization it is important to analyze them in truth and not in fiction.

Many criminal organizations are not only rational and orderly but also formally created as such. For example in his 1991 book, *Islands in the Streets*, Jankowski found that 22 out of 37 street gangs had written criminal constitutions. And street gangs are only one of numerous examples of criminal organizations having deliberately written and constructed constitutions that attempt to align the self-interests of its members to that of the group interest, such as eighteenth century pirates, prison gangs, the Mafia, etc. Additionally, other black markets without formal constitutions

will often still contain informal and implicit criminal codes that exist to do the same thing, such as the market for smuggled liquor in Detroit in the United States in the 1920s under alcohol prohibition as illustrated by historian Larry Engelmann's 1979 book, *Intemperance, the Lost War Against Liquor*.

The reason for the emergence of such criminal constitutions is simple. Black market firms are still firms and as such act in order to maximize profits. Economists Leeson and Skarbek, in a 2010 article, *Criminal Constitutions*, highlight three main reasons for the emergence of criminal constitutions. The first is that constitutions help to create common knowledge about what the organization expects from its members and the different members can expect from each other. This helps to align expectations among participants within the organization and in the broader market setting. Common knowledge is often created by having explicit rules, whether written or not, of how to behave within the organization and without. This has the advantage of members knowing exactly what they are getting themselves into by joining the organization in terms of duties, rewards, obligations to other members, etc. Rules may also consist of entrance requirements for potential members. By knowing the rules when you come into the organization, current members can have more confidence in the new members.

Second, criminal constitutions help to solve the principle agent problem within criminal organizations. It is often the case that members of the various criminal organizations are not necessarily the residual claimants of the organization, thus making them particularly vulnerable to divergent interests. The rules within the constitutions, and their enforcement, can help to make sure that divergence of interests does not take place. For criminal organizations this may be extremely important. For example, as Leeson explains in his 2009 book, *The Invisible Hook*, a pirate's profitable enterprise consists of plundering on the high seas. This is often hazardous work. In the act of successfully plundering a merchant vessel, pirates might be required to

engage in a dangerous battle. Individual crew members may recognize the high costs of participating in battle to themselves. Thus, a collective action problem emerges because the individual pirate's incentive is to stay out of the fight but the more pirates that think like this, the less likely the ship is going to be taken. Profit maximization requires pirates to take as many ships as possible. Therefore, pirates created rules and rewards to encourage all individuals to engage in battle when necessary, such as providing, what essential boiled down to, workman's compensation for injury and special rewards for particular courage.

Finally, criminal constitutions create information about member misconduct and help formulate rules and enforcement that prohibit such behavior. This final function is self-enforcing because conflict within any organization will necessarily cut into the profits of the illicit firm. If members of the gang are constantly fighting, then they are not engaged in productive efforts, which earn the organization profits. As Skarbek illustrates, in his 2014 book, *The Social Order of the Underworld*, prison gangs even utilize rules of conflict for preventing intra-gang conflicts. Violence is costly for business, so the gangs have an incentive to enforce rules among their own members to avoid unnecessary conflict even with competing gangs.

The main distinction facing illicit organizations from legal organizations is the lack of access to formal governance institutions necessary for markets to properly function and flourish. Given the illegal nature of their activities, criminal organizations are unable to turn to legal and social institutions, which help to enforce private property rights, contracts, and facilitate collective action agreements. It is for this reason that most black markets consist of many, very small, and ephemeral enterprises. Thus, contrary to conventional wisdom, large monopolies are often a rare occurrence within black markets. While the use of violence is often unlikely to have an effect on this structure, it is, however, likely to have an effect on profit opportunities for the different organizations within the market. Constitutional arrangements

are then able to help facilitate cooperation within and between organizations.

Ultimately, as profit maximizing organizations, criminal organizations use constitutions in order to help facilitate cooperation among its various members to help ensure higher profits for the organization. This certainly seems strange considering these organizations are made up of individuals who are openly flaunting the governments laws but when understanding that just like a legitimate firm, criminal organizations wish to maximize their profits, it becomes clear why such rules would emerge. Markets need rules to function and because of their illicit nature the government is not an option for providing such rules, thus the organizations will create these rules in order to facilitate the gains from trade.

Cross-References

- [Anarchy](#)
- [Crime and Punishment \(Becker 1968\)](#)
- [Prohibition](#)

References

- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Engelmann L (1979) Intemperance: the lost war against liquor. Free Press, New York
- Jankowski MS (1991) Islands in the street: gangs and American Urban Society. University of California Press, Berkeley
- Leeson PT (2009) The invisible hook: the hidden economics of pirates. Princeton University Press, Princeton
- Leeson PT, Skarbek D (2010) Criminal constitutions. *Glob Crime* 11(3):279–298
- Skarbek D (2014) The social order of the underworld: how prison gangs govern the American penal system. Oxford University Press, Oxford

Further Reading

- McCarthy DMP (2011) An economic history of organized crime: a national and transnational approach. Routledge, New York
- Reuter P (1985) The organization of illegal markets: an economic analysis. U.S. Department of Justice: National Institute of Justice, Washington, DC

Criminal Sanctions and Deterrence

J. J. Prescott

Law School, University of Michigan, Ann Arbor, MI, USA

Abstract

This entry defines criminal sanctions by distinguishing them from civil sanctions, and briefly surveys the major categories of criminal sanctions, both ancient and new. The entry then outlines the primary social justifications for using such sanctions—focusing on deterrence as a distinct purpose of punishment—and describes a basic model of criminal offending to highlight the conditions under which the threat of criminal sanctions can influence offender behavior in predictable ways. Next explored are the implications for deterrence of the different types of criminal sanctions. A brief discussion of the suggestive conclusions emerging from related empirical evidence follows.

Definition

Criminal sanctions are punishments (e.g., imprisonment, fines, infliction of pain, or death) imposed by governments on individuals or corporate entities for the violation of criminal laws or regulations. Deterrence, one of the principal theories used to justify the use of criminal punishment, occurs when the possibility that an individual will be subjected to a criminal sanction were he to commit a crime causes the individual to forgo or engage in less of the behavior in question. Deterrence is the primary justification for punishment in the field of law and economics.

Introduction

To address socially harmful (or potentially harmful) behavior, governments enact laws (or groups

of people develop practices) to sanction or punish individuals who choose to engage in it. If the problematic behavior rises to a particular level of harmfulness or has particular features (e.g., physical violence), the behavior may be classified as criminal, and upon conviction, a violator will, by definition, receive some sort of criminal sanction. The distinction between a criminal sanction and a civil sanction is largely one of degree and possibly just semantics, although many scholars have attempted to carefully delineate the two spheres. There is a credible argument that a crime is simply a harmful act that the government decides to label a crime (perhaps because it is seriously harmful but perhaps not), and a criminal sanction is simply a sanction the government applies to someone who has performed such an act.

This entry will begin by briefly exploring what may make a sanction “criminal” in nature and then by summarizing the key categories of criminal sanctions. Some types of criminal penalties are largely historical relics; others are innovative and rely on modern technology. Nevertheless, sanctions of more recent vintage, such as constant, real-time GPS monitoring, share many features with more traditional varieties. This entry will primarily focus, however, on deterrence—specifically, its definition and how the imposition of criminal sanctions can achieve this fundamental purpose of punishment. Basic features of the economic model of crime will be described (see Becker 1968), and a variety of topics relevant to criminal deterrence will be covered, including the model’s theoretical predictions for offender behavior, the implications for deterrence of employing different categories of modern criminal sanctions (including their unintended consequences and the potential for sanction complementarity when they are used simultaneously), and the consensus conclusions of the empirical literature.

Criminal Versus Civil Sanctions

What makes a sanction a “criminal” sanction? The cleanest answer to this question is the tautological response that “a criminal sanction is a sanction that seeks to punish someone for having committed a

crime,” thus begging the question: “What makes a particular act a crime?” One important attribute of a crime is that the act or its consequences be socially harmful. But the category of “harmful” behaviors is ultimately socially constructed through some aggregation of individual preferences and beliefs and is therefore determined simply by whether a government decides that the act in question is indeed harmful enough to be worthy of public censure. Some harmful acts are not subject to government sanctions for practical or philosophical reasons. When a government decides to discipline someone who has engaged in a harmful act, it selects between or some combination of “civil” measures (traditionally, fines) and criminal sanctions.

It is impossible to identify conditions that, without exception, distinguish a crime from another form of civil wrong, like a tort. But there are tendencies, or characteristics, that make a certain harmful act more likely to be labeled a crime. One important criterion is the mental state of the wrongdoer. Crimes are usually (but by no means always) done knowingly or intentionally or at least involve a blameworthy awareness of risk; otherwise innocent acts that accidentally result in harm, however, are rarely considered crimes, and sometimes even intentional acts that cause harm are merely intentional torts (Miceli 2009, p. 269).

Other suggested criteria relate to the difficulties that harmed individuals (or their friends or family) might face (as a group) in privately enforcing particular prohibitions. For example, when particular categories of wrongdoers might be difficult to identify (e.g., thieves), using specialists may make sense, and the average victim is certainly no specialist (Polinsky and Shavell 2007, p. 406). Likewise, if the nature of the harm is widespread, enforcement may be a public good. No single private individual may have the incentive to pursue the wrongdoer. As Miceli (2009, p. 270) points out, extreme examples of this category are “victimless” crimes (e.g., possessing child pornography or drugs) in which the harms are often thought to be the knock-on, indirect consequences of the activity.

Some argue that the civil/criminal wrong divide maps to the liability/property rule

distinction in the law (Miceli 2009, p. 273). Under this rubric (see Calabresi and Melamed 1972), civil wrongs are those for which there is, ultimately, a price, usually set by the court *ex post* in the form of damages awarded in a civil trial. By contrast, the enactment of a criminal prohibition can be viewed as the granting of an inalienable right to potential victims by the government. There is no dollar figure that an offender can pay to rectify the violation of this right. Rather, the offender must be punished, and the punishment has no necessary relationship to (and is usually much greater than) any damages sustained by the victim.

Even better, however, would be the prevention of the harm *ex ante*, since it is in some sense irreversible. This is consistent with criminal law often being publicly enforced by specialists (Shavell 2004, p. 575), as *ex ante* prevention by private individuals seems much less likely to succeed relative to winning an *ex post* lawsuit.

Types of Criminal Sanctions

Typical criminal sanctions in modern countries include fines, incarceration, and supervision (including probation and parole). The death penalty (or capital punishment) is a historically important criminal sanction, but it is employed rarely in practice in Western countries these days and is treated elsewhere in this collection.

Fines. When a fine is imposed, an offender is legally ordered to pay to the state or perhaps to a third party the levied amount, perhaps in lump sum or perhaps in installments under a payment plan. Fines are limited in their effectiveness as a punishment by the assets or potential assets of the offender (and by extension, the offender’s ability to work or otherwise access resources) and by the capacity of the government to collect the fine, through garnishment of wages, for example.

Incarceration. Imprisonment or incarceration more generally is perhaps the best-known and most common form of criminal sanction in the modern world, at least with respect to serious crimes. The “quantity” of imprisonment varies by the conditions of the imprisonment (e.g., amount

of space, cleanliness, quantity or quality of food, available or required activities, access to other prisoners, access to loved ones, and so on) and by the time the offender is sentenced to remain imprisoned.

Intermediate Sanctions. Intermediate or alternative sanctions are typically viewed as less serious than incarceration: the offender is released from (parole) or never sentenced to (probation) incarceration but remains free only by compliance with specified conditions, which can be affirmative obligations or negative restrictions.

Traditional forms of intermediate sanctions are increasingly common, at least in some jurisdictions—e.g., home detention, mandated community service, and mandatory drug treatment. Other forms of punishments are considerably less common. For example, rare at least in most Western countries is the imposition of the death penalty, the killing of an individual by the government as punishment for the commission of a crime. Corporal punishment, the deliberate infliction of pain as punishment for a criminal offense (e.g., caning or whipping), is still practiced, although today only in a small minority of countries and with significant limitations.

Governments, including many in the USA, also employ shaming as a criminal sanction—i.e., subjecting an offender to public humiliation as a form of punishment. Historically, these punishments were often accompanied by an element of physical discomfort or of physical or verbal abuse (e.g., stocks or pillory). Modern shaming punishments retain the humiliation element and perhaps allow verbal abuse by the public, but do not typically impose conditions that might result in extraordinary physical discomfort or pain. Finally, banishment, as a form of punishment, is the mirror image of incarceration. Rather than an offender being required to be in a particular place, he is required to leave the community altogether in order to isolate that individual from friends and loved ones and to eliminate any future threat to the community.

Many of the most innovative criminal sanctions in recent years have emerged in response to public concern over sex offenses, especially in the USA. In the 1990s, sex offenders were

believed to be highly likely to return to sex crime upon release from incarceration, and many such offenses are serious crimes; as a result, governments developed new criminal sanctions that manage to combine elements of incapacitation, shaming, and banishment.

Sex offender registration, for example, is an extremely weak form of incapacitation. When required to register, offenders must provide identifying information (e.g., name, date of birth, address, criminal history) to law enforcement to facilitate the latter's monitoring and, if necessary, apprehension of them. Consequently, committing a new crime becomes more difficult for an offender who in some sense weighs the costs and benefits of his options. Mandatory real-time GPS monitoring of an offender's location is a more technologically advanced version of registration. If monitoring equipment is visible to others when it is worn, shaming also occurs, much like the proverbial wearing of a scarlet letter.

Community notification is a stronger form of incapacitation, coupled with a different form of shaming. When an offender is made subject to notification, the government releases identifying registry information to the public, which increases the overall level of monitoring, prompts potential victims to take precautions (in theory creating a bubble around the offender), and subjects the offender to humiliation, to great difficulty in finding employment and housing, and potentially to more active forms of abuse at the hands of members of the public. Yet while GPS monitoring equipment may result in humiliation in every one-on-one interaction with another person if the equipment is visible, an individual living under notification, even when the notification is an active form (e.g., neighbors of an offender receive a card in the mail containing the offender's registration information), is not physically marked as a sex offender and so may have anonymous interactions without humiliation.

Sex offender residency restrictions are a targeted form of banishment. Offenders living under residency restrictions are not necessarily forbidden from a particular place under all conditions. However, offenders cannot lawfully live within a certain distance of places where potential victims are particularly likely to frequent. Other types of

restrictions—travel restrictions, employment restrictions, and so on—function in the same way, by banishing offenders either from places where or from roles in which they are assumed to pose the most threat.

Under many criminal justice systems, offenders are sentenced to combinations or bundles of different kinds of criminal sanctions. For example, sentences that include both fines and incarceration are common, as are sentences that begin with incarceration but are followed by parole or some other form of supervision. These combination sanctions may be attempts to pursue multiple purposes of punishment simultaneously (e.g., a fine can punish, but it cannot incapacitate, at least not to the extent that a prison cell can) or they may be an attempt to punish more efficiently. For example, as we will see below, fines can be more efficient as a form of punishment because they are a transfer, but most offenders are liquidity constrained and so fines have only limited utility in most circumstances. Finally, there may be economies of scope in using a cluster of criminal sanctions as a consequence of behavioral tendencies, including, e.g., hyperbolic discounting. Increasing a sentence from 5 to 6 years may be less effective at altering behavior than adding a comparable fine or other restrictions to a 5-year sentence.

Finally, while not formally sanctions, arrest and criminal process by themselves will result in many financial, psychic, and opportunity costs (see Eide 2000, pp. 351–352 for a discussion of these—e.g., lost employment and legal fees), and these costs have the potential to affect an individual's decision to engage in crime (e.g., Bierschbach and Stein 2005).

Purposes of Criminal Sanctions

The application of criminal sanctions to an offender is usually justified in one of two ways. Retributive theories are premised on the idea that, by committing the crime, the offender has become morally blameworthy and is deserving of punishment. A criminal sanction does justice (for society, for the offender, for the victim) by punishing the offender, with the degree of

punishment having a direct relationship to the seriousness of the offender's moral culpability (which in turn has some relationship to the seriousness of the harm), at least according to some retributive views. Utilitarian theories, by contrast, are forward looking, focused on future consequences: a criminal sanction is appropriate if the benefits that follow from imposing the sanction outweigh the costs and suffering of its imposition. The punishment that generates the most net benefit is morally preferable. Mixed theories of punishment draw from both sets of ideas, usually with the hope of accounting for our intuitions about whether and how much punishment is appropriate. For instance, certain limited forms of retributivism assume the consequences of criminal sanctions matter at the margin but that moral desert establishes constraints on the levels of punishment that are minimally required and maximally permitted.

Economists are primarily interested in utilitarian or consequentialist theories of punishment, and in setting punishments, they seek to maximize net social welfare by reducing the total harm that results from crime—including the administrative costs and consequences to offenders of imposing criminal sanctions, in addition to the harms suffered by victims of crime and society. In practice, criminal sanctions can reduce future harm in one of three ways, although these categories are not always precisely demarcated: rehabilitation, incapacitation, and most important from the perspective of economists, deterrence.

Rehabilitation involves using criminal sanctions to change the preferences of offenders or to improve an offender's range of legal choices, thereby making the commission of crime less attractive. Incapacitation entails increasing the costs to the offender of making illicit choices or simply limiting the number of illicit choices available (Freeman 1999, pp. 3540–3541). In the latter case, incapacitation can succeed even in the face of an entirely irrational offender. For a discussion of the economics of these purposes of punishment as well as brief notes on the economics of retribution, see Shavell (2004, pp. 531–539).

Deterrence assumes at least some offenders are at least somewhat rational, which means that

they must be capable of responding to incentives and to their environments generally, a proposition that seems obviously true (Shavell 2004, p. 504). When deterrence can work, it is particularly attractive because sanctions need only be employed when an individual breaks the law. In the limit, if deterrence is perfect and there are no law enforcement mistakes, then society can achieve the first best (at least if any benefits to an offender of committing a crime are not included in the calculus): no harm comes to victims and no harm comes to potential offenders. What is more, the nature of the criminal sanction is unimportant, so long as deterrence is complete. Outside of this extreme, however, the relative advantages of different criminal sanctions matter a great deal in determining the socially optimal punishment regime.

Economic Model of Crime and Deterrence

There are at least a few different ways to conceive of developing an economic model of crime. The most straightforward starts with a set of crimes fixed by assumption (e.g., by taking the set of criminal prohibitions enacted by a government as given and simply assuming it includes the harmful acts society ought to criminalize). Analysis then begins with the question: "What steps should the government take to minimize the aggregate net harm of these acts?" To simplify matters, the discussion below proceeds largely using this framing of the problem.

As suggested above, however, a more general approach to the economic model of crime begins first by endogenously identifying specific harmful acts (either because they are harmful in themselves or because they risk harm) that are socially undesirable (Shavell 2004, p. 471). According to the standard economic account, a socially undesirable act is one in which the expected social benefits that result from the act (which may or may not include gains that accrue to the offender) are outweighed by its associated harms. Theoretically, categorizing acts in this way requires comparing the social costs and benefits of the act under all potential enforcement regimes

(including no enforcement), making the problem anything but trivial. For certain harmful acts, even the most efficient method of enforcement will prove too costly (Shavell 2004, p. 486); in these cases, minimizing harm is best accomplished through decriminalization, leaving it either to private parties to enforce through civil enforcement mechanisms or to social norms (Polinsky and Shavell 2007, pp. 446–447).

Starting with the categories of acts it considers sufficiently undesirable to criminalize (or with all acts if the set will be endogenously determined), society seeks to minimize the net social harm (or maximize the net benefit) from these, potentially using criminal sanctions in combination with other policies.

The economic model of crime focuses on the individual offender's decision to commit one of these acts, asking whether, given the offender's preferences and environment, his individual benefits of proceeding outweigh his individual costs of doing so (Becker 1968). Most economic models of crime assume that at least some share of potential offenders are capable of responding in predictable ways to incentives created by criminal sanctions (Eide 2000, pp. 352–355). If individuals are incapable of rational behavior in this sense, then enforcement of any kind must be premised either on rehabilitation or incapacitation or both. By assumption, criminal prohibitions are enforced through the probabilistic application of criminal sanctions. The optimization problem is how best to select and structure the application of these sanctions, given their costs and the likely response of relevant actors. Again, this approach may mean allowing certain harmful acts to go unpunished.

Modeling Criminal Behavior

Suppose an individual considers committing a criminal act instead of engaging in some legal (perhaps income-generating) activity. Assume the relative benefits to that individual for carrying out the act *absent* a public enforcement regime (net of the forgone benefits from the competing legitimate activity) amount to b_o . Note that there may also be some social benefit to this act, either b_o (i.e., if society "counts" the gains to the offender and there are no other benefits) or some other lesser or

greater amount, which we may call b_s . By assumption, b_s will not affect the decision of the offender to engage in criminal activity, but b_s may affect whether society ought to label the act a crime as well as how best to enforce any prohibition. If b_o is positive, then by definition the individual will engage in the activity, absent some threat of a criminal sanction.

Now assume a public enforcement regime is in place, and if the offender is caught engaging in the now-prohibited activity, he will be apprehended and criminally sanctioned (e.g., fined or imprisoned for some length of time). This sanction is costly, both to the individual (s_o) and possibly also to society (s_s) (again, s_s may be larger or smaller than s_o).

Public enforcement of criminal prohibitions is infused with uncertainty. Not all criminal activity is detected; if it is detected, not all of it results in an arrest; if the activity results in an arrest, not every arrest results in a conviction that leads to a sanction (Nagin 2013, pp. 207–213). Furthermore, there is the possibility of plea bargaining—i.e., “settlements dilute deterrence,” as noted by Polinsky and Shavell (2007, p. 436). Even the formal criminal sanctions announced after conviction are sometimes overturned on appeal or adjusted down the road (e.g., early release). Often, there is also a considerable time lag between the commission of a crime and the application of any sanction to the offender.

At the same time, in some contexts, arrests and criminal process function in effect as additional punishment (conditional only on arrest) because dealing with the police and with criminal allegations is also costly, regardless of whether the offender is ultimately convicted of the crime. At an even more basic level, anxiety over possible detection increases the effective severity of any potential punishment. Criminal sanctions also result in other lost opportunities and collateral consequences not captured by the formal sanction itself (e.g., loss of employment).

Finally, uncertainty also cuts the other way: individuals innocent of any crime may be wrongfully convicted of a crime (Polinsky and Shavell 2007, pp. 427–429), which means even deciding to forgo committing a harmful act does not

completely insulate someone from criminal sanctions, although we assume that the probability of a criminal sanction is much higher if the person engages in the prohibited activity.

Let p represent the net increase in the probability that an individual is punished if he chooses to commit the criminal act in question, taking into account all of the considerations outlined above. In effect, p implicitly adjusts for the fact that a criminal sanction ultimately imposed (i.e., experienced) is very different from the one laid out in the law. With certainty, p will not only be less than one (unless detection is extremely high, as it may be for some crimes) but will also be greater than zero, at least in any society in which law enforcement does not arbitrarily mete out punishment. Importantly, p is likely to be specific to an individual; some individuals are better able to avoid detection and conviction, and others are presumably more likely to be wrongfully prosecuted or subjected to more onerous criminal process.

The phrase “deterrence is a perceptual phenomenon” (e.g., Nagin 2013, p. 215) refers to the ideas that (1) the offender must perceive the threat of being criminally sanctioned in order to reasonably expect the offender’s behavior to change and that (2) an offender will only be deterred to the extent of the *perceived* level of enforcement or severity of the criminal sanction. Legal systems typically assume that individuals are aware of the law as it appears on the books, and potential offenders typically have some sense of the potential consequences they face for engaging in criminal behavior, but any such knowledge is imperfect (Shavell 2004, p. 481). Allow $f(p \times s)$ to represent a function that captures how potential offenders perceive threatened criminal sanctions and public enforcement more generally. Again, this perception is specific to an individual (Eide 2000, p. 352), as some individuals will be more experienced, better educated, and so on.

These assumptions and this notation allow us to characterize the decision of the potential offender as willing to commit an offense (rather than engage in some other lawful activity) if the net expected utility (relative to alternatives) is positive: $u_o(f_o(p_o \times s_o), b_o) > 0$. Put in the simplest of terms, the potential offender weighs the

net benefits of harmful conduct (absent any threat of enforcement) and compares this to the perceived net individual costs of becoming subject to criminal enforcement. Note that this margin is, by assumption, net of competing legal opportunities that are otherwise available to the individual. Potential offenders who expect to receive relatively large net benefits from the illicit activity, those who underestimate public enforcement efforts, and those who do not expect to suffer as much as others from any sanction, for example, are more likely, all else equal, to choose to offend.

Assuming some level of offender rationality and access to information, the model tells us that society can influence an offender's calculation (and therefore behavior) by changing (1) the level or type of criminal sanction (s_o), (2) how it will be enforced and how likely it is to be enforced (p_o), (3) how the sanction and enforcement are perceived ($f(\cdot)$), and (4) the net benefit (relative to legal pursuits) to the offender of the activity (b_o). Deterrence research traditionally focuses on the first two variables (on the theory that they are more policy relevant), taking into account the potential offender's preferences and other relevant attributes (particularly, risk aversion and wealth or income levels). Reducing the returns to undetected crime also matters in this model, however, and society can always modify potential offender behavior by altering an individual's preferences, although the analysis is agnostic on how a society might accomplish this.

Importantly, a complete theory of optimal public enforcement incorporates this range of strategies and the expected response of offenders to each of them, but also much more that is beyond the scope of this entry's discussion. Identifying how to maximize a social welfare function requires not simply accounting for the effects on offender behavior of criminal sanctions and the degree of enforcement (deterrence). Also vital are the social costs of these activities (Polinsky and Shavell 2007; Miceli 2009), including consequences for victims. Here, the focus is not the optimal enforcement bundle, but deterrence alone—specifically, what models and empirical work can tell us about changing offender behavior by altering

criminal sanctions and, to a lesser extent, enforcement levels.

Theoretical Predictions

In a simple deterrence framework, the predictions of how changes in criminal sanctions or enforcement levels affect behavior follow fairly straightforwardly from traditional ideas about consumer choice (or labor supply) and decision making under uncertainty (see, e.g., Freeman 1999). As Eide (2000, p. 345) puts it, deterrence theory is “nothing but a special case of the general theory of rational behavior under uncertainty. Assuming that individual preferences are constant, the model can be used to predict how changes in the probability and severity of sanctions. . . may affect the amount of crime.” For the sake of brevity, the discussion below concentrates principally on the potential effects of increasing or decreasing criminal sanctions or the degree of enforcement effort and on the consequences of using different types of criminal sanctions.

To organize the discussion, observe the following: First, as theoretical work makes clear, a potential offender's attitude toward risk (whether risk averse, risk neutral, or risk preferring) has significant implications for criminal behavior in virtually every remotely realistic model of deterrence (e.g., Eide 2000, p. 350). Second, the consequences of different risk attitudes differ depending on the policy lever in question, whether it is the level of enforcement activity, the severity of the criminal sanction, or the type of criminal sanction.

Furthermore, it is important to recognize that theoretical work and empirical evidence on deterrence and offender behavior are most useful in the interior of the policy choice set, not at the extremes. Shavell (2004) has written that “optimal law enforcement is characterized by under deterrence—and perhaps by substantial under deterrence—due to the costliness of enforcement effort and limits on sanctions” (p. 488), implying, if it is not already obvious, that a nontrivial portion of potential offenders will not be deterred in any reasonably practicable regime. At the same time, no one doubts criminal sanctions deter in some basic sense if viewed from the perspective

of lawlessness, whether zero sanctions or no enforcement. As Robinson and Darley have stressed in a number of articles (e.g., 2003, 2004, among others), the relevant policy question is whether criminal sanctions deter effectively *at the margin* or whether long prison sentences deter more effectively *relative to* more moderate but still significant sentences.

Begin by stepping back from the model laid out above and consider perhaps the simplest choice environment available: the decision to allocate a fixed amount of money between legal and illegal activities (see Schmidt and Witt 1984, pp. 151–154). The legal investment provides a certain return; the illegal investment offers a higher return, but the possibility of apprehension (p) and a fine (s) makes it uncertain. The relative returns of these investments vary continuously across investors in such a way as to create two corner solution groups; some proportion makes only legal investments, while others make only illegal investments. Now, consider an increase in p . In all states of the world, the illegal investment offers a weakly lower return, and so under minimal conditions, some proportion of investors near the cutoff move from making illegal investments into making legal investments, regardless of risk preference. Similarly, an increase in the fine (s) also lowers the return on average and weakly in all states, and it does so in a way that generates significantly more risk; so assuming risk aversion or risk neutrality, deterrence also occurs (with fewer illegal investors). These outcomes align roughly with Becker's (1968) conclusions, which suggest that increasing fines rather than enhancing the degree of enforcement will prove more effective at deterring criminal behavior, assuming potential offenders are risk averse.

A more complicated scenario emerges from a somewhat more realistic framework that allows for leisure time as an element of the utility function, even with simplifying assumptions. Imagine a potential offender considering an act enforced by threat of a monetary fine. Assume the crime in question is fraud and the benefit is monetary and that the offender could otherwise spend the time lawfully employed, earning wages with certainty, or enjoying leisure. For ease, assume that u , f ,

and b and any nonlabor income are fixed and that the initial values of s and p would result in the offender choosing to commit at least some fraud but also realistically engaging in a minimum of some leisure.

What are the effects of increasing either the size of the fine (s) or the degree of enforcement (p)? A portfolio or labor supply analysis provides the answer: because an increase in either s or p would effectively reduce the "return to crime," the prediction is ambiguous. The substitution effect alone will typically point in the direction of less crime (or fewer people engaging in crime), but the income effect may point in the opposite direction, depending on offender attitudes toward risk. More precisely, when the question is which of these two effects dominates, the "tipping point" is determined by the curvature of the utility function (i.e., by the level of risk aversion).

In many deterrence models (Eide 2000, pp. 348–350), the results do wind up supporting casual deterrence intuitions—i.e., that raising p or s results in less time devoted to crime (Eide 2000, p. 347)—but a great deal always turns on risk attitudes. And, as the models become more realistic, other technical conditions, such as the effect on the illegal gain if the offender is apprehended, become important. Schmidt and Witte (1984, p. 160), for instance, report that even the *source* of a change in p (e.g., an increase in the arrest rate versus an increase in the conviction rate) can result in a stronger or weaker prediction. In their most complicated (read: realistic) deterrence model, Schmidt and Witte (1984, p. 164) acknowledge that unambiguous predictions are not possible without four strong assumptions regarding (1) attitudes toward risk, (2) net returns to crime and the net gains to crime if the individual is convicted, (3) how the marginal utility derived from one activity changes with the time allocated to another, and (4) the way in which the marginal utility of an activity changes with a change in income. They conclude their excellent review (on which Eide (2000) and this entry both draw heavily) by noting:

Economic models that allow either the level of sanctions or the time allocation to enter the utility function directly are more appealing intuitively. They allow individuals to have different attitudes

toward sanctions, work, and illegal activity, and they answer the question of how leisure is determined. However, this increased realism is not without its price. These models do not allow us to determine unambiguously how changes in criminal justice practices or opportunities to conduct legal activities will affect participation in illegal activities. The ultimate effect of stiffened penalties [(*s*)], increased probabilities of apprehension [(*p*)], or improved legal opportunities becomes an empirical question (Schmidt and Witt 1984, p. 183).

Still, as between increasing criminal sanctions (*s*) or increasing the degree of enforcement (*p*) as a means of reducing criminal activity, deterrence models support the relative robustness of increasing *p*, given the possibility that offenders may be risk loving. As Eide (2000, p. 347) generalizes, “[b]oth the probability and the severity of punishment are found to deter crime for a risk averse person. For risk lovers, the effect of severity of punishment is uncertain.” However, if *p* is disaggregated into the probability of arrest, the probability of conviction conditional on arrest, and so on, different conclusions are possible with respect to the “conditional” levers (Schmidt and Witt 1984, p. 160). It is worth reiterating that other technical assumptions also matter to whether these models can generate useful predictions, such as whether all costs and benefits can be monetized and the precise nature of the enforcement probabilities (Eide 2000, p. 350).

Finally, in most of these models of offending behavior, the level of risk aversion will affect how *p* and *s* can be combined to achieve a certain level of deterrence. When risk aversion levels are high, for example, increasing the severity of sanctions will be a comparatively easier strategy for increasing the offender’s costs of criminal activity (Shavell 2004, p. 480).

Types of Criminal Sanction: Implications for Deterrence

A rich theoretical literature exists on the relative merits of using monetary fines and nonmonetary sanctions, like incarceration. Many of the key results, however, turn on the social cost of imposing the sanction in question. For instance, all else equal, fines are superior to imprisonment because, theoretically, a fine is less expensive to impose

and is effectively just a transfer from the offender to society (i.e., “utility” is destroyed by incarceration, not transferred to someone else). Of course, comprehensive summaries of these results like Shavell (2004) and Polinsky and Shavell (2007) also acknowledge that sanction type is relevant to offender behavior in the first instance and therefore to social welfare calculations, but direct social costs typically feature more prominently. By contrast, this section will focus on a few implications for deterrence of choosing between different types of criminal sanctions.

While attractive on the grounds that they are less socially costly than other types of sanctions, monetary sanctions often face an essential practical difficulty: offenders often, if not usually, have too few assets for the optimal fine to be imposed (Piehl and Williams 2011, p. 117). When this is the case, often with respect to serious crimes, fines will underdeter relative to other sanctions. As the saying goes, “you can’t get blood from a stone”; the actual scope of punishment is therefore limited to what can reasonably be extracted from an offender. By definition (or at least the assumption is that), courts and corrections departments are unable to force a person to work (else, the fine becomes nonmonetary). In the limit, if a potential offender has no assets whatsoever and no realistic prospect of acquiring (or incentive to acquire) assets in the future that can be made subject to attachment, fines will produce no deterrence. Furthermore, once assets are exhausted, marginal deterrence is absent, also.

Technically, all sanctions have this feature. For instance, we only have so much time to live, and so a threat of incarceration will fail to deter someone who is certain he will die in the next 24 hours. But there are important reasons why the average potential offender’s circumstances are more likely to allow him to avoid (and know he can avoid) a monetary sanction. First, although not an iron law, most potential offenders are financially insecure. Indeed, in the case of crimes with a financial motive, the source of the benefit of the crime may be the offender’s lack of assets or alternative options for earning a living. Second, most potential offenders are comparatively young, which means that even if they have no financial resources, they

have other assets—years of their lives—that will serve as a better foundation on which to base a sanction. One can compensate for the limited assets of offenders by increasing the likelihood that any crime is detected, but this is a costly response and is insufficient when adequate deterrence requires a high fine.

Monetary sanctions—or at least significant ones—may also differ in other key ways from nonmonetary sanctions. Most importantly, levels of risk aversion—central in identifying deterrence model predictions—will almost certainly differ depending on whether the potential offender faces fines or nonmonetary sanctions. There is little evidence in support of the idea that potential offenders will be risk loving with respect to an increase in a threatened fine, whereas offenders may anticipate becoming accustomed to long imprisonment after so many years, increasing their relative preference for risk (Shavell 2004, pp. 503 n.17, 508).

Fines also differ from other forms of criminal sanctions in how potential offenders perceive them. Piehl and Williams (2011, p. 116) suggest that fines are viewed as “softer” than other forms of punishment—perhaps just a price, one that is not often paid—and that potential offenders’ responses to fines relative to other sanctions will be heterogeneous. Briefly stepping outside of the neoclassical economic framework, cognitive and behavioral biases clearly play an important role in distinguishing types of sanctions in the minds of offenders. Also worth noting is the fact that a monetary fine may be easier to set optimally when the crime is one that results in a financial benefit for the offender or a financial harm to a victim, as the fine imposed can be explicitly linked to these values.

Modern nonmonetary sanctions as a group tend to involve (1) constraints on where an offender can go, what he can do, who he can see, or where he can live (e.g., incarceration, residency restrictions, or banishment), (2) affirmative obligations that require time and effort (e.g., registration requirements), or (3) the imposition of psychic harm, usually through some form of humiliation (e.g., shaming penalties generally or sex offender community notification requirements). Of the possible nonmonetary sanctions, intermediate or alternative

sanctions—which were defined above—appear to be increasingly attractive from the government’s perspective because they are less costly than incarceration, because they do not require that an offender has wealth or income (Miceli 2009, p. 279), and because technological innovation offers the potential to reclaim otherwise forgone incapacitation benefits. At the same time, there are good reasons to surmise that offenders will perceive these sanctions in very different ways and that the effective cost of these sanctions to the offender will depend on characteristics more difficult to measure than the financial assets or wages of the offender.

Moreover, these sanctions, especially intermediate or alternative sanctions, may generate deterrence using more complicated mechanisms. For instance, relative to fines, we know incarceration may work differently across the board and with respect to specific people: a different asset (opportunities during a period of time) is confiscated from the offender, but money is typically easier to value and to relate to how the offender benefits from committing the crime or the social harm the crime caused. On the whole, though, all sanction types, if set appropriately and with an eye toward these concerns, should be capable of playing the generic role of *s*, shifting the perceived cost of committing a crime in a way that may cause the offender to use his time differently.

Nevertheless, there are at least two important reasons for explicitly considering the specific roles that the various types of criminal sanctions might play in any model of offending behavior, both of which can add to our understanding of criminal deterrence generally.

First, as pointed out at the end of the discussion of the types of criminal sanctions, the *simultaneous* use of two or more different kinds of sanctions may generate different deterrent effects than a comparable amount of punishment using a single sanction type. (Note: This proposition is distinct from the result from the public enforcement literature that an optimal sanction scheme involves first exhausting wealth through fines before turning to nonmonetary sanctions. This well-known idea turns on fines being superior but limited by the liquidity constraints of the offender.) Prescott and

Rockoff (2011) find evidence consistent with this idea in the sex offender law context, reporting surprisingly strong deterrent effects from the addition of post-release notification requirements (effectively, shaming) to be imposed at the end of already very long sentences. Although comparing the two is complicated, research that has examined simply the addition of more prison time at the end of a long sentence appears to suggest that “more of the same” had weaker effects (Kessler and Levitt 1999).

If there are diseconomies of scale or perhaps economies of scope in the deterrent effects of criminal sanctions, one likely mechanism is that offenders perceive the sanction regimes differently: adding distinct sanctions may seem worse than simply adding more of a sanction already in play. It is also possible that increasing the severity of punishment by adding a new type of sanction does not invite additional risk taking by offenders who are risk lovers, allowing the reduction in criminal activity by the risk averse to more clearly dominate the final deterrence tally.

Second, even if two criminal sanctions are thought to be equivalent in terms of their *deterrent* effects from the perspective of a potential offender, criminal sanctions may differ in how they influence ex post offender behavior down the road, leading potentially to different criminal activity levels in the aggregate. For example, the use of incarceration not only deters potential offenders ex ante but it incapacitates those who do choose to offend, which may offer separate social value in precluding subsequent offenses. There is no a priori reason to think that these indirect effects are always positive nor that they function solely in ways unrelated to deterrence.

Indeed, collateral effects of criminal sanctions may have indirect consequences for deterrence itself—and they may in fact *encourage* offending. Sex offender notification laws, for example, are intended to reproduce an upside of incarceration, essentially incapacitating potential recidivists by using information to aid potential victims in creating a barrier between themselves and released sex offenders. Unfortunately, notification also produces many deprivations in addition to the

incapacitation effect: finding employment and housing are difficult, public shaming results in strained relationships, and so on. Furthermore, these deprivations do not depend on whether the offender engages in criminal behavior. As a result, although the threat of notification may deter criminal offending ex ante by raising the costs of committing a sex crime, the application of the requirements ex post may hamper deterrence, by reducing the relative benefits of staying on the straight and narrow (Prescott and Rockoff 2011). Unless the incapacitation effect more than offsets this reduction in deterrence (or if notification increases p sufficiently through additional contact with law enforcement or public monitoring), criminal activity should increase, all else equal.

Empirical Evidence on Deterrence

There is a substantial and growing body of empirical evidence on criminal deterrence, and this entry only offers a brief summary of this research. For recent and detailed discussions of key papers and open questions, see generally Nagin (2013), Durlauf and Nagin (2011), Levitt and Miles (2007), Robinson and Darley (2004), Eide (2000), and Freeman (1999). Recognizing that empirical work must study changes from an existing framework of criminal sanctions and enforcement levels and that there is no reason to think deterrent effects should be homogeneous across the different departure points, much less across all types of crime, types of sanctions, and sanction levels, the work is decidedly mixed. Notwithstanding these caveats and all of the others implicit in empirical work in general, some basic points of conventional wisdom have emerged.

First, as a general matter, most scholars interpret the literature as strongly supporting the ability of criminal sanctions and enforcement to deter criminal behavior. The evidence in favor of “substantial” deterrent effects across “a range of contexts,” according to some, is “overwhelming” (Durlauf and Nagin 2011, p. 43). The details, however, reveal quite a bit of variation (in magnitudes, in particular), and others strongly disagree with this general conclusion—at least in

terms of the policy implication that an increase in s or p from present levels would result in less crime—on the basis of alternative evidence that appears inconsistent with it (like evidence that suggests many offenders are unaware of the size of criminal sanctions) and methodological objections (see, e.g., Robinson and Darley 2004), which are particularly strong with respect to early work and which are discussed in most reviews. Eide (2000, pp. 364–368) also highlights the importance of properly interpreting the empirical evidence.

Second, most scholars agree that increasing the degree of enforcement (p) (often referred to as “certainty” in the literature) appears to be more effective than a comparable increase in the severity of sanctions, at least given the current levels of each. Consistent with the models of Schmidt and Witte (1984, p. 160), Nagin (2013, p. 201) concludes that the evidence in favor of certainty is much stronger with respect to the probability of arrest and less so (or at least with more uncertainty) with respect to conviction and other later conditional enforcement probabilities. To the extent that deterrence is an important goal of criminal law, the take-away from these patterns is that crime policy should focus on how laws are enforced rather than on the severity of sanctions.

Third, all agree not only that increasing the severity of sanctions appears less effective than increasing the degree of enforcement but also that there is still some question whether, at current sanction levels, increasing severity will result in *any* reduction in crime (Durlauf and Nagin 2011). Studies of the effects of increased severity on crime are much more likely to produce estimates that are not statistically significant, and at times, the estimates are even positive (Eide 2000, p. 360). This body of work is actually remarkably consistent with what one would have expected to find given the conclusions of the theoretical work. In the standard models of criminal behavior, increasing sanction severity, for instance, was more likely to lead to more crime all else equal than increasing the certainty of enforcement, especially to the extent that the offenders in question are risk loving.

Concluding Remarks

The goal of this entry has been to contextualize the law and economics of criminal deterrence by linking research in economics to the range of criminal sanctions that exist in the real world. Some emphasis has been placed on the role intermediate or alternative sanctions might play in the deterrence framework, as there appears to be very little theoretical work on the subject. Unfortunately, many basic ideas from the research on deterrence are not reviewed in this entry. Optimal enforcement policy has only been mentioned in passing, and many key ideas in the deterrence literature have been omitted: marginal deterrence (see Shavell 2004, p. 518), specific deterrence (see Shavell 2004, p. 516), the roles that can be played by a fault standard or affirmative defenses in designing sanctions (see Shavell 2004, pp. 476, 494–495), and how best to deter repeat offenders (see Polinsky and Shavell 2007, p. 438).

References

- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Bierschbach RA, Stein A (2005) Overenforcement. *Georgetown Law J* 93:1743–1781
- Calabresi G, Douglas Melamed A (1972) Property rules, liability rules, and inalienability: one view of the cathedral. *Harv Law Rev* 85:1089–1128
- Durlauf SN, Nagin DS (2011) The deterrent effect of imprisonment. In: Cook PJ, Ludwig J, McCrary J (eds) *Controlling crime: strategies and tradeoffs*. University of Chicago Press, Chicago, pp 43–94
- Eide E (2000) Economics of criminal behavior. In: Boudewijn B, De Gerrit G (eds) *Encyclopedia of law and economics*, vol 5. Edward Elgar, Cheltenham, pp 345–389
- Freeman RB (1999) Chapter 52: The economics of crime. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, vol 3. North-Holland, Amsterdam, pp 3529–3571
- Kessler DP, Levitt SD (1999) Using sentence enhancements to distinguish between deterrence and incapacitation. *J Law Econ* 42(1):343–363
- Levitt SD, Miles TJ (2007) Chapter 7: Empirical study of criminal punishment. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*, vol 1. North-Holland, Amsterdam, pp 457–495
- Miceli T (2009) *The economic approach to law*, 2nd edn. Stanford Economics and Finance, Stanford

- Nagin DS (2013) “Deterrence in the twenty-first century”. *Crime and justice*. *Crime Justice Am* 1975–2025 42(1):199–263
- Piehl AM, Williams G (2011) Institutional requirements for effective imposition of fines. In: Cook PJ, Ludwig J, McCrary J (eds) *Controlling crime: strategies and tradeoffs*. University of Chicago Press, Chicago, pp 95–121
- Polinsky AM, Shavell S (2007) Chapter 6: The theory of public enforcement of the law. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*, vol 1. North-Holland, Amsterdam, pp 457–495
- Prescott JJ, Rockoff JE (2011) Do sex offender registration and notification laws affect criminal behavior? *J Law Econ* 54(1):161–206
- Robinson PH, Darley JM (2003) The role of deterrence in the formulation of criminal law rules: at its worst when doing its best. *Georgetown Law J* 91: 949–1002
- Robinson PH, Darley JM (2004) Does criminal law deter? A behavioral science investigation. *Oxf J Leg Stud* 23(2):173–205
- Schmidt P, Witte AD (1984) Chapter 9: Economic models of criminal behavior. In: *An economic analysis of crime and justice: theory, methods, and applications*. Academic, Orlando, pp 142–193
- Shavell S (2004) *Foundations of economic analysis of law*. Belknap Harvard.

Currency Demand

► Cash Demand

Custodian Liability

Jef De Mot¹ and Louis Visscher²

¹Postdoctoral Researcher FWO, Faculty of Law, University of Ghent, Belgium

²Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Definition

Liability of a person for losses caused by an object under that person’s control.

The Law on Custodian Liability

Custodian liability means that a person is liable for losses caused by an object under that person’s control. It has its origins in the French law (Wagner 2012). Article 1384(1) of the French Code Civil states that a person is responsible for the damage caused by things in his custody. Originally, this phrase was not intended as a separate cause of action but merely as a reference to the provisions 1385 and 1386 CC, which respectively hold keepers of animals and owners of buildings liable for damage caused by animals or defective buildings. Some landmark judgments by the Cour de Cassation however transformed this statement into the legal basis for making custodians liable for damages caused by objects of *any kind*. A claimant needs to show that the thing has contributed to the realization of the damage. A thing includes all inanimate objects. It can be dangerous or not dangerous, movable or immovable (e.g., a lift, a dyke), natural or artificial, and defective or nondefective. The contribution of the thing can be established in several ways (Fabre-Magnan 2010). If there has been contact between the claimant or his property and a *moving* thing (e.g., a falling tree or a rolling stone), it is simply assumed that the thing has contributed to the realization of the damage. The claimant does not need to show that the custodian contributed to the action of the thing. If there was contact between the claimant or his property and a *nonmoving* thing, the claimant needs to prove that the thing was instrumental in causing the damage in some way. He can do this by showing that the thing behaved in an abnormal way, more particularly that the thing was defective or was in an abnormal position. Finally, if there was no contact between the claimant or his property and the thing, Art. 1384(1) may apply, but the claimant needs to prove that the thing was the cause of his damage. The liable person in Article 1384(1) is the *custodian*: the person who had, at the time of the accident, the power to use, control, and direct the thing. Often this will be the owner of the thing, but also people other than the owner can be the custodian. The custodian can escape liability if he can prove an external cause of the damage which was both unforeseeable and unavoidable. The external cause

can be due to *force majeure*, the act of a third part or the act of the victim. This defense is generally rather difficult to prove. Also, the custodian can invoke the claimant's contributory negligence. If successful, the amount of compensation the custodian needs to pay is reduced.

Many legal systems that have adopted or integrated the French Code Civil have been influenced by the French model of custodian liability (Wagner 2012). The concept of custodian liability is known in, for example, Italy, Portugal, the Netherlands, Belgium, and Luxembourg. One important exception is Spain. Custodian liability is also unknown in German law, English common law, and more generally in countries outside the gravitational force of the Code Civil. Some countries which have introduced custodian liability limit this concept to cases where the object suffers from a defect (e.g., the Netherlands and Belgium). It is a general requirement that the object in question represents a source of abnormal danger to its surroundings. Case law in this respect frequently refers to the normal expectations in society with respect to the characteristics of the object.

The remainder of this contribution is structured as follows. Section “[Advantages of Strict Liability and Negligence in the Context of Liability for Harm Caused by Objects](#)” briefly discusses the relative advantages of strict liability and negligence for damage caused by objects under a person's control. Section “[Economic Analysis of Case Law on Custodian Liability](#)” provides an overview of how some European countries have actually applied the several concepts of custodian liability in an economically efficient way. Section “[Conclusion](#)” concludes.

Advantages of Strict Liability and Negligence in the Context of Liability for Harm Caused by Objects

Should custodian liability be an application of strict liability or negligence? In this section, we provide a brief overview of the main relative advantages of both rules (also see Schäfer and Müller-Langer 2009, pp. 3–45). There can be sound economic reasons to hold the custodian

strictly liable for losses caused by an object under his control:

- It provides incentives to the liable party to maintain his things, to regularly check and if necessary repair them, to replace components, etc.. In theory, negligence could also provide those incentives, but then the victim has to prove that the custodian did not take enough care and that he should have known about the existence of a defect. This will often be very difficult since many care measures may be non-verifiable (e.g., how often did the injurer check the brakes of his bicycle). Negligence would therefore frequently not result in liability and the behavioral incentives of liability would be frustrated. Therefore, the superior information of the custodian regarding which care measures he actually *has taken* pleads for strict liability. Especially with respect to immovables, the custodian will generally be better informed about the quality of the building or structure and about the care measures that are taken and that can be taken.
- Negligence can only give care incentives to the injurer in the dimensions that are incorporated in the negligence rule. Strict liability gives the injurer care incentives also in non-verifiable dimensions, which cannot be incorporated in a behavioral norm. On the other hand, when it is more important to provide care incentives in all dimensions to the victim rather than to the injurer, negligence is to be preferred: only then will the victim (being the residual loss bearer) receive care incentives in all dimensions.
- Similarly, strict liability is preferable when it is more important to control the activity level of the custodian than that of the victim (Shavell 1980, p. 2; Landes and Posner 1987, p. 61; Faure 2002, p. 366; Posner 2003, pp. 177, 178; Shavell 2004, p. 193). Many accidents can be regarded as unilateral because the victim has no influence on their incidence. In bilateral cases, where also the victim should take care, a defense of contributory or comparative negligence is necessary under strict liability to avoid that the victim takes no care at all because the injurer would be liable anyway.

- If the injurer has better information than the courts regarding which care measures he *can* take and what the costs and benefits of such measures are, strict liability is to be preferred. There the injurer himself weighs costs and benefits of care, whereas under negligence the court has to do that (Shavell 2004, pp. 18, 188; Landes and Posner 1987, p. 65).
- According to many law and economics scholars, the problem of judgment proofness is more severe under strict liability because the “reward” for being careful (reducing the accident probability) is lower than under negligence (fully escaping liability). Taking care under negligence is therefore also worthwhile for lower levels of wealth.
- If the injurer is a firm, strict liability induces loss spreading via the price of the product/service, whereas negligence leaves the loss (concentrated) with the victim.
- Insurance issues such as moral hazard and adverse selection often give a preference for negligence. Especially in situations which could be regarded as unilateral, an insured victim cannot display moral hazard, whereas an insured injurer could (Bishop 1983, p. 260; Faure 2005, p. 245, 261). In as far as liability insurance makes use of experience-rated premiums, deductibles, etc., the relative importance of this issue may be limited.
- Finally, administrative costs are generally assumed to be lower under strict liability, especially because more settlements are expected (there is no dispute over the true and due levels of care).

Economic Analysis of Case Law on Custodian Liability

De Mot and Visscher (2014) argue that Dutch and Belgian courts apply custodian liability in an economically efficient way (full references to the cases mentioned in this chapter can be found in De Mot and Visscher (2014)). First, the courts use the concept of defect to allocate costs to the least cost avoider. Courts often conclude that there is a defect when it is more important to provide

incentives to the injurer to take (often) unverifiable precautions and that there is no defect when it is more important to provide incentives to the victim to take such precautions. This way, the accident losses are allocated to the “least cost avoider.” This is especially clear in cases where the floor of a shop becomes slippery (or more dangerous) due to something laying on the floor, e.g., a vegetable leaf in the vegetable department of a supermarket. According to Belgian and Dutch courts, the mere presence of the leaf does not make the building defective. In a Dutch case, the court first noted that the floor is cleaned every morning, that there is always personnel present which is instructed to pick up litter if they observe it, and that there is active monitoring on littering. These measures are relatively easy to verify. What is much more difficult to verify is (1) whether the personnel always picks up litter when they observe it and (2) how careful the client walks in the shop. So the question becomes: if we can only give one party incentives to engage in non-verifiable measures, who do we choose to incentivize? Which party’s non-verifiable measures are most productive? Given that it is unreasonable (read: too costly) to require the shop to keep the floor clean all the time, clients have to realize that vegetable leaves (or comparable litter) may lie on the floor. It is thus important that clients are prudent. They are the least cost avoiders (with respect to the non-verifiable precautions). A comparable decision has been taken in Belgian and Dutch cases where someone slipped on the floor of a supermarket near the entrance which was wet due to the rainy weather and soaked carpets. Again, it is more difficult for the supermarket to avoid the floor from becoming wet and slippery also in situations of rain than it is for individuals to walk carefully in situations where the water and dirt were clearly visible. As the Belgian court stated: “During rainy weather, it is impossible for a supermarket to make sure the floor is dry continuously.” Furthermore, in cases in which the courts decided that there actually *was* a defect, it is clear that the customer was *not* the least cost avoider. The courts, for example, decided that the supermarket floor was defective in the following cases: ice cream on the floor of

the perfume department, oil on the floor of the food department, and a vegetable leaf on the floor in the textile department. Customers do not expect ice cream in the perfume department or vegetable leaves in the textile department, and they cannot be the least cost avoider. The shop owner is better placed to instruct his personnel to be watchful. As the court stated in the case of oil on the floor of the food department: “the (lower) court could lawfully decide that the floor of the food department was defective by establishing that the floor was exceptionally slippery due to an oil stain which was difficult to observe for a normally attentive customer and that an oil stain in this department is not an everyday phenomenon which clients reasonably have to expect and avoid.” De Mot and Visscher (2014) discuss many other slip and fall cases (e.g., a woman tripping over a gasoline hose at a gasoline station, a customer slipping over French fries in a chips shop, a bicyclist slipping when crossing wet train tracks) and other types of cases (e.g., a man ordering pheasant in a restaurant and damaging his teeth because there was still some lead in the pheasant, a shed catching fire while welding work was executed there) in which the courts make their decision in an economically efficient way.

Second, in some cases the discussion is not about whether the object is defective, but there is disagreement over who is the custodian. In these cases, the courts fill in the *concept of custodian* as if they place liability on the least cost avoider. For example, a player of a sports club got injured due to an iron bar sticking out from the field but hidden under the grass. The sports club had permission from the renter of the field to use the field for a couple of hours. The Belgian court concluded that the renter and not the owner of the field or the sports club was the custodian of the sports field, since the renter was the party who “called the shots” and took all the factual decisions with respect to the field; the sports club was only permitted to use the field after permission of the renter and only for a couple of hours and in the condition that the field was in. Hence, the renter was better placed to prevent the accident than the owner or the sports club. In a Dutch case, the victim fell off a stairway to the basement. He

was an employee of a telecom company who had to repair a connection in a shop, which was located at the top of a stairway. The employee walked down the stairway because the lid of the connection point was missing, and he wanted to check if it had fallen onto the basement floor. The fore-last step of the stairway turned out to be missing and the employee fell down the stairway. The court holds the renter of the building, who operated his shop there, strictly liable because there was a step missing and there was no adequate warning against this danger. The renter argued that the owner of the building should be held liable, but the court applies Art. 6:181 CC, which states that the business user rather than the possessor is the liable party. This makes economic sense: the renter of a building uses the building on a daily basis and is better informed about possible risks (such as a missing step) than the owner, who might not enter his rented building for substantial periods of time. It is therefore better to incentivize the renter to use his superior information (either by repairing the problem himself or by informing the owner that he should repair the stairway) than to require the owner to regularly check his buildings in which another party operates his business. More generally, in cases in which it is debated whether the letter or the renter was the custodian, Belgian courts have taken into account some elements that can obviously be linked to the least-cost-avoider concept. The following elements make it more likely that the letter is the custodian and not the renter (see Vansweevelt and Weyts (2009), pp. 485–487): limits on the use of the object for the renter (the object can only be used for a limited period, for a specific goal, or in a certain area), whether the letter is present when the object is used and whether the letter is responsible for the maintenance of the object. These considerations make economic sense. When one or more of these conditions is fulfilled, it is more likely that the letter and not the renter is the least cost avoider.

Conclusion

Economic analysis of tort law has shown under which circumstances strict liability creates better

incentives than negligence and vice versa. Often, whole classes of cases in which strict liability should apply can be described in a rather simple way (e.g., for transporting and/or working with dangerous substances). In other types of cases, however, like slip and fall cases, what is the optimal rule is more ambiguous. In some cases, strict liability leads to better results, and in other cases negligence is superior. In some European countries, this difficult exercise is performed by the courts through the interpretation of the concepts defect and custodian. The variety of solutions which is reached in this manner cannot be fully replicated with a binary choice between strict liability with a defense of contributory negligence and negligence with a defense of contributory negligence. If we choose a rule of strict liability with a defense of contributory negligence, we cannot reach optimal solutions in cases in which the victim's non-verifiable measures are very likely to be the most productive ones. And if we choose a rule of negligence with a defense of contributory negligence, we cannot reach optimal solutions in cases in which the injurer's non-verifiable measures are very likely to be the most productive ones. The interesting feature of the concept of a defect as it is interpreted by the Belgian and Dutch courts is that it enables the court to select the liability rule that is most suitable for the case at hand.

This field of tort law can still benefit from further economic research. For example, De Mot and Visscher (2014) have focused on Dutch and Belgian case law. Whether courts in other jurisdictions interpret the various concepts related to custodian liability in an economically efficient way is yet to be discovered. Also, it would be interesting to compare the concepts of defect and custodian with the concepts that are used in countries in which custodian liability is unknown and which may also have the effect of efficiently sorting between cases.

References

- Bishop W (1983) The contract-tort boundary and the economics of insurance. *J Legal Stud* 12:241–266
- De Mot J, Visscher LT (2014) Efficient court decisions and limiting insurers' right of recourse. The case of custodian liability in the Netherlands and Belgium. *Geneva Papers Risk Insur – Issues Practice* forthcoming, 39 (3):527–544
- Fabre-Magnan M (2010) *Responsabilité Civile et Quasi-Contrats*, 2nd edn. Presses Universitaires de France, Paris
- Faure MG (2002) Economic analysis. In: Koch BA, Koziol H (eds) *Unification of tort law: strict liability*. Kluwer Law International, Den Haag, pp 361–394
- Faure MG (2005) The view from law and economics. In: Wagner G (ed) *Tort law and liability insurance*. Springer, Vienna, pp 239–273
- Landes WM, Posner RA (1987) *The economic structure of tort law*. Harvard University Press, Cambridge, MA
- Posner RA (2003) *Economic analysis of law*, 6th edn. Aspen Publishers, New York
- Schäfer H-B, Müller-Langer F (2009) Strict liability versus negligence. In: Faure M (ed) *Tort law and economics*, vol 1, 2nd edn, *Encyclopedia of law and economics*. Edward Elgar, Cheltenham, pp 3–45
- Shavell S (1980) Strict liability versus negligence. *J Legal Stud* 9:1–25
- Shavell S (2004) *Foundations of economic analysis of law*. The Belknap Press of Harvard University Press, Cambridge, MA
- Vansweevelt T, Weyts B (2009) *Handboek Buitencontractueel Aansprakelijkheidsrecht*. Intersentia, Antwerp
- Wagner G (2012) Custodians liability. In: Basedow J, Hopt KJ, Zimmerman R, Stier A (eds) *Max Planck encyclopedia of European private law*. Oxford University Press, Oxford

Custom of Merchants

► [Lex Mercatoria](#)

Customary Law

Bruce L. Benson
Department of Economics, Florida State
University, Tallahassee, FL, USA

Abstract

Customary law, a system of rules of obligation and governance processes that spontaneously evolve from the bottom up within a community, guides behavior in primitive, medieval,

and contemporary tribal societies, as well as merchant communities during the high middle ages, modern international trade, and many other historical and current settings. Rules and procedures are recognized and accepted because of trust arrangements, reciprocities, mutual insurance, and reputation mechanisms, including ostracism threats. Negotiation (contracting) generally is the most important source for initiating change in customary law. Such agreements only apply to the parties involved, but others can voluntarily adopt the change if it proves to be beneficial. An individual also may unilaterally adopt behavior that others observe, come to expect, and emulate, or a dispute may arise that results in an innovative solution offered by a third party and voluntarily adopted by others. Many different third-party dispute resolutions procedures are observed in different customary communities, but they generally involve experienced mediators or arbitrators who are highly regarded community members. Some of these adjudicators may have leadership status, but leadership arises through persuasion and leaders do not have coercive authority. Since there is no coercive authority, protection and policing rely on voluntary arrangements, and community norms encourage and reward such activity. People who violate rules are generally expected to and have incentives to compensate victims for harms. Customary law is polycentric, with hierarchical arrangements to deal with intercommunity interactions. Customary law also may conflict with authoritarian law. When this occurs, a coercive authority may attempt to assert jurisdiction over a customary-law community, but this will have very different impacts depending on the options available to members of the community. The authority often adopts and enforces some customary rules in order to avoid conflict.

Synonyms

Informal law, Polycentric law, Private law, Stateless law

Definition

Customary law: a system of rules of obligation and governance processes that spontaneously evolve from the bottom up within a community.

“Customary law” sometimes refers to various immanent principles, so well established and widely recognized that sovereigns feel obliged to adopt them as law (e.g., as in common law recognition of “immemorial custom”), but it also can be applied to the legal arrangements within primitive communities. Another definition delineates the source of immanent customary principles, however, and encompasses primitive law: customary law is a system of rules of obligation and governance processes that spontaneously evolve from the bottom up within a community (Pospisil 1971). This definition is the focus of the following presentation. Those who contend that “law” consists of general commands of a sovereign will not consider this concept to be appropriately labeled as law, nor will those who contend that “law” applies to moral principles, whether logically derived or handed down by a higher being. Nonetheless, these rules and processes are real, and they influence behavior in very significant ways. This is true for the primitive (Pospisil 1971; Benson 1991), medieval (Friedman 1979), and contemporary tribal societies (Benson and Siddiqui 2014), merchant communities during the high middle ages (Benson 1989, 2014), modern international trade (Benson 1989), and many other historical and current settings (Fuller 1981).

Kreps (1990) stresses that socially or culturally derived experiences help players “know” what to do and predict what other players will do. In this context, Hayek (1973, pp. 96–97) observes that many issues of “law” are not “whether the parties have abused anybody’s will, but whether their actions have conformed to expectations which other parties had reasonably formed because they corresponded to the practices on which the everyday conduct of the members of the group was based. The significance of customs here is that they give rise to expectations that guide people’s actions, and what will be regarded as binding will therefore be those practices that everybody counts on being observed and which thereby

condition the success of most activities.” Similarly, Fuller (1981, p. 213) explains that “We sometimes speak of customary law as offering an unwritten code of conduct. The word *code* is appropriate here because what is involved is not simply a negation, . . . but of this negation, the meaning it confers on foreseeable and approved actions, which then furnish a point of orientation for ongoing interactive responses.” In this light, behavioral patterns individuals within a community are generally expected to adopt and follow in their various interdependent activities are considered to be legal rules below, whether those rules (expectations) arise through formal legislation or informal customary processes discussed below. Individuals are expected to follow such rules, and these expectations influence the choices made by other individuals. Customary law involves more than rules of behavior, however, as customary governance processes also evolve from the bottom up (Benson 1989, 2011, 2014; Pospisil 1971). Development and characteristics of both customary behavioral rules and procedural rules are discussed below.

Establishing and Recognizing Customary Rules

Vanberg and Buchanan (1990, p. 18) define rules of behavior toward others which individuals have positive incentives to voluntarily recognize as “trust rules” and explain that:

By his compliance or non-compliance with trust rules, a person selectively affects specific other persons. Because compliance and non-compliance with trust rules are thus “targeted,” the possibility exists of forming cooperative clusters.... Even in an otherwise totally dishonest world, any two individuals who start to deal with each other - by keeping promises, respecting property, and so on - would fare better than their fellows because of the gains from cooperation that they would be able to realize.

Game theory demonstrates that such cooperation can arise through repeated interactions, although the dominant strategy in bilateral games still depends on expected payoffs, frequency of interaction, time horizons, and other considerations (Ridley 1996, pp. 74–75). As

North (1990, p. 15) explains, however, game theory “does not provide us with a theory of the underlying costs of transacting and how those costs are altered by different institutional structures.” For example, if bilateral relationships form in recognition of the benefits from cooperation in repeated games, and if individuals in such relationships enter into similar arrangements with other individuals, a loose knit group with intermeshing reciprocal relationships develops. As this occurs, tit for tat becomes less significant as a threat, while expected payoffs from adopting trust rules increase. For instance, an exit threat becomes credible when each individual is involved in several different games with different players, in part because the same benefits of cooperation may be available from alternative sources (Vanberg and Congelton 1992, p. 426).

When the exit option becomes viable, Vanberg and Congleton (1992, p. 421) explain that a potential strategy is unconditional cooperation unless uncooperative behavior is confronted, and then imposition of some form of explicit punishment on the noncooperative player as exit occurs. They label such a strategy “retributive morality”; examples include the “blood feuds” of tribal (Benson and Siddiqui 2014) and medieval (Friedman 1979) societies. Retributive morality strengthens the threat against noncooperative behavior relative to tit for tat. Such violence is risky, however, and there is likely to be a better alternative. Since all community members have exit options, information about uncooperative behavior can be spread, creating incentives for everyone to avoid interacting with the untrustworthy individual. This suggests a strategy involving unconditional cooperation in all interactions with other community members, along with a refusal to interact with any individual who is known to have adopted noncooperative behavior with anyone in the group and the spread of information about such behavior. Vanberg and Congleton (1992) refer to this response as “prudent morality,” and given that reputation information spreads quickly and everyone spontaneously responds to information, the result is spontaneous social ostracism. Depending in part on severity of the offenses, however, ostracism may not be

absolute. Individuals might continue interacting with someone who has misbehaved, but only if certain conditions are met to reduce risk (e.g., a bond might have to be posted). Such conditions sanction an offender by raising costs or reducing his benefits in various interactions.

Many group-wide customary rules are simply commonly shared trust rules. Others, called “solidarity rules” by Vanberg and Buchanan (1990, pp. 185–186), are expected to be followed by all members of the group. The spontaneous development of social ostracism illustrates this. Solidarity rules produce community-wide benefits (Vanberg and Buchanan 1990, p. 115), including general deterrence and others discussed below.

Changing Customary Rules

Rules and communities evolve simultaneously: the evolution of trust rules leads to the development of a web of interrelationships that become a “close-knit” community, and the evolving web of interactions and expanding opportunities for interactions in turn facilitates the evolution of more rules. If conditions change or a new opportunity is recognized, for instance, and a set of individuals decide that, for their purposes, a new behavioral obligation arrangement will support more mutual benefits, they can voluntarily agree to accept (contract to adopt) the obligations. Such new obligations only apply for the contracting parties for the term of the contract. Individuals who interact with these parties learn about their contractual innovation, however, and/or members of the community observe its results. If the results are desirable, the new rules can be rapidly emulated. Such changes are initiated without prior consent of or simultaneous recognition by any other members of the relevant community, but they can be spread through voluntary adoption. Indeed, as Fuller (1981, pp. 224–225) explains, “If we permit ourselves to think of contract law as the ‘law’ that parties themselves bring into existence by their agreement, the transition from customary law to contract law becomes a very easy one indeed.” In fact, contract and custom are tightly intertwined and often inseparable:

if problems arise which are left without verbal solution in the parties’ contract these will commonly be resolved by asking what “standard practice” is with respect to the issues. . . . In such a case it is difficult to know whether to say that . . . the parties became subject to a governing body of customary law or to say that they have by tacit agreement incorporated standard practice into the terms of the contract.

. . . [Furthermore,] . . . the parties may have conducted themselves toward one another in such a way that one can say that a tacit exchange of promises has taken place. Here the analogy between contract and customary law approaches identity. (Fuller 1981, p. 176)

Negotiation (contracting) generally is the most important source for initiating change in customary law (Fuller 1981, p. 157), although there are others. An individual may unilaterally adopt behavior that others observe, come to expect, and emulate, for instance, or a dispute may arise that results in an innovative solution offered by a third party.

If new obligations are required to deal with the new situation, transaction costs may prevent individuals from agreeing on an appropriate arrangement. Similarly, one party may believe that a particular rule applies to a situation, while another may believe that a different rule is relevant. If direct negotiation fails, a solution may still be achieved if the parties turn to arbitration or mediation. Many different third-party dispute resolution procedures are observed in different customary communities (Benson 1991, 2011, 2014), but they generally involve experienced mediators or arbitrators who are highly regarded community members. They may be paid or they may voluntarily give their time in order to enhance their reputations. Whatever the process, a potential rule may be suggested by the dispute resolution (Fuller 1981, pp. 110–111). Unlike modern common law precedent, however, the resolution only applies to the parties in the dispute. If it suggests behavior that effectively facilitates desirable interactions, the implied rule can be adopted and spread through the community.

Reciprocity and Restitution

When retributive morality dominates, unilateral exaction of punishment can be very risky, as

noted above, and generate negative spillovers for the larger community. As a result, customary rules tend to evolve to reduce revenge seeking and encourage substitution of ostracism for retribution, as suggested above, but another option also commonly arises. To understand why, note that unregulated retaliation may result in greater costs for the alleged offender than the costs generated by the offender's initial offense, leading to an escalating chain of violence. In a dynamic society, this ongoing feud consumes resources, including human life, and, therefore, dissipates wealth. In subsistence societies the loss of such wealth can be devastating, but even as wealth expands, the opportunity cost of escalating feuds is high. Some of these costs are also likely to be external. Therefore, strong incentives arise among community members to constrain retribution. Parisi (2001) explains that rules generally evolve to specify who can pursue retribution (e.g., only the victim or a member of his extended family or support group discussed below) and to set an upper bound on the harm imposed on the offender based on the harm initially inflicted on the victim. Over time, the level of retribution moves to one of symmetry, "an eye for an eye." Once the severity of punishment is generally expected to be proportional, it becomes predictable, and the transaction costs of bargaining fall because the parties can bargain over a known "commodity." If the offender is willing to pay compensation that at least offsets the value the victim places on revenge, then violence can be avoided. In fact, in many customary-law systems, offenders reestablish membership in the community when appropriate payments are made. So-called blood money becomes increasingly prevalent.

History and anthropology suggest that customary restitution rules may be quite simple or very complex. For example, the *Hebrew Bible* dictates that "When anyone, man or woman, wrongs another. . . , that person has incurred guilt which demands reparation. He shall confess the sin he has committed, make restitution in full with the addition of one fifth, and give it to the man to whom compensation is due" (*The New English Bible*, Numbers 5: 6–7). Some customary legal systems include widely recognized rules detailing

payments for virtually every type of predictable harm (e.g., Goldsmidt 1951; Benson 1991, 2011; Barton 1967). In some societies, including medieval Iceland (Friedman 1979), the payment also depends in part on whether the offender tries to hide or deny the offense. If the offender admits guilt, thereby lowering the costs of enforcement, the payment is lower. Repeat offenders are also treated differently in many restitution-based systems. In Anglo-Saxon England, for instance, an offender could "buy back the peace" on a first offense, but for some kinds of illegal acts, a second offense was not be forgiven. Such an offender becomes an outlaw with no protection, making him fair game for anyone who wanted to attack him. Restitution-based systems also account for the problem of collections from potentially "judgment-proof" offenders. Payments do not necessarily have to be monetary, for instance, as labor services or other "goods" can serve as restitution. In Anglo-Saxon England offenders had up to a year to pay large awards, and if more time was required, they become "indentured servants" until the debt was worked off (Benson 2011, p. 25; 1994).

Bargaining power may differ between individuals, leading to variance in restitution for similar harms, so not surprisingly reciprocal mutual support groups typically develop in customary-law communities. These groups may consist of family members, for instance, but they also may be based on neighborhoods, religious affiliation, ethnicity, participation in the same commercial activity, or some other factor. Such groups accept reciprocal obligations to assist each other in the pursuit of justice, although they generally have many other reciprocal expectations, including social, religious, and joint production activities. Mutual support groups also may pay the restitution to a victim for someone in the group who cannot pay immediately, so the offender is obliged to pay members of his own support group. When such groups develop surety obligations they also may purchase indentured-servitude contracts so offenders work for members of their own groups rather than for their victims.

Clearly, hardships imposed on wealthy offenders who pay a fixed restitution are less

significant than the same payment for poor offenders. Thus, restitution may require a relatively large payment by a wealthy offender, enhancing their deterrence incentives. On the other hand, if a person is diverted from earning income to pursue and prosecute an offender, the value of lost time will be much higher for an individual earning a high income than for a low-income person. Therefore, restitution payments for the same offenses may be higher for the well-to-do victim in order to induce participation in the legal process. The schedule of payments in the *wergeld* or “man price” systems in Anglo-Saxon England reflects the status of the parties involved, as do some tribal customary-law systems (Barton 1967).

Negotiation between a victim and offender may be very difficult, if not impossible, of course, without additional options. As a consequence, a customary rule often develops which requires guilty parties to express remorse or repentance, thereby reducing the costs of negotiation. Substitutes for negotiation are also attractive. A community leader or group of leaders may come forward to encourage repentance and a truce so that negotiations can occur (Benson and Siddiqui 2014). Given the potential animosity between the parties, this third party might also mediate or arbitrate the dispute, or specialists in mediation or arbitration may be called upon. The primacy of rights also changes over time, as a right to restitution supersedes the right to retaliation. In early medieval England and Iceland, and in the large number of tribal societies, victims do not have the right to exact physical punish (retributive morality) unless and until the offender refuses to accept the customary adjudication procedures and/or to pay fair restitution (Benson 2011, pp. 11–30, 1991; Friedman 1979; Pospisil 1971; Goldsmidt 1951; Barton 1967). Similarly, ostracism occurs only when an offender refuses to adjudicate or to pay restitution.

Mutual Insurance

In part, to encourage people to continue to recognize customary behavioral rules, customary-law

communities develop mutual insurance arrangements to aid individuals who find themselves in significant risk as a consequence of mistakes, unanticipated natural disasters, warfare, theft, or general bad luck. Johnsen’s (1986) analysis of the potlatch system of the Southern Kwakiutl Indians provides an insightful example. He explains that “In order to provide the incentives of would-be encroachers to recognize exclusive property rights, and thus to prevent violence, those Kwakiutl kinship groups whose fishing seasons were relatively successful transferred wealth through the potlatch system to those groups whose seasons were not successful.... Although potlatching thereby served as a form of insurance, the relevant constraint in its adoption and survival was the cost of enforcing exclusive property rights rather than simple risk aversion” (Johnsen 1986, p. 42). Sharing norms such as hospitality and potlatching are common practices in customary-law societies all over the world (Ridley 1996, pp. 114–124). Therefore, even those who may find themselves in desperate situations know that recovery is possible, so they have relatively strong incentives to live up to customary obligations.

Leadership

Customary-law communities do not have executives with coercive power to induce recognition of law. Leaders arise to facilitate various kinds of cooperation, but they lead through persuasion and example. Leadership also is conditional, with no specified terms or binding claims to loyalty. Leaders generally serve as a nexus of the voluntary relationships that dominated the internal life of a community. Therefore, a very important leadership characteristic is a reputation for making good decisions (wisdom) that benefit his followers. Anyone who acquires a leadership role is likely to be a mature, skilled individual with considerable physical ability and intellectual experience and perhaps, more importantly, someone who has a history of cooperative behavior. The importance of maturity and experience often mean that leaders are relatively old community members (elders). They earn respect (Benson

1991; Pospisil 1971), but they also must continue to earn it. Individuals who achieve leadership positions but then make decisions that turn out to be undesirable to followers lose those followers. Wisdom and respect also are not sufficient to attract substantial numbers of followers (Pospisil 1971, p. 67). A self-interested entrepreneurial leader within a close-knit community rationally chooses to pursue activities that benefited others and/or generously spreads the wealth he gains. “The way in which [leadership or social] capital is acquired and how it is used make a great difference; the [members of the community] . . . favor rich candidates who are generous and honest” (Pospisil 1971, p. 67). Indeed, in customary-law societies, the honor of being recognized as a leader is often “purchased” through repeated public displays of generosity demonstrated at occasions such as marriages. Third-party mediation or arbitration often develops in close-knit groups, as noted above, and individuals seeking recognition as leaders also may offer to help resolve disputes and provide mediation or arbitration advice free of charge (Pospisil 1971; Benson 1991). “Fair” nonviolent dispute resolution is attractive to community members because it avoids the spillover costs associated with violent dispute resolution, and it is attractive to the “suppliers” of such dispute resolution even when they are not explicitly paid, because it enhances their prestige. In other words, both generous gift and advice giving (e.g., dispute resolution) signal that the individual is wise, successful, cooperative, and trustworthy. As Ridley (1996, p. 138) puts it, such acts “scream out ‘I am an altruist; trust me.’” Leaders engage in such displays of generosity because they expect to benefit in the future through reciprocal obligations and cooperation in joint production. Since leadership positions are available to anyone who can persuade a group to accept his decisions and guidance, customary communities often have multiple leaders. Competition for followers arises, and community members may change their primary allegiances if they feel that a leader has failed to perform well or that another individual offers what appear to be superior options. Specialization among leaders also is not uncommon. A community may have one or

more individuals serving as arbitrators or mediators, and/or as religious, hunt, or war leaders, and so on.

Warriors

Joint production of mutual defense against enemies evolves if outside threats are perceived, and such threats often mean an important part of an individual’s belief system will be “a concept of them and us” (e.g., tribalism, patriotism). In fact, an external enemy can strengthen incentives for intragroup cooperation (Ridley 1996, p. 174). There is no centralized authority in a customary-law system, however, so individuals cannot be forced to become soldiers. They have to be persuaded to take on this role. This explains the incentives to propagate cultural beliefs (rules) about honor from bravery and skill in warfare in primitive and medieval communities. Members of customary-law communities facing outside threats (e.g., spouses or potential spouses of warriors, fathers and mothers of potential warriors, elders who can no longer fight) have strong incentives to encourage young males to be brave and to be skilled in combat (Benson and Siddiqui 2014). Thus, the successful warrior is honored, but he also receives many personal rewards. A warrior who backs down from a threat or who does not pursue retribution will be seen as a coward, and this generally will be the greatest fear instilled in boys and young men. The prestige associated with abilities in warfare also encourages entrepreneurial war leaders to organize attacks on enemies. After all, given the belief that some other group is made up of enemies, aggression can easily be rationalized – “the best defense is a good offense” – particularly when the expected gains exceed the expected costs.

The discussion of conflict suggests a question: why warfare rather than cooperation (negotiation, diplomacy) with outsiders. In fact, as explained below, cooperation between members of different communities is also widely observed in customary-law systems. Given the high transaction costs for multiple-community collective action, however, the benefits of

intergroup cooperation have to be high for stable cooperation to develop (e.g., high enough for warriors to refrain from aggressive acts even when very attractive opportunities to attack arise). One potentially large benefit from intergroup cooperation arises when two or more groups face the same relatively powerful enemy. In this context, however, it should be noted that terms like “alliance,” which suggest established protocol, permanence, and formality, do not describe military relationships between such communities. These cooperative activities generally are expedient combinations in which distinct and autonomous individuals and groups work toward common but limited aims. Indeed, military coalitions generally are temporary spontaneous arrangements, and they can fall apart when the common threat loses power and/or one of the groups in the coalition gain power relative to others (Benson 2006).

Conflicts between customary-law communities need not be violent. When two different commercial communities have disputes or try to capture each other’s markets, they generally do not go to war with one another. Their warriors may be lawyers who pursue objectives through adversarial adjudication processes that arise, for instance, or lobbyists who compete for political favors in an authoritarian legal system, as discussed below. Conflict is certainly not inevitable, however, if benefits of intercommunity cooperation are significant.

Polycentric Law

Customary rules and procedures often facilitate voluntary cooperative interactions such as trade between members of different customary-law communities (Benson 1988). Groups need not formally “merge” and accept a common set of rules, however, in order to achieve intercommunity cooperation on some dimensions. Individuals only have to expect each other to recognize specific rules pertaining to the types of intergroup interactions that evolve. Indeed, a “jurisdictional hierarchy” may arise wherein each group has its own rules and procedures for intra-community

relationships, while a separate more narrowly focused set of behavioral and/or procedural rules applying for intercommunity relations (Pospisil 1971; Benson 1988). For instance, intra-community recognition of rules is likely to be largely based on reciprocities, trust, and reputation, along with ostracism threats (prudent morality), while intergroup recognition may require bonding, strong surety commitments, and/or threats of retribution (retributive morality). A jurisdictional hierarchy also does not create a higher order of law, as intergroup institutions typically have no role in any community’s internal relationships. Thus, customary law is “polycentric,” with multiple parallel “local” jurisdictions, as well as overlapping jurisdictions supporting intercommunity interactions. Many intragroup rules will be common across different groups, of course, as individuals in different groups discover similar ways to deal with an issue. Emulation also will occur where differences initially exist but individuals perceive superior arrangements among other groups (Benson 1988, 1989).

Custom Versus Authority

Pospisil (1971) distinguishes between “legal” arrangements that evolve from the top down through command and coercion, which he called “authoritarian law,” and customary-law systems that evolve from the bottom up through voluntary interaction. Similarly, Hayek (1973, p. 82) distinguishes between “purpose-independent rules of conduct” that evolve from the bottom up and rules that are designed for a purpose and imposed by “rulers.” Individuals may be members of different specialized customary-law communities (e.g., diamond merchants (Bernstein 1992) or trade associations (Benson 1995), a neighborhood (Ellickson 1991), and so-called informal or underground communities (de Soto 1989)) while simultaneously being subjected to authoritarian law. Both Hayek’s and Pospisil’s distinctions suggest that customary law also may conflict with authoritarian law. Indeed, as Hayek (1973, p. 82) stresses, “the growth of the purpose-independent

rules of conduct which can produce a spontaneous order will . . . often have taken place in conflict with the aims of the rulers who tended to turn their domain into an organization proper.” When this occurs, an authority backed by coercive power may attempt to assert jurisdiction over a - customary-law community, but this will have very different impacts, depending on the options available to members of the community. If the authority is strong enough, a community might be forced to accept the commands, although new customary rules and procedures often evolve that allow a customary-law community to avoid some and perhaps most authoritarian supervision (Bernstein 1992; Benson 1995). This may involve moving “underground,” making the customary rules and procedures difficult to observe, and raising the cost of enforcement for the authority. If a customary-law community (and its wealth) is geographically mobile, however, members may simply move outside an authority’s jurisdiction. The threat to move can significantly constrain authoritarian attempts to displace a customary system (Benson 1989). A sovereign who wants to avoid an exodus may even offer to assist in enforcing the customary rules, and even explicitly codify them. Indeed, a sovereign may offer special privileges to members of a highly mobile customary community in order to capture benefits including revenues, perhaps directly from tribute or taxes, but also indirectly because of a positive impact that this community has on other less-mobile sources of wealth (e.g., land) that can more easily be controlled and taxed. If customary behavioral rules are absorbed and customary procedures atrophy, an authority may amend or replace many customary rules, although the ability to do so depends on the costs of reinvigorating customary institutions, the mobility of wealth for members of the community, and the privileges granted to the community’s members.

Cross-References

► [Lex Mercatoria](#)

References

- Barton RF (1967) Procedure among the Ifugoa. In: Bohannon P (ed) *Law and warfare. Natural History Press, Garden City*, pp 161–181
- Benson BL (1988) Legal evolution in primitive societies. *J Inst Theor Econ* 144:772–788
- Benson BL (1989) The spontaneous evolution of commercial law. *South Econ J* 55:644–661
- Benson BL (1991) An evolutionary contractarian view of primitive law: the institutions and incentives arising under customary American Indian law. *Rev Austrian Econ* 5:65–89
- Benson BL (1994) Are public goods really common pools: considerations of the evolution of policing and highways in England. *Econ Inquiry* 32:249–271
- Benson BL (1995) An exploration of the impact of modern arbitration statutes on the development of arbitration in the United States. *J Law Econ Organ* 11: 479–501
- Benson BL (2006) Property rights and the buffalo economy of the Great Plain. In: Anderson T, Benson BL, Flannagan T (eds) *Self determination: the other path for Native Americans*. Stanford University Press, Palo Alto, pp 29–67
- Benson BL (2011) *The enterprise of law: justice without the state*, 2nd edn. Independent Institute, Oakland
- Benson BL (2014) Customary commercial law, credibility, contracting, and credit in the High Middle Ages. In Boettke P, Zywicki T (eds) *Austrian law and economics*. Edward Elgar, London (forthcoming)
- Benson BL, Siddiqui ZR (2014) Pashtunwali – law of the lawless and defense of the stateless. *Int Rev Law Econ* 37:108–120
- Bernstein L (1992) Opting out of the legal system: extra-legal contractual relations in the diamond industry. *J Leg Stud* 21:115–158
- de Soto H (1989) *The other path: the invisible evolution in the third world*. Perennial Library, New York
- Ellickson RC (1991) *Order without law: how neighbors settle disputes*. Harvard University Press, Cambridge, MA
- Friedman D (1979) Private creation and enforcement of law: a historical case. *J Leg Stud* 8:399–415
- Fuller L (1981) *The principles of social order*. Duke University Press, Durham
- Goldsmith W (1951) Ethics and the structure of society: an ethnological contribution to the sociology of knowledge. *Am Anthropol* 53:506–524
- Hayek FA (1973) *Law, legislation, and liberty*, volume I: Rules and order. University of Chicago Press, Chicago
- Johnsen DB (1986) The formation and protection of property rights among the Southern Kwakiutl Indians. *J Leg Stud* 15:41–68
- Kreps DM (1990) *Game theory and economic modeling*. Oxford University Press, Oxford, UK

North DC (1990) *Institutions, institutional change and economic performance*. Cambridge University Press, Cambridge, UK

Parisi F (2001) The genesis of liability in ancient law. *Am Law Econ Rev* 3:82–124

Pospisil L (1971) *Anthropology of law: A comparative theory*. Harper and Row, New York.

Ridley M (1996) *The origins of virtue: human instincts and the evolution of cooperation*. Viking Penguin, New York

Vanberg VJ, Buchanan JM (1990) Rational choice and moral order. In: Nichols JH Jr, Wright C (eds) *From*

political economy to economics and back. Institute for Contemporary Studies, San Francisco, pp 175–236

Vanberg VJ, Congleton RD (1992) Rationality, morality and exit. *Am Political Sci Rev* 86:418–431

Customary Law of Merchants

► [Lex Mercatoria](#)

D

Data Privacy

► [Privacy](#)

De Jure/De Facto Institutions

Jacek Lewkowicz and Katarzyna Metelska-Szaniawska
Faculty of Economic Sciences, University of
Warsaw, Warsaw, Poland

Abstract

Recent works in Law and Economics distinguish between the so-called *de jure* and *de facto* institutions. We define these two types of institutions, as well as indicate their place in the broad institutional system, in particular relative to the formal/informal and external/internal distinctions applied in (new) institutional economics. We also mention the possible interrelationships between *de facto* and *de jure* institutions, linking them to economic outcomes, and provide examples of *de jure/de facto* analyses in Law and Economics. Finally, we reflect on controversies and lacunas in the literature and present an outlook for future research.

Definition

Recent works in law and economics distinguish between the so-called *de jure* and *de facto* institutions. *De jure* means a state of affairs that is in accordance with the law, i.e., *de jure* institutions constitute a subclass of formal institutions and must necessarily be external in nature. *De facto* means a state of affairs that is true in fact but does not have to be officially sanctioned, i.e., *de facto* institutions may be formal or informal, as well as external or internal, provided that they are operative. *De facto* and *de jure* institutions are not antonyms; they may overlap.

Introduction

Recent works in law and economics increasingly emphasize the distinction between *de jure* and *de facto* institutions, e.g., in relation to constitutional rights and freedoms (including property rights), judicial independence, central bank independence, or the independence of regulatory agencies (e.g., Law and Versteeg 2013; Melton and Ginsburg 2014; Voigt et al. 2015; Hanretty and Koop 2013). Similarly, studies in political economy use the *de jure/de facto* distinction in reference to political power and its role for economic growth and development (e.g., Acemoglu and

Robinson 2006a, b). Such a distinction is not as common in (new) institutional economics, a major background for law and economics, where a broad body of literature exists on formal and informal institutions. Research in this field is, however, also concerned with enforcement mechanisms and recently pays particular attention to the distinct measurement of de jure and de facto institutions (e.g., Voigt 2013; Shirley 2013; Robinson 2013). This entry presents the conceptualization of de jure and de facto institutions and discusses their relevance for law and economics.

De Jure and De Facto Institutions Versus Other Classifications of Institutions

Most generally, institutions are perceived in the literature as systems of established social rules that structure social interactions (Hodgson 2006). They constrain behavior and are permanent or stable (Glaeser et al. 2004). In the words of D. C. North, they are certain “rules of the game,” i.e., “humanly devised constraints that shape interaction” (North 1990, p. 3), encompassing both formal and informal systems, as well as, importantly, enforcement mechanisms. Voigt (2013) is particularly clear in emphasizing the difference between rules per se and their enforcement. According to this approach, institutions are “commonly known rules used to structure recurrent interaction situations that are endowed with a sanctioning mechanism” (Voigt 2013, p. 5).

Several classifications of institutions exist in economics, the most popular one distinguishing between formal and informal institutions. Formal institutions are laws (including constitutions), policies, regulations, rights, etc. that are enforceable by official authorities (i.e., with respect to them, there exists an official sanctioning mechanism). Informal institutions are social norms, traditions, and customs that may also shape social behavior even though they are not enforced by any official authority (Berman 2013) but by means of, e.g., social control or self-enforcement. As it is unclear how formalized a rule needs to be to qualify as a formal one, in response to this

important caveat of the formal/informal distinction, Voigt (2013) proposed another classification of institutions, i.e., internal and external institutions distinguished based on the underlying enforcement mechanism. According to this approach, when sanctioning is privately organized (i.e., by members of the group or society within which a given institution functions), the institution is internal, and when sanctioning is public, it is classified as external.

As mentioned earlier, recently a new classification of de jure/de facto institutions has emerged and becomes increasingly popular in economics and other social sciences (see, e.g., Voigt 2013; Lewkowicz and Metelska-Szaniawska 2016). De jure stands for a state of affairs that is in accordance with the law. Classical works define the law as a “rule laid down for the guidance of an intelligent being by an intelligent being having power over him” (Austin 1885, p. 86) or a “rule of conduct, prescribed by the supreme power in a state, commanding what is right and prohibiting what is wrong” (Blackstone 1979, p. 44). Being a type of norms, legal norms are “generally accepted, sanctioned prescriptions for, or prohibitions against, others’ behavior, belief, or feeling, i.e. what others ought to do, believe, feel – or else. . .” (Morris 1956, p. 610) and always include sanctions. De jure institutions are, therefore, formal and external institutions. However, as the broadest approach to institutions adopted here also encompasses formal rules governing the functioning of organizations as well as formal policies which, as such, need not be rooted in the legal system (i.e., formal policy documents may exist that are sources of constraints shaping interaction but lack the status of law), de jure institutions are a subclass of formal institutions. While the set of de jure institutions covers the entire set of external institutions (i.e., a de jure institution must necessarily be external), it may also be that a given (de jure) institution has both an external and an internal nature (i.e., the same institution is sanctioned by the state, as well as by a social mechanism). De facto institutions are those observed in actual human interactions – in the market and social practice. While fulfilling the condition of being actually operative (effective),

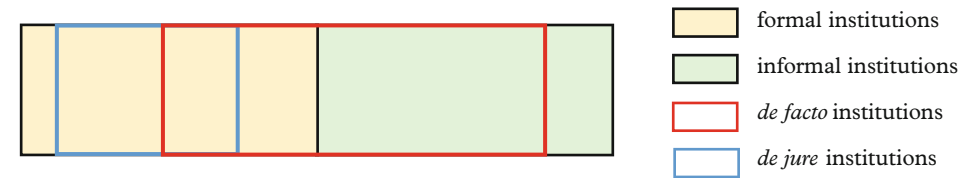
de facto institutions may be of varying nature – formal or informal. The enforcement mechanism behind the factual operation of these institutions may be both private and public, i.e., these institutions may be both of an internal and an external type.

De jure and de facto institutions are clearly not antonyms. Figure 1 presents the different classifications of institutions. Part (a) focuses on formal, informal, de jure, and de facto institutions. The sets of formal and informal institutions are disjoint, and together they form the complete set of existing institutions. As argued earlier, de jure institutions constitute a subclass of formal institutions. De facto institutions, in turn, may be either formal (de jure) or informal, provided that they are operative. A subclass of de jure institutions that

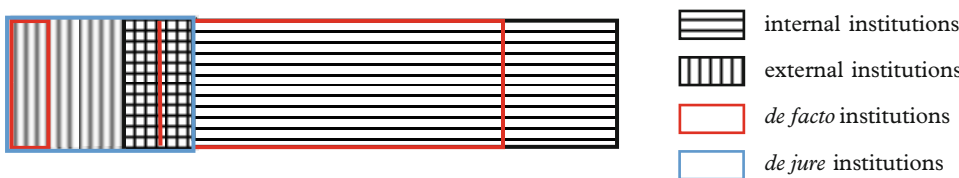
are perfectly enforced will simultaneously constitute de facto institutions. In effect, the de jure/de facto distinction produces sets with an overlap which do not cover the entire spectrum of institutions, i.e., there exist both formal and informal institutions which are neither de jure nor de facto, such as unenforced policies based on documents which are not law (formal) or normative beliefs when conceived as social norms (informal).

Part (b) of Fig. 1 presents the sets of external, internal, de jure, and de facto institutions. The set of de jure institutions is identical to the set of external institutions, including the latter’s intersection with the set of internal institutions. De facto institutions cover part of the de jure institutions set (including institutions of solely external

(a) sets of formal, informal , *de jure* and *de facto* institutions



(b) sets of external, internal, *de jure* and *de facto* institutions



(c) sets of formal, informal , external and internal institutions



De Jure/De Facto Institutions, Fig. 1 Classifications of institutions. (a) Sets of formal, informal, de jure, and de facto institutions. (b) Sets of external, internal, de jure, and de facto

institutions. (c) Sets of formal, informal, external, and internal institutions (Source: own elaboration)

nature, as well as those being at the same time external and internal) and part of the internal institutions set. While some *de jure* (external institutions) are neither *de facto* nor internal ones, we can also identify institutions that qualify as external, *de jure* and *de facto*, or even external, internal, *de jure* and *de facto*, at the same time.

Finally, part (c) of Fig. 1 demonstrates the relative positions of the sets of formal, informal, internal, and external institutions. While the sets of formal and external institutions are nearly identical, this is not the case for informal and internal institutions. The latter set intersects with the set of formal institutions covering those institutions which, as argued earlier, have both an external and an internal nature. Furthermore, as internal institutions include formal private rules, the set of internal institutions also expands beyond the informal institutions set.

Interrelationships Between De Jure and De Facto Institutions

Institutions usually interplay with each other. De jure and de facto institutions may boost each other when they lead to commonly desired behavior or inhibit each other when this is not the case. In dynamic settings convergence and divergence may be observed, as well as a crowding out effect between de jure and de facto institutions. Economic effects of the interrelationships between these institutions result, depending on their nature, in decreasing or increasing the level of transaction costs connected with implementing and enforcing legislation. They may also affect the aims that the legislator strives to achieve by imposing new laws, as well as government's credibility vis-à-vis economic actors.

De Jure and De Facto Institutions in Law and Economics

Analysis of the interrelationships between de facto and de jure institutions is present primarily in studies confined to individual rules. Judicial independence is an area where this distinction

has been most pronounced in law and economics over the recent years, in particular since Feld and Voigt (2003) proposed distinct measures of de jure and de facto judicial independence and provided empirical evidence confirming the relevance of the de facto, not de jure, measure for economic growth (a finding later confirmed by Voigt et al. 2015). Much debate arose in the literature concerning the relationships between de jure and de facto judicial independence, with some studies finding no relationship and others – a tight correlation (see, e.g., Herron and Randazzo 2003; Hayo and Voigt 2007). Melton and Ginsburg (2014) summarized this literature and offered an explanation, which de jure protections actually enhance de facto judicial independence. The de jure/de facto distinction has also been applied in studies focusing on independence of central banks (e.g., Hayo and Voigt 2008), prosecutors (Aaken et al. 2010), and regulatory agencies (e.g., Hanretty and Koop 2013). Other examples of institutions that have been studied with regard for the de jure/de facto distinction are, inter alia, competition policy (Voigt 2009) and direct democracy mechanisms (Blume et al. 2009).

A second area of law and economics research, where the focus has been on the de jure/de facto distinction, concerns protection of rights. Some studies have focused on explaining the determinants and/or effects of the de jure de facto gap for various categories of constitutional rights and freedoms (e.g., Law and Versteeg 2013), while others concentrated on disentangling the relationships between de jure and de facto rights (e.g., Melton 2013; Chilton and Versteeg 2016). Other types of rights analyzed from the de jure/de facto perspective are property (land) rights and the related problem of natural resource management (Robbins 2000; Alston and Mueller 2007; Alston et al. 2012; Bellemare 2013).

Future Outlook

Until recently there only existed rare empirical studies of de jure/de facto institutions, confined to individual rules, with no commonly accepted definitions and no general-level analysis of

relationships between these two types of institutions. In recent years this topic raised increased interest leading to conceptualization and a more systematic analysis from an economic perspective. This may serve as a theoretical underpinning for subsequent empirical studies, focusing on other law and economics contexts, beyond independence and rights. The discussion concerning identification of de facto institutions (i.e., ascertaining when an institution can be deemed operative/effective) will also certainly intensify alongside the stimulating debate on measurement of institutions. Another significant extension could concentrate on providing a detailed dynamic account of the interrelationships between de jure and de facto institutions. Such studies are of high policy relevance as they yield recommendations on the design of more effective legal institutions.

Cross-References

- [Constitutional Political Economy](#)
- [Institutional Complementarity](#)
- [Institutional Economics](#)

References

- Acemoglu D, Robinson JA (2006a) Economic origins of dictatorship and democracy. Cambridge: Cambridge University Press
- Acemoglu D, Robinson JA (2006b) De facto political power and institutional persistence. *Am Econ Rev* 96(2):326–330
- Alston L, Mueller B (2007) Legal reserve requirements in Brazilian forests: path dependent evolution of de facto legislation. *Economia* 8(4):25–53
- Alston L, Harris E, Mueller B (2012) The development of property rights on frontiers: endowments, norms, and politics. *J Econ Hist* 72(3):741–770
- Austin J (1885) Lectures on jurisprudence; or the philosophy of positive law. London: J. Murray Publishing
- Bellemare MF (2013) The productivity impacts of formal and informal land rights: evidence from Madagascar. *Land Econ* 89(2):272–290
- Berman S (2013) Ideational theorizing in the social sciences since ‘Policy paradigms, social learning and the state’. *Governance* 26(2):217–237
- Blackstone W (1979) Commentaries on the laws of England. Chicago: The University of Chicago Press
- Blume L, Müller J, Voigt S (2009) The economic effects of direct democracy: a first global assessment. *Public Choice* 140(3):431–461
- Chilton A, Versteeg M (2016) Do constitutional rights make a difference? *Am J Polit Sci* 60(3):575–589
- Feld LP, Voigt S (2003) Economic growth and judicial independence: cross-country evidence using a new set of indicators. *Eur J Polit Econ* 19(3):497–527
- Glaeser EL, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) Do institutions cause growth? *J Econ Growth* 9(3):271–303
- Hanretty C, Koop C (2013) Shall the law set them free? The formal and actual independence of regulatory agencies. *Regul Gov* 7:195–214
- Hayo B, Voigt S (2007) Explaining de facto judicial independence. *Int Rev Law Econ* 27(3):269–290
- Hayo B, Voigt S (2008) Inflation, Central Bank Independence, and the legal system. *J Inst Theor Econ* 164(4):751–777
- Herron ES, Randazzo KA (2003) The relationship between independence and judicial review in post-communist courts. *J Polit* 65:422–438
- Hodgson GM (2006) What are institutions? *J Econ Issues* 11(1):1–25
- Law DS, Versteeg M (2013) Sham constitutions. *Calif Law Rev* 101:863–952
- Lewkowicz J, Metelska-Szaniawska K (2016) De Jure and de facto institutions – disentangling the interrelationships. *Latin Am Iber J Law Econ* 2(2), forthcoming
- Melton J (2013) Do constitutional rights matter? The relationship between de jure and de facto human rights protection, working paper
- Melton J, Ginsburg T (2014) Does De Jure judicial independence really matter. A reevaluation of explanations for judicial independence. *J Law Courts* 2:187–217
- Morris RT (1956) A typology of norms. *Am Sociol Rev* 21:610–613
- North DC (1990) Institutions, institutional change, and economic performance. New York: Cambridge University Press
- Robbins P (2000) The rotten institution: corruption in natural resource management. *Polit Geogr* 19(4), 2000:423–443
- Robinson JA (2013) Measuring institutions in the Trobriand Islands: a comment on Voigt’s paper. *J Inst Econ* 9(1):27–29
- Shirley MM (2013) Measuring institutions: how to be precise though vague. *J Inst Econ* 9(1):31–33
- van Aaken A, Feld L, Voigt S (2010) Do independent prosecutors deter political corruption? An empirical evaluation across seventy-eight countries. *Am Law Econ Rev* 12(1):204–244
- Voigt S (2009) The effects of competition policy on development: cross-country evidence using four new indicators. *J Dev Stud* 45(8):1225–1248
- Voigt S (2013) How (not) to measure institutions. *J Inst Econ* 9(1):1–26
- Voigt S, Gutmann J, Feld LP (2015) Economic growth and judicial independence, a dozen years on: cross-country evidence using an updated set of indicators. *Eur J Polit Econ* 38:197–211

Death Penalty

John J. Donohue III

Stanford Law School and National Bureau of
Economic Research, Stanford, CA, USA

Abstract

The issue of the death penalty has been an area of enormous academic and political ferment in the United States over the last 40 years, with the country flirting with abolition in the 1970s, followed by a period of renewed use of the death penalty and then a period of retrenchment, reflected in a declining number of death sentences and executions and a recent trend leading six states to abolish the death penalty in the last 6 years. Internationally, there is a steady movement away from the death penalty, which has been abolished throughout the European Union, although certain states in the Middle East (Saudi Arabia, Iran) and Asia (China, Singapore, Japan) have continued to use it frequently.

Definition

A system of punishment involving the execution of individuals convicted of a capital crime.

Introduction

Death penalty statutes vary some across the 32 states with current enforceable statutes in the United States, but capital punishment is universally reserved for a relatively narrow category of “first-degree” murders that involve one or more aggravating circumstances. These aggravating circumstances typically include murdering a police officer or witness, murder for hire, multiple murders, and murders that are “heinous, atrocious, or cruel.” Even cases that are initially treated as capital eligible are very unlikely to result in a death sentence; defendants often plead guilty in return for receiving a noncapital

sentence, some defendants are found guilty of a lesser charge in jury trials, and some juries decide not to impose the death sentence even when the law would permit them to do so (National Research Council 2012).

This entry will review some of the major social science studies evaluating the issue of whether the death penalty deters, which have largely failed to provide any convincing evidence of deterrence, as well as the major studies exploring racial bias in the administration of the death penalty.

Procedural Issues Associated with the Death Penalty

The modern death penalty era in the United States began in 1972 with *Furman v. Georgia* (408 U.S. 238 1972). In that case, the US Supreme Court was concerned that the unchanneled discretion of prosecutors, judges, and juries led to the arbitrary administration of the death penalty. As a result, the Court struck down every then-existing sentence of death, stating “that the imposition and carrying out of the death penalty in these cases constitute cruel and unusual punishment in violation of the Eighth and Fourteenth Amendments.” While the Justices apparently believed that their decision would lead to the abolition of the death penalty, it ironically propelled its strong revival.

Most states, including some that had not had the penalty before, responded to *Furman* by enacting specific death penalty statutes that were designed to address the Court’s articulated concerns, including the implementation of a pretrial capital hearing along with separate guilt and sentencing trials; the consideration of so-called aggravating and mitigating circumstances that enhance or undermine the case for execution, respectively; and the automatic appeal of death sentencing decisions. Four years later, in *Gregg v. Georgia* (428 U.S. 153, 169 1976), the Supreme Court held that “the punishment of death does not invariably violate the Constitution” and indicated that several of these new statutes were facially constitutional.

However, there is a substantial and growing body of evidence that challenges the notion that the procedural changes that were deemed facially constitutional in *Gregg* have adequately addressed the concerns of *Furman* in practice. Indeed, in October 2009, the American Law Institute (ALI) voted overwhelmingly to withdraw its death penalty framework from the Model Penal Code because that framework had proved to be woefully inadequate in application (Liebman 2009). The April 15, 2009, “Report of the Council to the Membership of the American Law Institute On the Matter of the Death Penalty” expressed concerns over the inability to avoid the arbitrary, discriminatory, or simply erroneous invocation of this irrevocable punishment. These difficulties are compounded by the need to construct a system that, on the one hand, allows for consistent sentencing outcomes but simultaneously gives the fact finder sufficient leeway to consider the circumstances surrounding each crime.

Despite concerns about the administration of the death penalty, it has been broadly popular in the United States, especially during periods such as the late 1980s and early 1990s when crime in the United States was particularly high. As a result, politicians such as New York Governor Mario Cuomo failed to win reelection in part due to their opposition to the death penalty, while other governors, such as George W. Bush of Texas, were launched into national prominence because of their strong support for the death penalty.

Deterrence and the Death Penalty

The simplest law and economics assessment would suggest that the death penalty should deter murder, as it enhances the maximum penalty associated with killings. It is now recognized that this simple theoretical analysis is inadequate. First, capital punishment is invoked rarely and after years of delay due to the post-*Furman* implementation of long and complicated legal procedures that must be exhausted before any death sentence can be administered, thereby undercutting any deterrent potential. Second, the costs of

running a death penalty system are staggering, and any attempt to estimate the effect of capital punishment on crime must also take into account the fact that implementing a capital punishment system may draw resources away from other more effective law enforcement projects. A recent study from California estimated that the state’s system for prosecuting death penalty cases cost taxpayers \$4 billion between 1976 and early 2011, even though only 13 executions were carried out over the same time period (Alercon and Mitchell 2011). Cook (2009) similarly concluded that North Carolina’s execution system costs \$11 million annually, while a regression analysis performed using Maryland case data calculated that the full array of added expenses associated with pursuing a capital prosecution rather than a non-capital trial was approximately \$1 million (Roman et al. 2009).

The former Manhattan District Attorney Robert Morgenthau spoke out eloquently about the “terrible price” inflicted by the presence of capital punishment in the hopes that his words might forestall New York’s 1995 launch of a death penalty system:

Some crimes are so depraved that execution might seem just. But even in the impossible event that a statute could be written and applied so wisely that it would reach only those cases, the price would still be too high.

It has long been argued, with statistical support, that by their brutalizing and dehumanizing effect on society, executions cause more murders than they prevent. “After every instance in which the law violates the sanctity of human life, that life is held less sacred by the community among whom the outrage is perpetrated.”¹ (Morgenthau 1995)

When the New York death penalty law was adopted over Morgenthau’s opposition, he simply refused to seek the death penalty in the borough of Manhattan, as did his fellow District Attorney in the Bronx. From the implementation of the New York death penalty law in 1995 until 2004 (when it was judicially abolished), the murder rate

¹Quote is from Robert Rantool Jr. in 1846. Titus J. (1848) Reports and Addresses Upon the Subject of Capital Punishment. New York State Society for the Abolition of Capital Punishment. pg. 48.

dropped in Manhattan by 64.4% (from 16.3 to 5.8 murders per 100,000) and in the Bronx by 63.9% (from 25.1 to 9.1 per 100,000). Another New York City borough with the same laws and police force and with broadly similar economic, social, and demographic features as Manhattan and the Bronx – Brooklyn – had a top prosecutor who issued the largest number of notices of intention to seek the death penalty, and yet Brooklyn experienced only a 43.3% decline in murders over this period (from 16.6 murders to 9.4 per 100,000 in 1995) (Kuziemko 2006; Donohue and Wolfers 2009). Strong causal inferences cannot be drawn from Manhattan and the Bronx's homicide decline, but this example illustrates the lack of an apparent capital punishment deterrent effect in the crime patterns across these counties.

While a steady stream of papers beginning with Ehrlich (1975) have tried to make the empirical case that the death penalty has been a deterrent to murder in the United States, these studies have largely been rejected by the academic community (National Research Council 1978; Donohue and Wolfers 2005, 2009; Kovandzic et al. 2009; National Research Council 2012). The mounting evidence undermining the view that the death penalty deters crime has begun to influence the US Supreme Court's discussion of the constitutionality of the death penalty, as illustrated by the debate between Justices Stevens and Scalia in *Baze v. Rees* (553 U.S. 35 2008). Justice Stevens agreed with the majority in that case that the cocktail of drugs used by Kentucky to execute prisoners did not violate the Eighth Amendment, but went on to argue that there was "no reliable statistical evidence that capital punishment in fact deters potential offenders."

Justice Scalia responded sharply to Justice Stevens's concurrence. Referencing an article by Cass Sunstein and Adrian Vermeule (Sunstein and Vermeule 2005), Scalia argued, "Justice Stevens' analysis barely acknowledges the 'significant body of recent evidence that capital punishment may well have a deterrent effect, possibly a quite powerful one.'" However, knowledgeable researchers believe that this so-called significant body of evidence was based on outdated or invalid econometric techniques and models. Indeed,

shortly after the decision was handed down, Sunstein indicated that he had changed positions in the wake of the scholarly demolition of the existing pro-deterrence literature, writing, "In short, the best reading of the accumulated data is that they do not establish a deterrent effect of the death penalty" (Sunstein and Wolfers 2008).

In April of 2012, a panel of the National Research Council (NRC) issued a report on deterrence and the death penalty, concluding that previous research was "not informative about whether capital punishment decreases, increases, or has no effect on homicide rates." The panel based this decision on two main factors. First, they noted that existing studies did not adequately model the effect of noncapital punishment on crime, which would bias estimates of the effect of capital punishment on crime if common factors influenced both the frequency of death sentences and the severity of noncapital punishment. To accurately capture the deterrent effect, researchers would need to show the deterrent effect of the death penalty in comparison to other common sanctions.

Second, the panel wrote that existing research did not adequately model "potential murderers' perceptions of and response to the capital punishment component of a sanction regime." The NRC panel noted that potential offenders cannot be deterred by the death penalty unless they are aware of the threat, and there is a large body of literature confirming that the general public is very poorly informed about the actual likelihood of the imposition of death penalty sentences. The NRC report also determined that many earlier studies used "strong and unverifiable assumptions" in their identification strategies (National Research Council 2012).

Racial Bias in the Implementation of the Death Penalty

There is an expansive empirical literature on whether race affects prosecutors' and jurors' death penalty decisions, with nearly all recent studies finding that race does influence capital sentencing outcomes. David Baldus's 1990 study

of capital sentencing in Georgia, *Equal Justice and the Death Penalty*, was a landmark study in the empirical evaluation of the impact of race on the administration of the death penalty. Similar studies conducted in states, counties, and cities across the United States confirm these findings, and a number of studies have used controlled experiments to pinpoint precisely how race affects capital outcomes. The results of these studies further corroborate the findings of the econometric literature: race inappropriately influences the administration of the death penalty even after controlling for legitimate case characteristics.

The Baldus Study on Capital Sentencing in Georgia

Baldus's Georgia study investigated the effect of race on decisions throughout the charging and sentencing process by analyzing a large, stratified random sample of 1,066 defendants selected from the universe of 2,484 defendants who were charged with homicide and subsequently convicted of murder or voluntary manslaughter in Georgia between March 28, 1973, and December 31, 1979. The researchers then weighted this sample, which included 127 defendants who had been sentenced to death, to evaluate the effect of race on capital sentencing in the case universe as a whole. The researchers reviewed each defendant's case files and collected data on the circumstances of the offense and the characteristics of the defendant. Using both linear probability and logit models, the Baldus team conducted an extensive regression analysis investigating the main effect of race on capital sentencing in Georgia. Eight models differing in their method and in their explanatory variables were presented, each of which indicated that victim race inappropriately influenced which defendants were sentenced to die and which were permitted to live.

This comprehensive analysis showed that defendants convicted of murdering a white victim were statistically significantly more likely to be sentenced to death than defendants convicted of murdering a black victim. A logistic regression model from this study showed that the odds of being sentenced to death were 4.3 times greater for a defendant convicted of murdering a white

victim than for a defendant convicted of murdering a black victim. This became the core piece of evidence regarding the race-of-victim discriminations in *McCleskey v. Kemp* (481 U.S. 279 1987). This difference was statistically significant at the 0.005 level. Victim race exerted a greater influence on capital sentencing outcomes than numerous legitimate factors, such as whether the offense was coupled with kidnapping, whether the victim was frail, or whether the victim was an on-duty law enforcement officer.

Baldus et al. also used two different approaches to evaluate whether death sentencing decisions were influenced by the *interaction* of the race of the defendant and the victim. First, the authors examined narrative summaries of cases that were death-eligible under the state's contemporaneous-felony statutory aggravating circumstance. This statutory aggravating factor served as rough proxy for death eligibility, because death penalties were imposed primarily in cases involving contemporaneous felonies. The researchers classified the 438 cases involving a contemporaneous felony into various crime subcategories and then compared the death sentencing rates for similar types of cases involving different combinations of defendant and victim race.

As shown in Table 1, controlling for the type of contemporaneous felony revealed that the race of the victim strongly influenced capital sentencing. The interaction between defendant and victim race was particularly pronounced for armed robbery cases, as a black defendant was over six times more likely to be sentenced to death if convicted of murdering a white victim than if convicted of murdering a black victim. The disparity between the death sentencing rates of cases involving a black defendant and white victim and cases involving a black defendant and a black victim is statistically significant at the 0.01 level. The disparity between these racial combinations is also significant at the 0.05 level for burglary and/or arson cases and the 0.01 level for armed robbery cases.

Second, the researchers employed regression techniques to examine how the race of the defendant and victim interacted to influence capital

Death Penalty, Table 1 Race of defendant/victim and death sentencing in Georgia by contemporaneous felony

Contemporaneous felony	% of cases with a <i>black</i> defendant and a <i>black</i> victim that result in a death sentence	% of cases with a <i>black</i> defendant and a <i>white</i> victim that result in a death sentence	% of cases with a <i>white</i> defendant and <i>black</i> victim that result in a death sentence	% of cases with a <i>white</i> defendant and a <i>white</i> victim that result in a death sentence	Ratio of probability of a death sentence for a black-on-white murder to probability of death sentence for a black-on- black murder (controlling for contemporaneous felony)
	(A)	(B)	(C)	(D)	Ratio of (B)/(A)
All death-eligible cases involving a contemporaneous-felony statutory aggravating circumstance	14.4 % (15/104)	37.5 % (60/160)	21.4 % (3/14)	32.5 % (52/160)	2.6
Armed robbery	5.3 % (3/57)	34.1 % (42/123)	27.3 % (3/11)	27.4 % (23/84)	6.49
Rape	44.4 % (8/18)	50 % (8/16)	0 % (0/1)	58.8 % (10/17)	1.13
Kidnapping	28.6 % (2/7)	60 % (3/5)	0 % (0/1)	45.0 % (9/20)	2.1
Burglary and/or arson	0 % (0/8)	62.5 % (5/8)	–	38.5 % (5/13)	Infinite
Another murder	28.6 % (2/7)	33.3 % (2/6)	–	27.8 % (5/18)	1.17
Aggravated battery	0 % (0/7)	0 % (0/2)	0 % (0/1)	0 % (0/8)	Undefined

Death Penalty, Table 2 Race of defendant/victim and death sentencing in Georgia by egregiousness categories

Predicted chance of a death sentence, from 1 (low) to 8 (high)	% sentenced to death for murders with a <i>black</i> offender and a <i>black</i> victim	% sentenced to death for murders with a <i>black</i> offender and <i>white</i> victim	% sentenced to death for murders with a <i>white</i> offender and <i>black</i> victim	% sentenced to death for murders with a <i>white</i> offender and <i>white</i> victim	Ratio of (B)/(A)
	(A)	(B)	(C)	(D)	
1	0 % (0/19)	0 % (0/9)	–	0 % (0/5)	Undefined
2	0 % (0/27)	0 % (0/8)	0 % (0/1)	0 % (0/19)	Undefined
3	11.1 % (2/18)	30.0 % (3/10)	0 % (0/9)	2.6 % (1/39)	2.7
4	0 % (0/15)	23.1 % (3/13)	–	3.4 % (1/29)	Infinite
5	16.7 % (2/12)	34.6 % (9/26)	–	20 % (4/20)	2.08
6	5.0 % (1/20)	37.5 % (3/8)	50.0 % (2/4)	15.6 % (5/32)	7.5
7	38.5 % (5/13)	64.3 % (9/14)	0 % (0/5)	38.5 % (15/39)	1.67
8	75 % (6/8)	90.9 % (20/22)	–	89.3 % (25/28)	1.21

sentencing outcomes. The research team began by conducting a multiple regression analysis that considered the (nonracial) circumstances of each case to produce an estimate of the probability that it would result in a death sentence. They then used the results of this regression analysis to construct an eight-point egregiousness scale based on the

estimated probability that each case would result in a death sentence. Finally, the research team placed the 472 most egregious cases of the total sample into an eight-level egregiousness scale and compared the racial characteristics of actual sentencing rates within each level. The results of this regression-based analysis are provided in Table 2.

Table 2 shows that controlling for egregiousness, cases involving black defendants and white victims were substantially more likely to result in a death sentence than cases involving other combinations of defendant and victim race. Other than at the two lowest levels of the egregiousness scale (where no death sentences were imposed), a black defendant convicted of murdering a white victim was substantially more likely at each egregiousness level to be sentenced to death than either a black defendant convicted of murdering a black victim or a white defendant murdering a white victim. (As the authors noted, the racial disparities shrink at the highest egregiousness level – level 8 – since most defendants received the death penalty.)

Subsequent Studies of Race and Capital Sentencing

The Baldus team's regression models uniformly demonstrate that race infected the administration of capital punishment in Georgia during the study's sample period. Well-controlled studies using more recent data from jurisdictions across the country have similarly found that the race of the victim influences who is sentenced to die. This finding is consistent across studies and permeates both the pre- and post-1990 literature. An overview of the pre-1990 literature on the role of race in post-*Furman* capital sentencing was captured in a 1990 report of the US General Accounting Office (GAO), which issued a clear assessment of a set of studies conducted by 21 sets of researchers and based on 23 distinct datasets: "Our synthesis of the 28 studies shows a pattern of evidence indicating racial disparities in the charging, sentencing, and imposition of the death penalty after the *Furman* decision." The report also concluded:

In 82 percent of the studies, race of victim was found to influence the likelihood of being charged with capital murder or receiving the death penalty, i.e., those who murdered whites were found to be more likely to be sentenced to death than those who murdered blacks. This finding was remarkably consistent across data sets, states, data collection methods, and analytic techniques. The finding held for high, medium, and low quality studies. . . [Our] synthesis supports a strong race of victim influence.

The GAO noted that "The race of victim influence was found at all stages of the criminal justice system process, although there were variations among studies as to whether there was a race of victim influence at specific stages."

Findings that race influences the administration of capital punishment are similarly robust in the post-1990 literature. Table 3 presents the regression results of ten methodologically rigorous recent studies on the effect of victim race on capital sentencing outcomes, which have found that defendants convicted of murdering a white victim are significantly more likely to be sentenced to death than similarly situated defendants convicted of murdering a black victim (Donohue 2013). The relative probabilities presented in this table were estimated by regression models that controlled for variables that are expected to affect capital sentencing decisions.

In addition, studies that have examined the interaction between defendant and victim race have generally confirmed that black defendants are remarkably more likely to be sentenced to death if their victim is white rather than black. Table 4 displays these unadjusted rates of racial disparity. For example, in the Baldus et al. Georgia study, black defendants were 17.2 times more likely to be sentenced to death if the victim was white rather than black. The figures in this table are just overall percentages, not regression-adjusted estimates, but their uniformity is revealing.

National-Level Studies Examining Race and Capital Sentencing

In a sophisticated national-level study including 99.4% of persons admitted to death row in the United States between 1977 and 1999, researchers Blume et al. (2004) analyzed data on murders and the composition of death row from the 31 states that admitted ten or more defendants to death row during this time period. The researchers obtained data on the characteristics of murders, the racial composition of death row, and the legal and political characteristics of different states. They then compared the overall population of murderers to the death row population to determine which factors are related to the probability of being

Death Penalty, Table 3 Regression analyses on the race-of-victim effect and capital sentencing

Location	Period of study	Cases analyzed	Relative probability of being sentenced to death for killing a white victim rather than a black victim (controlling for other relevant variables)	Statistical significance
<i>Panel A: states</i>				
California	1990–1999	Reported homicides	2.46	<0.001
Georgia	1973–1979	Defendants charged with homicide and subsequently convicted of murder or voluntary manslaughter	4.3	<0.005
Florida	1976–1987	Homicides	3.42	<0.001
Illinois	1988–1997	Defendants convicted of first-degree murder	2.48	<0.01
Maryland	1978–1999	Death-eligible first- or second-degree murder cases	3.7	
Missouri	1977–1991	Nonnegligent homicides	2.61	<0.10
North Carolina	1980–2007	Homicides	2.96	<0.001
Ohio	1981–1994	Homicide	1.66	<0.01
<i>Panel B: counties</i>				
East Baton Rouge, LA	1990–2008	Defendants convicted of homicide	37.04	<0.005
Harris County, TX	1992–1999	Defendants indicted for capital murder	1.63	n/a

convicted of capital murder and placed on death row.

The researchers found that variation in black representation on states' death rows across the country can be largely predicted by three variables: (1) the overall proportion of murders committed by blacks, (2) the proportion of all murders involving a black offender and a white victim, and (3) whether a state is a former confederate state (where the large proportion of murders involve black defendants and victims). The finding that black-on-white murders were treated more harshly than other types of murders was statistically significant at the 0.01 level. Variables such as whether a judge imposes the final sentence, the amount of political pressure on judges, and state Supreme Court Justices' political ideology were not related to the proportion of blacks on death row.

Blume et al. (2004) also calculated the rate at which murder cases involving different

combinations of defendant and victim race resulted in death sentences for the eight states for which they had complete data for the period from 1977 to 2000. Table 5 displays this data and shows that cases involving a black offender and a white victim are far more likely to result in the offender being placed on death row than cases involving other combinations of offender and victim race. The combination of a black offender and a white victim leads to a death sentence roughly 3–23 times more frequently than the rate associated with black offender-black victim cases.

A New Test of Racial Bias in Capital Sentencing

In a recent working paper, Alberto F. Alesina and Eliana La Ferrara propose a novel test of racial bias in capital sentencing based on whether reversals of death sentences vary depending on the race of the defendant and victim. The authors model the behavior of the trial court as minimizing the weighted sum of the probability of sentencing an

Death Penalty, Table 4 Unadjusted rates of death sentencing in various states and counties by race of defendant/victim

Location	Period of study	Type of case	% of cases with a <i>black</i> defendant and a <i>black</i> victim that result in a death sentence (A)	% of cases with a <i>black</i> defendant and a <i>white</i> victim that result in a death sentence (B)	% of cases with a <i>white</i> defendant and a <i>black</i> victim that result in a death sentence (C)	% of cases with a <i>white</i> defendant and a <i>white</i> victim that result in a death sentence (D)	Ratio of (B)/(A)
Panel A: states							
California	1990–1999	Reported homicides	0.7 % (36/5,355)	3.5 % (34/984)	0 % (0/244)	1.9 % (79/4,206)	5.14
Florida	1976–1987	Homicides	0.8 % (36/4,428)	12.6 % (92/731)	3.4 % (9/264)	4.9 % (227/4,645)	15.48
Georgia	1973–1979	Defendants charged with homicide and subsequently convicted	1.2 % (18/1,443)	21.5 % (50/233)	3 % (2/60)	7.8 % (58/748)	17.2
Illinois	1988–1997	Defendants convicted of first-degree murder	1.1 % (27/2,526)	4.7 % (17/363)	4.8 % (3/59)	4.8 % (23/458)	4.38
Maryland^a	1978–1999	Death-eligible first- or second-degree murder	2.30 %	13.80 %	4.60 %	8.90 %	6
Missouri	1977–1991	Nonnegligent homicides	1.2 % (24/2,033)	7.1 % (17/239)	3.3 % (3/90)	3.9 % (58/1,488)	6.03
Nebraska	1973–1999	Death-eligible homicides	8.7 % (2/23)	18.2 % (4/22)	20.0 % (1/5)	21.0 % (13/62)	2.09
Ohio	1981–1994	Homicides	2.3 % (77/3,337)	10.8 % (56/517)	4.3 % (8/184)	5.5 % (130/2,385)	4.69
Panel B: counties							
East Baton Rouge, LA	1990–2008	Defendants convicted of homicide	8.3 % (11/132)	30 % (9/30)	0 % (0/3)	12 % (3/25)	3.6

^aRaw numbers not available for Maryland

innocent defendant to death and that of letting a guilty defendant free (the inevitable trade-off between type I and type II error). The authors suggest that racial bias exists when the relative weight of these two types of errors is a function of defendant and/or victim race. Thus, if decision makers throughout the criminal justice system consider minority on white crimes to be more serious, the relative weighting of the burdens of type I and type II error might shift in favor of a greater likelihood of erroneous conviction for defendants accused of these crimes. Under the

assumption that higher courts are less likely to be affected by racial bias, one can predict that the combination of defendant and victim race will be only correlated with reversals if lower courts are affected by racial bias. The authors test this prediction by looking nationwide at all capital appeals that became final between 1973 and 1995 and by gathering information on the race of the defendant and victim(s) in these cases. They find robust evidence of bias in minority on white murders: in direct appeal and habeas corpus cases, the probability of error is 3 and 9 percentage

Death Penalty, Table 5 Capital sentencing rates by race of defendant and victim in eight states (1977–2000)

	% sentenced to death for murders with a <i>black</i> offender and a <i>black</i> victim	% sentenced to death for murders with a <i>black</i> offender and <i>white</i> victim	% sentenced to death for murders with a <i>white</i> offender and <i>black</i> victim	% sentenced to death for murders with a <i>white</i> offender and <i>white</i> victim	Ratio of (B)/(A)
	(A)	(B)	(C)	(D)	(A)
<i>State</i>					
Georgia	0.5 (35/7,091)	9.9 (72/726)	2.1 (4/187)	4.2 (114/2,734)	20.1
Indiana	0.6 (12/2,151)	4.2 (16/375)	0.0 (0/100)	2.2 (49/2,272)	7.6
Maryland	0.2 (10/4,174)	5.2 (25/479)	0.7 (1/137)	1.4 (20/1,429)	21.8
Nevada	2.5 (11/442)	10.1 (18/178)	1.3 (1/80)	3.7 (46/1,244)	4.1
Pennsylvania	1.8 (112/6,310)	4.9 (46/947)	1.2 (4/335)	2.2 (90/4,055)	2.7
South Carolina	0.3 (14/4,784)	6.8 (50/738)	5.0 (9/179)	2.7 (72/2,654)	23.2
Virginia	0.4 (18/4,975)	6.5 (46/713)	2.3 (5/217)	1.8 (58/3,167)	17.8
Arizona^a	0.5 (13/2,416)	4.8 (19/400)	2.8 (7/247)	5.9 (95/1,613)	8.8

^aNote: the data for Arizona combines blacks and Hispanics into a single “minority” category. Thus, the numbers in the last row of the table for Arizona black offender and black victim also include Hispanic offenders and victims.

points higher for minority on white murders, respectively, than for minority on minority murders.

Controlled Experiments and Social Science Evidence on the Pathways of Racially Biased Decision Making in Capital Sentencing

Some social science research has tried to illuminate the mechanisms leading to racially biased capital sentencing decisions. For example, a study by Mona Lynch and Craig Haney (2009) investigated how the process of juror deliberation can generate racially biased death penalty sentences. In their study, Lynch and Haney recruited 539 mock jurors to participate in video-simulated death penalty trials. Each juror viewed identical videos of the case, varying only the race – through both appearance and voice – of the defendant and victim. As part of their data collection, Lynch and Haney quantified “verdict certainty” by asking mock jurors to assess, both before and after deliberation, with what level of certainty they felt that the defendant deserved the death penalty for the particular homicide committed. Lynch and Haney found that after collective deliberation, not only did all jurors favor the death penalty more frequently, but the tendency to sentence black defendants to death more often than white defendants was exacerbated among white

jurors and jurors with poor instruction comprehension. Additionally, white jurors felt more certain that the nature of the homicide merited the death penalty when the defendant was black rather than white. The 2009 Lynch and Haney study also shed important light on how capital jurors evaluate mitigating and aggravating factors. Lynch and Haney found that white male jurors were less likely to consider mitigating evidence for black defendants; there was no comparable effect for women and nonwhite jurors when treated as a separate group, but the “white male dominance” of deliberation sessions nonetheless led to biased sentencing outcomes. Similarly, they conclusively found that jurors gave less weight to two categories of mitigating factors – namely, psychiatric problems and substance abuse issues – when the victim was white than when the victim was black.

Their 2009 findings accord with those of their previous study (Lynch and Haney 2000) that investigated whether juror comprehension of the judge’s instructions was a factor in sentencing bias. The authors again created a video-simulated trial that altered only the race of the victim and defendant for a “midrange” robbery-murder case. They recruited 402 jury-eligible participants to watch the videotaped trial and answer a series of questionnaires. Finally, each juror completed an

instructional comprehension test on the judicial instructions guiding their sentencing decision.

Lynch and Haney's results from 2000 revealed a bias against black defendants among those with a low comprehension of sentencing instructions. In particular, jurors who did not understand the role of mitigating and aggravating circumstances were more likely to treat mitigating factors as aggravating in black than white defendant cases. For example, when a defendant had psychiatric problems, a mitigating factor, jurors mistakenly used this as an aggravating factor for 18% of black defendants but only 9% of white defendants. Further, even when mitigation was defined properly, the evidence was regarded as "significantly less mitigating" for black than for white defendants.

Based on this finding, Lynch and Haney concluded that black defendants faced the most pronounced discrimination by those who least understood the judge's instructions, and this discrimination was manifested in a misapplication of circumstances that led to a harsher view of the crime. Haney has elsewhere identified this as the inevitable result of an "empathic divide" between white jurors and black defendants (Lynch and Haney 2009). This divide can lead jurors to engage in what Haney refers to as "moral disengagement" to separate themselves from the defendants they sentence (Haney 1997).

An important 2006 study analyzed over 600 death-eligible cases in Philadelphia, Pennsylvania, between 1979 and 1999 and showed how arbitrary this type of racial bias can be (Eberhardt et al. 2006). Forty-four of the cases involved a black defendant and white victim; another 308 had a black defendant and a black victim. Over 40 (mostly white) Stanford undergrads rated "the stereotypicality of each Black defendant's appearance" using whatever indication they felt appropriate. The study found that for cases in which Blacks Killed Whites, "24.4% of those Black defendants who fell in the lower half of the stereotypicality distribution received a death sentence, whereas 57.5% of those Black defendants who fell in the upper half received a death sentence." That this finding represents racial bias in the capital punishment regime is

underscored by the fact that when a black defendant was accused of killing a black victim, the defendant's "stereotypical blackness" did *not* predict a sentence of death. In other words, it is not something intrinsic to "stereotypical black" defendants that makes them more likely to be sentenced to death but rather how the system processes their cases when race becomes salient, as it apparently is in cases involving a black defendant who killed a white victim.

Conclusion

Despite decades of attempts to show that capital punishment deters murder, no study that purports to reach that finding has been deemed to meet the standards of modern empirical research (National Research Council 1978; Donohue and Wolfers 2005, 2009; Kovandzic et al. 2009; National Research Council 2012). Given the impossibility of employing randomized executions, it is not clear whether any stronger refutation of the deterrence hypothesis is possible.

At the same time, a large and growing literature suggests that the probability of being sentenced to death is powerfully influenced by the interaction of the race of the defendant and victim (Baldus et al. 1990; United States General Accounting Office 1990; Lynch and Haney 2000; Blume et al. 2004; Eberhardt et al. 2006; Lynch and Haney 2009; Donohue 2013). Most studies find that killers of white victims are far more likely to receive the death penalty than killers of minority victims. In addition, homicide cases with black defendants and white victims are significantly more likely to receive death penalty sentences, even when controlling for the egregiousness of the crime.

Other arenas of death penalty research in the United States have also focused on the financial cost of a death penalty system (Cook 2009; Roman et al. 2009; Alercon and Mitchell 2011) and the high rates of sentencing reversals (Liebman et al. 1999; Alesina and La Ferrara 2011), two factors that have fueled criticism of the current system. With US crime rates at a relative low after the crime surge of the late

1960s through early 1990s and with a series of DNA exonerations of death row inmates, enthusiasm for the death penalty in the United States appears to be on the decline. Whether these factors coupled with a perceived lack of deterrent benefit, the scourge of racial discrimination in implementation, and the substantial costs of a death penalty system will further the recent trends of state abolitions or be overwhelmed by a counter-insurgency by pro-death penalty forces will be one of the interesting features of the criminal justice landscape over the next decade and beyond.

References

- Alercon A, Mitchell P (2011) Executing the will of the voters?: a roadmap to mend or end the California legislature's multi-billion-dollar death penalty debacle. *Loyola Law Rev* 44:S41–S224
- Alesina AF, La Ferrara E (2011) A test of racial bias in capital sentencing (No. w16981). National Bureau of Economic Research, Cambridge, MA
- American Law Institute (2009) Report of the council to the members of the American Law Institute on the matter of the death penalty. Available at http://www.ali.org/doc/Capital%20Punishment_web.pdf
- Baldus D, Woodworth G, Pulaski C (1990) Equal justice and the death penalty: a legal and empirical analysis. Northeastern University Press, Boston
- Blume J, Eisenberg T, Wells MT (2004) Explaining death row's population and racial composition. *J Empir Leg Stud* 1(1):165–207
- Cook P (2009) Potential savings from abolition of the death penalty in North Carolina. *Am Law Econ Rev* 11(2):498–529
- Donohue J (2013) Capital punishment in connecticut, 1973–2007: a comprehensive evaluation from 4686 murders to one execution. Available at http://works.bepress.com/john_donohue/87
- Donohue J, Wolfers J (2005) Uses and abuses of empirical evidence in the death penalty debate. *Stanf Law Rev* 58:791–846
- Donohue J, Wolfers J (2009) Estimating the impact of the death penalty on murder. *Am Law Econ Rev* 11(2):249–309
- Eberhardt JL, Davies PG, Purdie-Vaughns VJ, Johnson SL (2006) Looking deathworthy: perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychol Sci* 17(5):383–386
- Ehrlich I (1975) The deterrent effect of capital punishment. *Am Econ Rev* 65(3):397–417
- Haney C (1997) Violence and the capital jury: mechanisms of moral disengagement and the impulse to condemn to death. *Stanf Law Rev* 49:1447–1463
- Kovandzic TV, Vieraitis LM, Boots DP (2009) Does the death penalty save lives? *Criminol Public Policy* 8:803–843
- Kuziemko I (2006) Does the threat of the death penalty affect plea bargaining in murder cases? Evidence from New York's 1995 reinstatement of capital punishment. *Am Law Econ Rev* 8(1):116–142
- Liebman L (2009) ALI withdraws section 210.6 (capital punishment) of the model penal code. American Law Institute. Available at http://www.ali.org/_news/10232009.htm
- Liebman JS, Fagan J, West V, Lloyd J (1999) Capital attrition: error rates in capital cases, 1973–1995. *Tex Law Rev* 78:1839–1865
- Lynch M, Haney C (2000) Discrimination and instructional comprehension: guided discretion, racial bias, and the death penalty. *Law Hum Behav* 24(3):337
- Lynch M, Haney C (2009) Capital jury deliberation: effects on death sentencing, comprehension, and discrimination. *Law Hum Behav* 33(6):481–496
- Morgenthau R (1995) What prosecutors won't tell you. *The New York Times*, 7 Feb 1995
- National Research Council (1978) Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates. The National Academies Press, Washington, DC
- National Research Council (2012) Deterrence and the death penalty. The National Academies Press, Washington, DC
- Roman JK, Chalfin AJ, Knight CR (2009) Reassessing the cost of the death penalty using quasi-experimental methods: evidence from Maryland. *Am Law Econ Rev* 11(2):530–574
- Sunstein CR, Vermeule A (2005) Is capital punishment morally required? Acts, omissions, and life-life tradeoffs. *Stanf Law Rev* 58:703–750
- Sunstein CR, Wolfers J (2008) Op-ed: a death penalty puzzle. *The Washington Post*, 30 June 2008
- United States General Accounting Office (1990) Report to senate and house committees on the judiciary: death penalty sentencing: research indicates pattern of racial disparities. Available at <http://archive.gao.gov/t2pbat11/140845.pdf>

Deception Games

► Counterfeit Money

Deduction

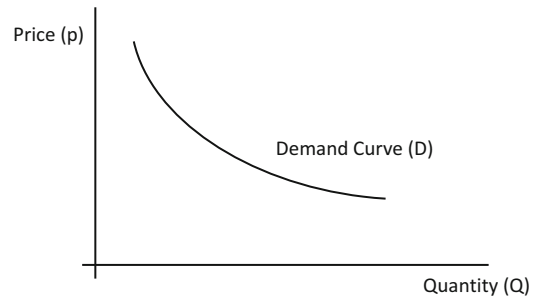
► Rationality

Demand

Silvia Ručinská¹, Ronny Müller¹ and
Jannik A. Nauerth²

¹Faculty of Public Administration, Pavol
Jozef Šafárik University in Košice, Košice,
Slovakia

²Faculty of Business and Economics,
University of Technology Dresden,
Dresden, Germany



Demand, Fig. 1 The demand curve (Hardes et al. 1995, p. 15)

Abstract

The term demand describes the willingness to buy a fixed quantity of goods or services at a specific price. This relation depends on the income and preferences of an individual but is typically expressed as an overall economic aggregate. The disposition to buy normally alters in contrary to the price.

Introduction

Demand represents the relation between demanded, purchased quantity of a good and the market price. Generally it is to express that with an increasing price of good, the willingness to buy is declining and also in the opposite way, that with a decreasing price of good, the willingness to buy a good is increasing. This relation is called the law of downward-sloping demand and can be described also through a graph in the form of a demand curve. The quantity of a good is on the horizontal axis and the price on the vertical axis, so the quantity and price are inversely related. That means the dependent volume is quantity and the independent volume the price (Samuelson and Nordhaus 1992). The demand curve slopes downward (Fig. 1):

The first, who graphically defined the demand, was in the nineteenth century A. Marshall.

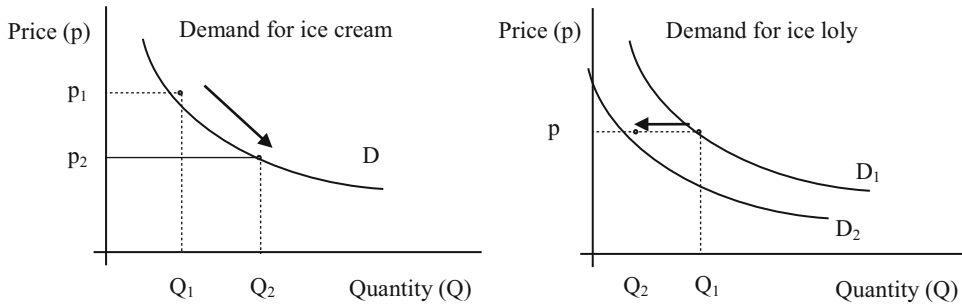
When describing and analyzing the demand, it is important to distinguish the type of demanded good, that's why we are speaking about the demand for consumer goods and the demand for production factors.

Demand for Consumer Goods

Demand for consumer goods is affected besides the price also by several non-price factors, which effect we can express, if we consider the price as a fixed value (compliance with the condition of *ceteris paribus*). Factors other than price influencing demand are price change of substitute goods, price change of complementary goods, change in the number of households on the market, change of the buyer's income, change of consumer's preferences, expectation of a future good's price change, or exceptional circumstances.

The substitute good is a good which replaces the original good and which satisfies the same need, for example, glasses and contact lenses or ice cream and ice lolly. If there is a price change of a substitute good, it affects the original good's demand, for example, if the price of a substitute good (price of an ice cream) decreases, then the demand for the original good (ice lolly) drops without a change in the price of the original good (ice lolly) (Fig. 2).

Complementary goods represent goods, in which consumption is interlinked or supplemented, for example, computer and software or skis and sky shoes. In this kind of goods, it is also applied that if the price of a complementary good changes, it affects also the original good's demand. For example, if the price of software of several companies increases, not only the demand for the software decreases due to the effects of the law of decreasing demand but also the demand for computers. Thus price increases of a complementary good will cause a



Demand, Fig. 2 Price change of a substitute good as a factor influencing demand (own; Samuelson and Nordhaus 1992, p. 59; Frank and Bernanke 2011/2009, p. 75)

decrease of demand for a complementary good and also for an original good. If the price of a complementary good decreases, the demand for complementary good will increase and also the demand for an original good will increase.

A change of the number of households can affect the demand in a following way. Should the number of buyers increase, for example, due to migration, then the demand at such a market will increase and the demand curve moves to the right. Other way round, should the number of buyers decrease, there will be less subjects who will buy and that's why also the total demand at the market will decrease and the whole demand curve moves to the left.

A change in the income of consumers is also an important factor affecting the demand at the market. When the income of consumers is increasing, their willingness to buy is also increasing and that's why the demand is increasing and the demand curve moves to the right. If the income of consumers is decreasing, this factor affects in the opposite way.

Change in preferences of consumers is a reaction of the consumers, for example, to fashion trends (tight jeans) or to the change of lifestyle (preference of a healthy lifestyle); they are related also to hobbies or as a reaction on seller's marketing tools. If any impulse causes a growth in preferences of specific products, there is also an increase of demand and the demand curve of such products moves to the right.

Consumers can have future expectations that the price of a good will increase or decrease. If they expect that the future price will be lower, for

example, sellouts and after-Christmas discounts, today they will demand less and the demand curve moves to the left. If they expect that the future price will grow, today they will demand more and buy on stock and the demand curve moves to the right.

Demand is affected also by many exceptional factors and unexpected circumstances, for example, epidemics and floods, which increase the demand for concrete goods.

Distinguishing Between the Movement on the Curve and the Movement of the Curve

Considering demand one has to distinguish between a shift of demand and a change along the demand curve. For example, demand shifts if the preferences of individuals vary or if the future expectations change. In contrary to that, a change in the price induces movements along the demand curve without shifting it.

Income and Substitute Effect

The existence of the law of decreasing demand is connected with and explained by two reasons (effects): the income and substitute effect. If the price of a good is increasing, then the substitute effect expresses the consumer's effort to replace – substitute the consumption of an original good, in which price has increased, with a substitute good. When increasing price of a good, also

the income effect appears. It describes the behavior of consumers, so the consumers are feeling poorer when the price of goods is increasing. Thus the consumers will buy and demand less of a good, which has become more expensive (Frank and Bernanke 2011/2009, p. 65)

Giffen's Goods and the Demand of Inferior and Luxury Goods

The already described law of decreasing demand applies in a large extent, but it does not apply absolutely. However, it is possible that an increase in price level leads to an increase in demand, so-called Giffen paradox. It was found first by R. Giffen, who observed the demand for bread and meat of poor people in Ireland (Marshall 1997). In his considerations bread was an inferior product. This means that a higher income leads to lower demand for bread and higher demand for meat. He discovered that an increase in the bread price lead to an increase in the demand for bread. A modern approach to discuss Giffen's goods is delivered by Jensen and Miller (2008).

Demand Elasticity

Although the law of decreasing demand suggests that with the growth of price, the market demand will decrease, such a statement is not clear about how much the demand will decrease. This is answered by the direct price elasticity, which expresses what the percentage change of demanded goods will be, if the price of a good changes. Direct price elasticity is measured by the elasticity coefficient, which can be in the interval $<0, \infty>$ (Siebert 1996, pp. 79–82). The higher the number of the coefficient, the bigger is the reaction of demand to the price change and thus the bigger is the elasticity of demand. The more the demand is elastic, the more the demand curve is horizontal and vice versa; the less elasticity, the more vertical the demand curve. Besides the direct price demand elasticity, also income demand elasticity and indirect price demand elasticity exist.

Individual, Market, and Aggregate Demand

If we are analyzing one consumer's demand and how this consumer chooses demanded quantities related to different prices, such a demand is referred to as an individual demand. It expresses the relation between different prices of the good and of quantity what one consumer would demand related to the price. His demand is expressed by the individual demand curve, and it can be derived through an indifference analysis when examining the consumer's optimum through indifference curves and budget lines.

The sum of individual demands will create a market demand, which is the demand of all consumers for one good. Aggregate demand is a demand of all subjects for all goods in the economy.

Production Factor's Demand

Demand for consumer goods determines how the demand for production factors will be; that's why we are saying that the demand for production factors is a derived demand.

Conclusion

Demand analysis is always connected to concrete goods or to a concrete group of consumers or to concrete markets. If we want to better understand the functioning of markets through the demand analysis, it is appropriate to link the demand analysis with supply.

References

- Chauhan SPS (2009) Microeconomics: an advanced treatise. PHI Learning, New Delhi
- Frank RH, Bernanke BS (2011/2009) Principles of microeconomics, brief edition. Mc Graw-Hill/Irwin, New York
- Hardes HD et al (1995) Volkswirtschaftslehre – problemorientiert. J.C.B. Mohr (Paul Siebeck), Tübingen
- Jensen RT, Miller NH (2008) Giffen behavior and subsistence consumption. *Am Econ Rev* 98(4):1553–1557

- Marshall A (1997) Principles of economics. Prometheus Books, Amherst
- Samuelson PA, Nordhaus WD (1992) Economics, 14th edn. Mc Graw-Hill
- Siebert H (1996) Einführung in die Volkswirtschaftslehre. W. Kohlhammer, Stuttgart/Berlin/Köln

Detention

► Prisons

Development and Property Rights

Peter J. Boettke¹ and Rosolino A. Candela²

¹Department of Economics, George Mason University, Fairfax, VA, USA

²Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

Abstract

This chapter argues that economic development originates not from the gains from trade and specialization under a division of labor but fundamentally from an institutional framework of property rights which permits the gains from trade and innovation that emerge on a societal wide scale. It is this framework that enables the transition from small-scale trading and capital accumulation to medium-scale trading and capital accumulation and finally to large-scale trading and capital accumulation. All of humanity was once poor; those societies that have been able to escape from poverty are those that were able to get on this development path by adopting the institutional framework of property, contract, and consent. We argue that well-defined and exchangeable private property rights yield economic growth by operating as a filter on economic behavior – the establishment of property rights embedded in the rule of law weeds out unproductive entrepreneurship and the corresponding politicized redistribution of property rights, with rent-

seeking and predation as its consequence, and engenders instead productive entrepreneurship and a more efficient allocation of property rights and with that a realization of the gains from trade and the gains from innovation. The fundamental cause of economic development, we argue, is the institution of private property, as it is this institutional framework that results in productive specialization and peaceful cooperation among diverse and disparate individuals.

Introduction

In the introduction to their book, *How the West Grew Rich*, Rosenberg and Birdzell (1986) argue that little controversy exists that institutions, namely, private property rights as well as the rule of law and enforcement of contracts, are the fundamental determinant to economic growth. There is a general consensus among economists that the unprecedented gains in labor productivity and innovation beginning in the nineteenth-century England cannot be attributed exclusively to the proximate causes of growth, such as the expansion of international trade, the accumulation of physical capital, or utilization of the economies of scale, all of which by themselves exhibit diminishing returns to scale (Phelps 2013, pp. 5–8; McCloskey 2010, pp. 133–177). Rather, the fundamental cause of modern economic growth is the institutional framework that makes possible the increasing specialization and widening circle of exchange. The “virtuous cycle” is implied in Adam Smith’s famous dictum that the “division of labor is limited by the extent of the market.” Increased possibilities of trade result in increasing specialization and a more extensive division of labor, which in turn increases the productive capacity of individuals and leads to great trading opportunities. With specialization and trade, there is also great scope of opportunities for innovation.

An understanding of how private property generates economic development also provides a perspective of the processes that emerge from such an institutional environment, which is necessary for prosperity. Observation of countries around the

world indicates that those countries with an institutional environment of secure property rights have achieved higher levels of various measures of human well being, including not only higher GDP per capita, but also lower infant mortality rates and higher rates of education. Private property rights structure human interaction by providing individuals three main mechanisms of social coordination and conflict resolution: (1) excludability, (2) accountability, and (3) exchangeability.

Well-defined and exchangeable private property rights yield economic growth by operating as an entrepreneurial filter. By structuring the costs and benefits of exchange, private property rights economize on the emergence of certain patterns of behavior by (1) filtering in productive entrepreneurship, leading to a more efficient partitioning of property rights and technological innovation as its outcome, and (2) filtering out unproductive entrepreneurship that leads to a politicized redistribution of property rights with rent-seeking and predation as its consequence.

The basic thesis of this entry is that the process of economic development goes as follows: the only way to achieve sustained increases in real income is to increase real productivity. Such increases in real productivity come from investments and technological innovation that increase and improve physical and human capital. However, because of the heterogeneity of capital, capital accumulation is a necessary, though not a sufficient, condition for economic growth. Such capital formation can only be undertaken through a decentralized price system, which coordinates the particularized insights of entrepreneurs about opportunities for gains from trade and gains from innovation. This manifests itself in technological change by discovering new and improved combinations of land, labor, and capital to satisfy the most valued consumer demands. The production plans of some must mesh with the consumption demands of others, and this is accomplished through the guiding influence of monetary calculation and the weighing of alternative investment decisions. Moreover, the allocation of entrepreneurship into productive activities that improve productivity and increase real income depends upon an institutional framework that widens the

extent of the market for entrepreneurs “where they can take advantage of increasing returns to ability” (Murphy et al. 1991, p. 510). An institutional framework of secure and exchangeable private property rights is sufficient for the emergence of a decentralized price system that provides profit and loss signals to entrepreneurs to discover new technological opportunities, generating economic growth.

We Were All Once Poor

Since at least the days of Adam Smith, economists have debated why certain societies have grown rich while others have remained stagnant and poor. Despite the unprecedented economic growth that has transformed the West and more recently China and India, many parts of the world today, particularly sub-Saharan Africa, are still poverty stricken. Many development economists, most notably Jeffrey Sachs, have argued that sub-Saharan Africa has been stuck in a “poverty trap,” resulting in unsustainable levels of economic growth that are not robust enough to bring Africa out of poverty. Africa’s extreme poverty levels lead to low savings rates, which in turn lead to low or negative economic growth, which cannot be offset by large inflows of foreign capital. Therefore, an investment strategy focusing on specific *interventions*, defined broadly as the provision of goods, services, and infrastructure, would be required, including improved education, which in turn leads to reductions in income poverty, hunger, and child mortality. The concept of a “poverty trap” has been a long-standing hypothesis in theories of economic growth and development (Sachs et al. 2004).

The underlying premise behind the poverty trap hypothesis is that the conditions of poverty are unique to Africa and other developing regions around the world and that the West has been uniquely endowed with economic wealth. The poverty trap hypothesis leads to the presumption that the West can save Africa (Easterly 2009), particularly through increasing transfers of foreign aid. However, development economist Peter Bauer, one of the most outspoken critics of

modern development economics in the twentieth century, wrote the following:

To have money is the result of economic achievement, not its precondition. That this is so is plain from the very existence of developed countries, all of which originally must have been underdeveloped and yet progressed without external donations. The world was not created in two parts, one with ready-made infrastructure and stock of capital, and the other without such facilities. Moreover, many poor countries progressed rapidly in the hundred years or so before the emergence of modern development economics and the canvassing of the vicious circle. Indeed, if the notion of the vicious circle of poverty were valid, mankind would still be living in the Old Stone Age. (2000, p. 6)

The engine of growth that transforms a society from “subsistence to exchange” (Bauer 2000, p. 3) is trading activity, leading to what Adam Smith recognized as “the greatest improvement in the productive powers of labour, and the greater part of the skill, dexterity, and judgment with which is any where directed, or applied, seemed to have been the effects of the division of labour” (Smith 1776[1981], p. 13). The absence of exchange opportunities precludes social cooperation under the division of labor and the emergence of specialized skills and crafts.

Smith pointed out that the division of labor was limited by the extent of the market. By widening its extent, individuals could capture increasing returns from specialization and trade. While Smith had emphasized the role of international trade in promoting economic growth, Bauer focused on the neglect among development economists of “internal trading activity” which in emerging economies leads to “not only the more efficient deployment of available resources, but also the growth of resources” (2000, p. 4). When individuals exercise their comparative advantage, not only are they able to produce goods and services beyond their subsistence level of consumption, but such surplus consumption can be deferred as savings and investment, not only in physical capital but also in human capital. Through increasing investments in physical and human capital, economies become more productive.

What is lost among First World observers is that in the developing world, much of this investment takes place in nonmonetary forms. As most

production in the developing countries is labor intensive and agriculturally based, “these investments include the clearing and improvement of land and the acquisition of livestock and equipment. Such investments constitute capital formation” (Bauer 2000, p. 11). Because much of this investment is not calculated in money prices, these forms of investment “are generally omitted from official statistics and are still largely ignored in both the academic and the official development literature” (Bauer 2000, p. 11).

The fundamental basis of economic development from subsistence to exchange entails well-defined, enforceable, and exchangeable property rights. Most of the developing world today remains poor because governments are predatory or because governments are unable to enforce private property rights, precluding the advance from subsistence to exchange. Just as Bauer pointed out that the “small-scale operations” of trade and nonmonetary investment are required for economic development, analytically speaking, what allowed for the birth of economic development in the West and in those emerging economies embarking in economic growth today was the development of various types of property rights arrangements:

We ought not to be surprised if we find that in the relatively short history of man, he has already devised, tested, and retained an enormous variety of allocations and sharing of property rights. The history of the law of property reveals an overwhelming and literally incomprehensible variety. (Alchian 1961[2006], p. 33)

The absence of tried and tested mechanisms of private property rights and their enforcement would have thwarted modern economic growth in the West and those developing countries emerging from poverty today. Within the framework of private property rights under the rule of law, individuals are able to form reliable expectations about how their land, labor, capital, and entrepreneurial talent can be permissibly utilized. In rich countries, property rights provide a framework of rules that provide a degree of legal certainty so that individuals reliably coordinate their actions amid the flux and “throng of economic possibilities that one can only dimly perceive” (Mises

1922[2008], p. 117) over an uncertain economic horizon.

Some Development Economics of Property Rights

James Buchanan stated “the economist should not be content with postulating models and then working within such models. His task includes the derivation of the institutional order itself from the set of elementary behavioral hypotheses with which he commences. In this manner, genuine institutional economic becomes a significant and important part of fundamental economic theory” (Buchanan 1968[1999], p. 5). However, throughout most of the twentieth century, neoclassical economists have largely neglected the framework within which exchange and production takes place. In the textbook neoclassical model, individuals are presumed to have perfect information and are able to engage in costless exchange without incurring any externalities, or third-party effects, on other individuals. Property rights are the given background of analysis of competitively perfect markets, but a theoretical framework cannot account for the innumerable contractual arrangements, such as firms and money, that emerge in order to reduce the transaction costs of engaging in market exchange.

However, the economic analysis of property rights, although neglected, was not completely overlooked. Scholars working within the property rights, law and economics, public choice, and Austrian market process perspectives all took property rights out from underneath the cover of the “given background” to analyze the evolution and allocation of property rights and how alternative institutional arrangements of property rights will have different consequences on the pattern of exchange and production. The leading twentieth-century economists who emphasized the importance of private property rights to economic theory were Ludwig von Mises, Friedrich Hayek, James Buchanan, Ronald Coase, Armen Alchian, and Harold Demsetz. Their research emphasized how different delineations of property rights lead to different economic outcomes. Because of

scarcity of knowledge and other resources, competition among individuals emerges in all societies. However, the manner in which competition manifests itself was *institutionally contingent* to the cost-benefit structure of property rights. But as neoclassical economics grew increasingly focused on static equilibrium analysis after the 1930s, what emerged was an institutionally anti-septic theory of choice. This preoccupation with the properties of static equilibrium shifted theoretical attention away from the institutional context of choice, namely, how private property rights structure the marginal costs and benefits of choice that generate a dynamic tendency toward equilibrium (Boettke 1994[2001], p. 236). Economist Svetozar Pejovich defines property rights in this way:

Property rights are relations among individuals that arise from the existence of scarce goods and pertain to their use. They are the norms of behavior that individuals must observe in interaction with others or bear the costs of violation. Property rights do not define the relationship between individuals and objects. Instead, they define the relationship among individuals with respect to all scarce goods. The prevailing institutions are the aggregation of property rights that individuals have. (italics original, 1998, p. 57)

Private property rights structure human interaction by providing individuals three main mechanisms of social coordination and conflict resolution: (1) excludability, (2) accountability, and (3) exchangeability. Excludability means that individuals are free to use and dispose of their property rights over a particular resource and exclude other individuals from utilizing their property rights so long as they do not violate the property rights of other individuals, namely, the physical properties of their body and the resources they own.

Accountability assigns residual claimancy over the costs and benefits of an action initiated by an individual. Without private ownership, when a person uses resources, they impose a cost on everyone else in the society, leading to a tragedy of the commons. Therefore, private property rights provide accountability over the costs and benefits of individual's actions through the internalization of positive and negative

externalities (Demsetz 1967, p. 350). Through such internalization, private property rights over resources provide the incentive for individuals to maximize the present value of their resources by taking into account alternative future time streams of benefits and costs and selecting that one which he believes will maximize the present value of his resources. Private property incentivizes individuals to economize on resource use and maintain capital for future production because the user bears the costs of their actions. Poorly defined and enforced property rights lead to overuse and depletion of resources since the decision maker of a particular action does not bear the full cost of his action.

Exchangeability of property rights not only allows individuals to make trades that both parties believe will make them better off. When rights over private property are transferable, it also provides an institutional framework within which a system of money prices emerges. The emergence of money prices provides the information to calculate the relative scarcity of different resources, such that “prices can act to coordinate the separate actions of different people” by communicating the dispersed and particular knowledge of millions of individuals (Hayek 1945, p. 526). People are able to observe prices and determine whether they value the property they have more than the money they could receive for it. Changes in price signals drive the movements in the demand and supply for different goods and services. These price changes provide the information to entrepreneurs as to what products are most urgently demanded and what inputs can be combined to most cheaply produce them. Absent the free exchange of private property rights, this *contextual* information embodied in money prices is not generated (Mises 1920[1975]). Since entrepreneurs have a property right, or residual claimancy, over their profits and losses, they also have every incentive to use resources to satisfy these most highly valued demands.

The crucial link between private property rights and such unprecedented economic growth lies with increasing returns to the division of knowledge embodied in entrepreneurial activity. As the extent and complexity of the market widen,

so do the complexity and specialization of knowledge within the market. The effective partitioning of property rights enables individuals to specialize in applying their particularized knowledge of time and circumstance in the discovery of previously unnoticed profit opportunities conducive to capital investment and technological progress (Alchian 1965[2006], p. 63). Through this process, entrepreneurship effectively leads to greater productivity, higher real wages, an expansion of output, and an overall increase in human welfare.

Property Rights and Entrepreneurship

Certain institutional frameworks encourage the spontaneous order of the market economy, as well as its entrepreneurial drive towards economic growth, while others erect barriers to growth and pervert the incentives of entrepreneurs towards rent-seeking and predation. Private property rights and their exchangeability ensure the emergence of a spontaneous order, in which entrepreneurs are driven by consumer preferences and encouraged to invest in enterprises that spur innovation and create wealth. Through purposeful actions of entrepreneurs, economic resources and knowledge, which are dispersed and particular to time and place, are coordinated through the incentives of the price system. However, how entrepreneurs coordinate economic knowledge and resources depends heavily on the institutional framework, or the rules of the game, that happen to prevail in the economy.

The prosperity or stagnation of societies rests on the allocation of entrepreneurship (Baumol 1990). The institutions that constrain human behavior within a particular society largely influence how entrepreneurial activity will be allocated and the nature of their purpose, which may be productive or unproductive in result. In prosperous societies, in which exchangeable private property rights have prevailed, entrepreneurship has been driven by consumer preferences and led the market process to more efficient outcomes, leading to economic growth. Poor and stagnating societies are characterized by institutions that are interventionist and arbitrary, leading to the

politicization of entrepreneurial activity. Such an environment encourages rent-seeking and predation and discourages innovation, capital investment, and economic growth.

It is not only because individuals have limited means to satisfy their innumerable wants that property rights structure the rewards and costs of human interaction but more fundamentally because knowledge about how to *discover* such means is scarce as well. Property rights structure the costs and rewards of utilizing particularized knowledge in the application and specialization of particular forms of entrepreneurial talent, both productive and unproductive.

Moreover, private property rights yield economic growth by operating as an entrepreneurial filter. By structuring the costs and benefits of exchange, private property rights economize on the emergence of certain patterns of behavior by filtering in productive entrepreneurship, leading to technological innovation and enhanced productivity, and filtering out unproductive entrepreneurship, which leads to rent-seeking and predation. In a world of uncertainty, the means by which individuals pursue different economic ends are unknown and must be discovered through entrepreneurship.

According to Israel Kirzner (1988, p. 179), entrepreneurship refers to the process of individuals acting upon profit opportunities “that could, in principle, have been costlessly grasped earlier.” In *Competition and Entrepreneurship*, Kirzner further elaborates on the process of entrepreneurship:

The entrepreneur is someone who hires the factors of production. Among these factors may be persons with superior knowledge of market information, but the very fact that these hired possessors of information have not *themselves* exploited it shows that, in perhaps the truest sense, their knowledge is possessed not by them, but by the one who is hiring them. It is the latter who “knows” whom to hire, who “knows” where to find those with the market information needed to locate profit opportunities. Without himself possessing the facts known to those he hires, the hiring entrepreneur does nonetheless “know” these facts, in the sense that his alertness – his propensity to know where to look for information – dominates the course of events. (Kirzner 1973, p. 68, *italics original*)

Kirzner also states “the discovery of a profit opportunity *means the discovery of something obtainable for nothing at all*. No investment at all is required; the free ten-dollar bill is discovered to be already within one’s grasp” (Kirzner 1973, p. 48, *emphasis in original*). The entrepreneur’s role in the production process is to earn pure profit based on his “alertness” of where to find market data under uncertainty (Kirzner 1973, p. 67). It entails that the entrepreneur possesses the right knowledge at the right time for discovering new combinations of technological inputs for the production of new goods and services to their most valued uses. The entrepreneur does not mechanically respond to profit opportunities as a calculative, maximizing *homo economicus*. Rather, he is “alert” to price discrepancies between existing commodities and to discovering previously unknown opportunities for mutually beneficial exchange. It is the entrepreneurial element in each individual “that is responsible for our understanding of human action as active, creative, and human rather than as passive, automatic, and mechanical” (Kirzner 1973, p. 35). It is through the discovery of profit opportunities that entrepreneurs discover how resources must be allocated to satisfy their most valued uses.

The Smithian growth process that was described above rests not only on passive capital accumulation but more importantly on the increasing returns to knowledge that enlarge the extent for entrepreneurial activity. The emergence of knowledge externalities through the entrepreneurial pursuit of pure profit opportunities links the fundamental relationship between private property rights and economic growth (Holcombe 1998, pp. 51–52). Demsetz states that:

Property rights develop to internalize externalities when the gains of internalization become larger than the cost of internalization. Increased internalization, in the main, results from changes in economic values, changes which stem from the development of new technology and the opening of new markets, changes to which old property rights are poorly attuned. (1967, p. 350)

Following Kirzner, Holcombe goes further to state that entrepreneurship drives the changes in relative prices and technology, from which:

Knowledge externalities occur when the entrepreneurial insights of some produce entrepreneurial opportunities for others. Increasing returns occur because the more entrepreneurial activity an economy exhibits, the more new entrepreneurial opportunities it creates. (Holcombe 1998, p. 58)

The key to understanding the engine that drives the market economy towards efficient outcomes is the fact that today's inefficiencies are tomorrow's profit opportunities for entrepreneurs to seize upon such externalities of knowledge. However, this entrepreneurial market process requires that private property rights assign to entrepreneurs residual claimancy over the costs and rewards of their actions in the form of monetary profits and loss. The consequence of poorly defined property rights will be the destruction of wealth. Without secure property rights, economic calculation will break down, as money prices will not reflect the relative economic profitability of using different quantities and qualities of scarce inputs, such as land, labor, and capital. As a result, insecure property rights will shrink the extent of the market for productive entrepreneurship. Murphy et al. (1991) also point out that unproductive entrepreneurship is more prevalent in countries with poorly defined property rights since the market for rent-seeking is larger and more lucrative there. As they argue, "rent seeking pays because a lot of wealth is up for grabs," (1991, p. 519) particularly for those entrepreneurs who are successful at defining property rights through bribery, theft, or litigation. Such entrepreneurial activity is wasteful since entrepreneurs are committing their time, knowledge, and resources not to creating more efficient ways of producing goods and services but in transferring wealth or resisting other entrepreneurial competitors from capturing their rents (Tullock 1967, p. 228).

Conclusion

Economic development originates from trading activity, specialization, and social cooperation under a division of labor. But the *cause* of economic development is inextricably linked with a framework of well-defined, enforceable, and exchangeable private property rights. Economic

growth and development, driven by entrepreneurship, cannot be explained independent of its institutional context, namely, private property rights. Entrepreneurship by itself cannot be the fundamental cause of economic development, since scarcity, competition, and entrepreneurship exist in all societies. Rather, the manner in which entrepreneurship manifests itself is a *consequence* of the structure of property rights (Boettke and Coyne 2003). The fundamental cause of economic development is the adoption of well-defined and exchangeable private property rights, which incentivizes entrepreneurship to act on profit opportunities that facilitate increasing gains from exchange and specialization, spurring capital investment, increased labor productivity, and higher real income.

Recognizing the link between property rights, entrepreneurship, and economic growth has important implications not only for economic theory but also for economic policy as well. Assuming away the institutional differences in property rights arrangements across countries leads to misleading policy advice about how poor nations can emerge from poverty. In the wake of the fall of the Berlin Wall and the collapse of communism in Eastern Europe, Peter Murrell asked whether neo-classical economics could underpin the reform of centrally planned economies. As he wrote, "reformers need a filter that interprets the experience of capitalist and socialist systems" (Murrell 1991, p. 59). Such a filter refers to a comparative institutional analysis of property rights that have emerged to fit the historical and cultural context of a particular time and place.

By focusing on the accumulation of production inputs, such as physical and human capital, to increase productivity and real income, economic policymakers have focused on investment as well as research and development to spur economic growth. However, failing to take account of the framework of property rights misleadingly places factor accumulation as a fundamental cause of economic development, rather than as a proximate cause. Capital investment by itself does not cause economic growth but emerges in response to productive entrepreneurship incentivized by a framework of private property.

References

- Alchian AA (1961[2006]) Some economics of property. In: Property rights and economic behavior. The collected works of Armen A. Alchian, vol 2. Liberty Fund, Indianapolis
- Alchian AA (1965[2006]) Some economics of property rights. In: Property rights and economic behavior. The collected works of Armen A. Alchian, vol 2. Liberty Fund, Indianapolis
- Bauer P (2000) From subsistence to exchange and other essays. Princeton University Press, Princeton
- Baumol WJ (1990) Entrepreneurship: productive, unproductive, and destructive. *J Polit Econ* 98(5):893–921
- Boettke PJ (1994[2001]) The political infrastructure of economic development. In: Calculation and coordination: essays on socialism and transitional political economy. Routledge, New York
- Boettke PJ, Coyne C (2003) Entrepreneurship and development: cause or consequence? *Adv Austrian Econ* 6:67–87
- Buchanan JM (1968[1999]) The demand and supply of public goods. The collected works of James M. Buchanan, vol 5. Liberty Fund, Indianapolis
- Demsetz H (1967) Toward a theory of property rights. *Am Econ Rev* 57(2):347–359
- Easterly W (2009) Can the West save Africa? *J Econ Lit* 47(2):373–447
- Hayek FA (1945) The use of knowledge in society. *Am Econ Rev* 35(4):519–530
- Holcombe RG (1998) Entrepreneurship and economic growth. *Q J Austrian Econ* 1(2):45–62
- Kirzner IM (1973) Competition & entrepreneurship. University of Chicago Press, Chicago
- Kirzner IM (1988) Some ethical implications for capitalism of the socialist calculation debate. *Soc Philos Policy* 6(1):165–182
- McCloskey DN (2010) Bourgeois dignity: why economics can't explain the modern world. University of Chicago Press, Chicago
- Mises L (1920[1975]) Economic calculation in the socialist commonwealth. In: Hayek FA (ed) *Collectivist economic planning*. August M. Kelley, Clifton
- Mises L (1922[2008]) *Socialism: an economic and sociological analysis*. Ludwig Von Mises Institute, Auburn
- Murphy KM, Shleifer A, Vishny RW (1991) The allocation of talent: implications for growth. *Q J Econ* 106(2):503–530
- Murrell P (1991) Can neoclassical economics underpin the reform of centrally planned economies. *J Econ Perspect* 5(4):59–76
- Pejovich S (1998) *Economic analysis of institutions and systems*, revised 2nd edn. Kluwer, Norwell
- Phelps E (2013) *Mass flourishing: how grassroots innovation created jobs, challenge, and change*. Princeton University Press, Princeton
- Rosenberg N, Birdzell LE Jr (1986) *How the West grew rich: the economic transformation of the industrial world*. Basic Books, New York
- Sachs JD, McArthur JW, Schmidt-Traub G, Kruk M, Bahadur C, Faye M, McCord G (2004) Ending Africa's poverty trap. *Brook Pap Econ Act* 2004(1):117–216
- Smith A (1776[1981]) *An inquiry into the nature and causes of the wealth of nations*. Liberty Fund, Indianapolis
- Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. *West Econ J* 5(3):224–232

Difference-in-Difference

J. L. Jiménez¹ and J. Perdiguerro²

¹Facultad de Economía, Empresa y Turismo, Departamento de Análisis Económico Aplicado. Despacho D.2-12, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

²Departament d'Economia Aplicada, Research Group of "Economia Aplicada" (GEAP), Universitat Autònoma de Barcelona, Bellaterra, Spain

Abstract

The difference-in-difference (DiD) is one of the most popular approaches to evaluate causal effects of programs or policies. The idea is very simple: a treatment group is affected by an external change in one period, and the main aim is to evaluate how this treated group changes after the policy, regarding a control group that is not affected. So it controls the double difference (changes over time and over control group). Although some assumptions have to be assumed, its flexibility and requiring a relatively small volume of data yield to a large number of papers and documents that use it on topics as competition policy, merger evaluations, political economy, and so on.

The difference-in-difference (DiD) is one of the most popular approaches to evaluate causal effects of programs or policies. This empirical strategy has become widespread not only in economics but also in other social sciences. Although it shows some assumptions and it is not without criticism, a lot of academic research and private and public documents use this technique.

The idea is very simple: We have data on two similar groups of interest in two different periods. One of them is affected by an external policy in the second period (the treated group). The main aim is to evaluate how the treated group changes after the policy, regarding the other, i.e., the control group. The DiD not only considers differences by group but by time. In fact, as Imbens and Wooldridge (2009) state, in this double difference, the average gain over time in the control group is subtracted from the gain over time in the treated group.

The Structure of a Difference-In-Difference Analysis

The setting of this model can be summarized as follows. We have two groups (treated and control) in two different periods (1 or before and 2 or after). In period 2, a treatment occurs and affects only to the treated group. Our aim is to evaluate how the outcome (prices, quantities, wages, GDP, or any variable of interest) changes due to the treatment. So we have to consider both the time effect that incides in two groups and the previous differences by group.

Table 1 includes the outcome in each situation (subscript 0 is before; Superscript TRT is Treated group and CRT is Control group). On one hand, if we calculate the differences between after and before by groups, we are not considering differences by group (last column). On the other hand,

if we only consider the difference between treated and control group, we are not considering time effects (last row). For these reasons, the cornerstone is to evaluate these two differences simultaneously, i.e., how treated changes regarding how control changes (last cell in Table 1).

Equation (1) shows the basic model to estimate in a pool database:

$$\begin{aligned} \text{Endogenous}_{it} = & \beta_0 + \beta_1 \text{After}_t + \beta_2 \text{Treated}_i \\ & + \beta_3 \text{After} * \text{Treated} (\text{Interaction})_{it} \\ & + \sum_{j=4}^n \beta_j X_{it} + \varepsilon_{it} \end{aligned} \quad (1)$$

Estimations simultaneously have to include three binary variables: after, treated, and the interaction of these two covariates (the coefficient β_3). The former takes value 1 for two groups in the period 2. The variable treated takes value 1 if the observation corresponds to the treated group, regardless the period. The latter attempts to assess the causal effect of the treatment in the treated group.

Empirically, this effect is summarized in Table 2.

Assumptions and Robustness Checks

One of the pillars of the difference-in-difference is that the policy or change analyzed was an

Difference-in-Difference, Table 1 Explanation of coefficients for Eq. (1) and subsequents

		Time effect		Difference (Af–Be)
		Before (Be)	After (Af)	
Group effect	Control (Co)	Y_0^{TRT}	Y_1^{TRT}	$Y_1^{\text{TRT}} - Y_0^{\text{TRT}}$
	Treated (Tr)	Y_0^{CRT}	Y_1^{CRT}	$Y_1^{\text{CRT}} - Y_0^{\text{CRT}}$
Difference (Tr–Co)		$Y_0^{\text{TRT}} - Y_0^{\text{CRT}}$	$Y_1^{\text{TRT}} - Y_1^{\text{CRT}}$	$Y_1^{\text{TRT}} - Y_1^{\text{CRT}} - (Y_0^{\text{TRT}} - Y_0^{\text{CRT}})$

Difference-in-Difference, Table 2 Explanation of coefficients for Eq. (1)

		Time effect		Difference (Af–Be)
		Before (Be)	After (Af)	
Group effect	Control (Co)	β_0	$\beta_0 + \beta_1$	β_1
	Treated (Tr)	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_3$
Difference (Tr–Co)		β_2	$\beta_2 + \beta_3$	β_3

exogenous change, as Lafontaine and Slade (2008) point out. It has been called as a “natural experiment,” which refers to an analysis that fulfills these three conditions (exogenous change, a group affected by the change, and an unaffected group).

Another of the main basic assumptions of the difference-in-difference models is that the temporal effect in the two groups is the same in the absence of the exogenous change. This has been called the “identifying assumption.” So, we first have to test whether both treatment and control group show the same trend before the change. In order to check it, we estimate a similar equation than Eq. (1) but we substitute DiD estimator by separate dummies for treatment and control groups, in order to check whether the time trends in the pretreatment period were the same.

The empirical strategy is the following one: firstly we create time dummies for both control and treatment group. Then, we estimate each equation replacing DiD variables for these previous variables generated. Finally, we test whether coefficients for each group of time dummies are equal or not (see Albalade 2008, for further explanation of this empirical strategy).

Simplicity aside, its great advantage is its potential to avoid many of the problems of endogeneity that habitually arise when carrying out comparisons among heterogeneous individuals (see Bertrand et al. 2004). Nevertheless, these authors also argue that DiD is in practice subject to a possibly serial correlation problem.

A second possible criticism is the potential endogeneity in the change in the market. When the change that occurs in the market is not exogenous, the DiD estimator will be biased and inconsistent. For example, it is clear that the decisions of merger between two or more firms are not exogenous and depend on their pricing decisions. The endogeneity problem is explained and discussed in-depth by Dafny (2009).

Some robustness checks are usually implemented in this empirical strategy. These placebo tests affect both date of treatment, treated group, and/or endogenous variable. Regarding the former, new estimations can be made changing the date of the treatment. If the DiD variable

shows the same result as in the original estimations, some problem arises.

The second placebo test can be to replace the treated group for some control groups (and vice versa). As in the previous test, we expect empirical changes in the outcome. Finally, new endogenous variables must be considered in order to control for general effects on all variables after treatment regarding the control group.

Where Has DiD Been Applied? Some Empirical Findings

The DiD estimator is extremely flexible and requires a relatively small volume of data, so it has been applied increasingly in the empirical analysis of many different aspects. It is impossible to summarize the whole set of empirical applications that have used this empirical approach, so we will only indicate some of the main aspects related to the competition policy analyzed through this indicator.

A first element is the analysis of the effects produced by the horizontal mergers. Although the authorization or not of the horizontal mergers requires an ex ante analysis, which it is impossible to realize through a methodology like the DiD, this empirical approximation can be very useful to observe the real effects that have been, and if the ex ante analysis was, accurate or not. In this sense, Peters (2006) pointed out that the simulations performed ex ante underestimated the potential effects of the mergers, as showed the results of the DiD estimator in the ex post analysis.

Equally, Hosken et al. (2017) expose the benefits of using ex post merger evaluation for ex ante analysis. Concretely they expose that ex post merger evaluations can be used to evaluate the accuracy of merger simulations by comparing predictions versus evaluations. European Commission (ex post analysis of two mobile telecom mergers: T-Mobile/tele.ring in Austria and T-Mobile/Orange in the Netherlands) and Austrian Competition Authority (BWB 2016, The Austrian market for mobile telecommunication services to private customers – An ex-post evaluation of the mergers H3G/Orange and TA/Yees!) provided two examples of it. Both documents use a difference-in-difference analysis to measure the effect of the evaluated merger.

There is a large number of examples that analyze the effects of horizontal mergers, obtained through the DiD estimator, in different sectors: in the air sector (Kim and Singal 1993; Peters 2006; Dobson and Piga 2013; or Fageda and Perdigueró 2014), in the petrol market (Taylor and Hosken 2007; Simpson and Taylor 2008; or Jiménez and Perdigueró *forthcoming*), in the scientific journals market (McCabe 2002), in the banking sector (Prager and Hannan 1998; Focarelli and Panetta 2003), or in the health sector (Connor et al. 1998; Vita and Sacher 2001; Dafny 2009). A good example of their versatility among different sectors is the paper by Ashenfelter and Hosken (2010), which analyzes the effect on prices in five different industries (female hygiene products, alcoholic drinks, lubricating oil, cereals, and breakfast syrups).

Although DiD has been used to a greater extent in the analysis of the ex post effects of horizontal mergers, it has also been used for the analysis of other equally important issues in competition policy, such as entry (Bernardo 2016), although respecting the DiD assumption of randomness in group formation is difficult, or restricting vertical relations (Vita 2000).

Outside the field of competition policy, DiD is increasingly being used to measure the impact of a broad range of public policies. Some examples are the impact of water service privatization on infant mortality (Galiani et al. 2005), the effect of the reduction of maximum permitted levels of alcohol and number of traffic accidents (Albalade 2008), the effects of the “Cash for Clunkers” programs (Jiménez et al. 2016), the effects of certain infrastructures on tourism (Albalade and Fageda 2016), the alignment of parties and the intergovernmental transfers (Solé-Ollé and Sorribas-Navarro 2008), or the effects of corruption on voters (Costas-Pérez et al. 2012) or on municipal budgets (Artés et al. 2016).

References

- Albalade D (2008) Lowering blood alcohol content levels to save lives: the European experience. *J Policy Anal Manage* 27(1):20–39
- Albalade D, Fageda X (2016) High-speed rail and tourism: empirical evidence from Spain. *Transp Res A Policy Pract* 85:174–185
- Artés J, Jiménez JL, Perdigueró J (2016) The effects of revealed corruption on local finances. Unpublished paper
- Ashenfelter O, Hosken D (2010) The effect of mergers on consumer prices: evidence from five mergers on the enforcement margin. *J Law Econ* 53(3):417–466
- Bernardo V (2016) The effect of entry restrictions on price. Evidence from the retail gasoline market. Unpublished paper
- Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *Q J Econ* 119:249–275
- BWB (2016) An ex-post evaluation of the mergers H3G/Orange and TA/Yees!
- Connor RA, Feldman RD, Dowd BE (1998) The effects of market concentration and horizontal mergers on hospital costs and prices. *Int J Econ Bus* 5:159–180
- Costas-Pérez E, Solé-Ollé A, Sorribas-Navarro P (2012) Corruption, voter information, and accountability. *Eur J Polit Econ* 28(4):469–484
- Dafny LS (2009) Estimation and identification of merger effects: an application to hospital mergers. *J Law Econ* 52:523–550
- Dobson P, Piga C (2013) The impact of mergers on fares structures: evidence from European low-cost airlines. *Econ Inq* 51:1196–1217
- Fageda X, Perdigueró J (2014) An empirical analysis of a merger between a network and low-cost airlines. *JTEP* 48(1):81–96
- Focarelli D, Panetta F (2003) Are mergers beneficial to consumers? Evidence from the market for bank deposits. *Am Econ Rev* 93:1152–1172
- Galiani S, Gertler P, Schargrodsky E (2005) Water for life: the impact of privatization of water services on child mortality. *J Polit Econ* 113(1):83–120
- Hosken D, Miller N, Weinberg M (2017) Ex post merger evaluation: how does it help ex ante? *Journal of European Competition Law & Practice* 8(1):41–46
- Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J Econ Lit* 47(1):5–86
- Jiménez JL, Perdigueró J (*forthcoming*) Mergers and difference-in-difference estimator: why firms do not increase prices?. *Eur J Law Econ*
- Jiménez JL, Perdigueró J, García C (2016) Evaluation of subsidies programs to sell green cars: impact on prices, quantities and efficiency. *Transp Policy* 47:105–118
- Kim H, Singal V (1993) Mergers and market power: evidence from the airline industry. *Am Econ Rev* 83:549–569
- Lafontaine F, Slade M (2008) Exclusive contracts and vertical restraints: empirical evidence and public policy. In: Buccirossi P (ed) *Handbook of antitrust economics*. MIT Press, Cambridge, pp 319–414
- McCabe MJ (2002) Journal pricing and mergers: a portfolio approach. *Am Econ Rev* 92:259–269

- Peters C (2006) Evaluating the performance of merger simulations: evidence from the U.S. airline industry. *J Law Econ* 49:627–649
- Prager RA, Hannan TH (1998) Do substantial horizontal mergers generate significant price effects? Evidence from the banking industry. *J Ind Econ* 46:433–452
- Simpson J, Taylor C (2008) Do gasoline mergers affect consumer prices? The Marathon Ashland petroleum and ultramar diamond shamrock transaction. *J Law Econ* 51:135–152
- Solé-Ollé A, Sorribas-Navarro P (2008) The effects of partisan alignment on the allocation of intergovernmental transfers. Differences-in-differences estimates for Spain. *J Public Econ* 92(12):2302–2319
- Taylor C, Hosken D (2007) The economic effects of the Marathon-Ashland joint venture: the importance of industry supply shocks and vertical market structure. *J Ind Econ* 55:419–451
- Vita MG (2000) Regulatory restrictions on vertical integration and control: the competitive impact of gasoline divorce policies. *J Regul Econ* 18:217–233
- Vita MG, Sacher S (2001) The competitive effects of not-for-profit hospital mergers: a case study. *J Ind Econ* 49:63–84

Digital Piracy

Paul Belleflamme¹ and Martin Peitz²

¹CORE and LSM, Université catholique de Louvain, Louvain-la-Neuve, Belgium

²Department of Economics, University of Mannheim, Mannheim, Germany

Definition

Digital piracy is the act of reproducing, using, or distributing information products, in digital formats and/or using digital technologies, without the authorization of their legal owners.

Introduction

The objective of this entry is to provide a comprehensive and up-to-date overview of digital piracy (this entry summarizes and updates Belleflamme and Peitz (2012)). Although we put the emphasis on the economic analysis, we also briefly present the legal context and its recent

evolution. As digital piracy consists in infringing intellectual property laws, it is important to start by understanding the rationale of such laws. That allows us to define more precisely what is meant by digital piracy. We can then move to the economic analysis of piracy. We start with the basic analysis, which explains why piracy is likely to decrease the profits of the producers of digital products; we also examine how the producers have reacted to digital piracy when it started to grow. We review next more recent contributions that point at possible channels through which piracy could improve the profitability of digital products. These channels have inspired new business models for the distribution of digital products, which we describe in the last part of the entry. Throughout the entry, we report the results of some of the most recent empirical studies, so as to quantify the impacts of digital piracy.

The Intellectual Property (IP) Protection of Information Products

Information products (such as music, movies, books, and software) are often characterized as being hardly excludable, in the sense that their creators face a hard time excluding other persons, especially non-payers, from consuming these products. This feature may undermine the incentives to create, because of the difficulty in appropriating the revenues of the creation. The production of information products may then be insufficient compared to what society would deem as optimal. One solution to this so-called “underproduction” problem is to make intellectual creations excludable by legal means. This is the objective pursued by intellectual property (IP) laws, which most countries have adopted. IP refers to the legal rights that result from intellectual activity in the industrial, scientific, literary, and artistic fields. IP laws generally distinguish among four separate IP regimes, which are targeted at different subject matters: information products (and more generally literary, musical, choreographic, dramatic, and artistic works) are protected by copyrights; the three other regimes (patents, trade secrets, and trademarks) aim at

protecting industrial property (such as inventions, processes, machines, brand names, industrial designs).

It is important to note, from an economic perspective, that IP laws alleviate the “underproduction” problem at the cost of exacerbating an “underutilization” problem. To understand this second problem in the context of information products, we need to refer to another characteristic of information products, namely, their non-rivalness, which refers to the property that their consumption by one person does not prevent their consumption by another person (for instance, the fact that some person listens to the performance of an artist does not reduce the possibility for anyone else in the audience to listen to the same performance). A consequence of this nonrivalness is that the marginal cost of production of information products is zero (i.e., taking the artist’s viewpoint in the previous example, once the show has started, it costs nothing to have one extra spectator viewing it). From the point of view of static efficiency, the price of information goods should therefore be equal to zero. However, because IP laws endow them with some market power, the creators of information products are able to set a positive price, which reduces social welfare by preventing those consumers with a low, but positive, valuation of the information products from consuming them.

In other words, IP laws aim at striking the balance between providing incentives to create and innovate while promoting the diffusion and use of the results of creation and innovation. To do so, IP rights are granted only for a limited period of time and for a limited scope. In particular, copyright protection usually lasts for a number of years (currently, 70 in both the European Union and the United States) after the creator’s death; in terms of scope, copyrights protect only the expression but not the underlying ideas.

Defining Digital Piracy

IP laws are effective only if they are properly enforced and respected. Yet, as far as information products are concerned, one observes a large-

scale violation of the laws protecting them, a phenomenon known as “piracy.” What is striking is that the illegal reproduction and distribution of copyrighted works is not only the act of criminal organizations (so-called commercial piracy) but also the act of the consumers themselves (so-called end-user piracy). (We do not review here factors that influence the piracy decision; one can indeed wonder what motivates such large-scale violation of IP laws by individuals who are normally law-abiding citizens; for a review of the literature on this topic, see Novos and Waldman (2013).)

Commercial piracy does not need much analysis, as the motivation is easily understood: criminal organizations are simply attracted by the high profit margins that the large-scale reproduction and distribution of copyrighted products generate. On the other hand, end-user piracy raises a number of issues that the fast penetration of the Internet and the digitization of information products have made much more pressing. Digital technologies have indeed drastically reduced the cost of making and distributing illegal copies while increasing their quality; thereby, they have deeply modified the interaction between end users, copyright holders, and technology companies. End-user piracy in the digital age, or for short *digital piracy*, is thus a major phenomenon that requires a thorough analysis.

The Basic Economics Analysis: Digital Piracy Decreases Profits

The main consequence of digital piracy is that it seriously limits copyright owners in their ability to control how information products get to consumers. As a result, the availability of digital copies is likely to reduce the copyright owner’s profits. This is the prediction that can be drawn from the basic *theoretical modeling* of piracy (see, for instance, Novos and Waldman 1984, Johnson 1985, and the references in Belleflamme and Peitz 2012). These models simplify the analysis by focusing on the market for a digital product supplied by a single producer. One can justify this assumption by arguing that digital products within

a given category are highly differentiated in the eyes of the consumers; the demand for any product is therefore hardly affected by the prices of other products. Even though the copyright owner acts as a monopoly, he/she faces nevertheless the competition exerted by the availability of (illegal) digital copies. Copies are seen as imperfect substitutes for the original digital product, insofar as their quality is generally lower than the quality of original products. In particular, the quality of copies primarily depends on technological and legal factors, which can be affected by public authorities (through the definition and the enforcement of IP protection) and/or by the copyright owner himself or herself (through technical protective measures). In this setting, it is possible to analyze the copyright owner's decisions about the pricing and the technical protection of original products, as well as public policy regarding IP laws.

The main results of these analyses can be summarized as follows. First, because consumers with a low cost of copying or with a low willingness to pay for quality prefer copies to original products, the copyright owner is forced to charge a lower price (than in a world where digital piracy would not exist). That clearly decreases the copyright owner's profits but increases the surplus of the consumers of original products; moreover, a number of consumers who were not willing to purchase the original product at the monopoly price get now some utility from the pirated copies. As the increase in consumer surplus outweighs the profit reduction, digital piracy results in an improvement of welfare from a static efficiency point of view (like any erosion of market power does). However, the lower profits may reduce the incentives of copyright owners to improve the quality of existing products or to introduce new products on the market; this is detrimental to welfare from a dynamic perspective. Moreover, total welfare may decrease because of a number of avoidable costs that digital piracy entails (e.g., the costs for producers to implement technical protective measures or the costs for public authorities to enforce copyrights).

Looking at the profits of copyright owners, it is an undisputed fact that they started to decrease when end-user piracy started to grow (i.e., around

1999 with the launch of Napster, a peer-to-peer file-sharing service). This was particularly acute in the music industry where physical music sales (that is to say, CDs) dropped significantly. Numerous *empirical studies* (for a survey, see Waldfogel 2012a) have tried to estimate the extent to which this decrease in sales could be attributed to digital piracy. These studies converged on the conclusion, now widely accepted, that digital piracy has “displaced” physical sales (i.e., legal purchases were substituted for, mainly, illegal downloads). However, it is also established that the estimated “displacement rate” is slightly above zero and nowhere near unity, reflecting the observation that the vast majority of goods that were illegally consumed would not have been purchased in the absence of piracy (contrary to what the recording industry would have liked the general public to believe by counting any download as a lost sale).

Very little empirical work has been devoted to the long-term effects of piracy (i.e., to dynamic efficiency considerations). One notable exception is Waldfogel (2012b), who tries to estimate the extent to which digital piracy has affected the incentives to bring forth a steady stream of valuable new products. To address this issue, he uses three different methods to assess the quality of new recorded music since Napster. The three resulting indices of music quality show no evidence of a reduction in the quality of music released since 1999; two indices even suggest an increase. One explanation could be that the digital technologies that have made piracy easier have also reduced the costs of bringing creative works to market and that the latter effect is at least as important as the former.

Reactions of Copyright Owners

In the face of digital piracy and of the reduction of sales, the first reaction of copyright owners was to try and prevent the existing business models from crashing down. As these models were relying on controlled distribution and broadcast channels, the main strategies consisted (i) in pursuing more heavily copyright infringers, (ii) in using

digital technologies as protective measures, and (iii) in lobbying for more restrictive IP laws.

The music industry started the fight against illegal downloading. In 2001, the Recording Industry Association of America (RIAA) obtained the closure of Napster, but the victory proved short-lived as a number of other file-sharing systems (such as Kazaa, LimeWire, and Morpheus) quickly replaced Napster. The industry started then a campaign of litigation against individual P2P file sharers: between 2003 and 2008, legal proceedings were opened against about 35,000 people. The software and the movie industries also engaged in similar legal battles.

As far as technical measures are concerned, a common tactic was to protect digital products through so-called digital rights management (DRM) systems, which inhibit uses of digital content not desired or intended by the content provider. DRM systems were meant to fight digital piracy but also, more generally, to manage how digital products can be used. Well-known examples of DRM systems are the Content Scrambling System (CSS) employed on film DVDs since 1996, so-called “copy-proof” CDs introduced by Bertelsman in 2002 (which could not be played on all CD players and were later abandoned), and the FairPlay system used by Apple on its iTunes Music Store. Such systems were gradually abandoned in the music industry (but are still used in other industries, such as in the case of ebooks).

Finally, lobbying efforts were met with success as stronger copyright laws were passed in a number of countries. In the United States, in 1998, the Copyright Term Extension Act extended the duration of existing copyrights by 20 years, while the Digital Millennium Copyright Act reinforced copyright protection by making it a crime to circumvent the technological measures that control access to copyrighted work. In Europe, a number of EU directives led EU member states to harmonize their national copyright laws in the first half of the 1990s; also, the European Union Copyright Directive (EUCD) of 2001 required member states to enact provisions preventing the circumvention of technical protection measures. In the late 2000s, some countries (led by France and the United Kingdom) passed

so-called three-strikes antipiracy laws, which authorize the suspension of Internet access to pirates who ignored two warnings to quit. Finally, actions were also directly taken against platforms that were hosting and sharing illegal content (the most famous cases being the shutdowns of Napster in 2001, of Megaupload in 2012, and of the Pirate Bay in 2013).

In sum, the first reaction of copyright owners in the face of digital piracy was to enforce and reinforce both the legal and technical excludability of their products. However, these measures turned out to be of little effectiveness and sometimes even counterproductive. On the one hand, technical measures were not only quickly circumvented but they also irritated legitimate consumers, thereby decreasing their willingness to pay for copyrighted products. Zhang (2013) gives an indirect proof by showing that the decision by various labels to remove DRM from their entire catalogue of music increased digital music sales by 10%. To establish this point, she compares sales of similar albums with and without DRM before and after DRM removal; her sample includes a large selection of hits and niche albums, from all four major record labels and from multiple genres.

On the other hand, a number of empirical studies have tried to assess the effectiveness of anti-piracy interventions by governments on the sales of digital products. The results obtained so far are rather mixed. For instance, two papers examine the impacts of French “three-strikes antipiracy law” (known as HADOPI law) introduced in 2009 and reach opposite conclusions: Danaher et al. (2014) find that the law caused a 20–25% increase in music sales in France, whereas Arnold et al. (2014) conclude that the law was ineffective not only in deterring individuals from engaging in digital piracy but also in reducing the intensity of illegal activity of those who did engage in piracy. Similarly, different approaches to estimate the impacts of the shutdown of Megaupload in 2012 lead to contrasting conclusions. Peukert et al. (2013) compare box office revenues before and after the shutdown for two sets of movies with matching characteristics but presenting one main difference: the first set could be accessed illegally through Megaupload, while the second set could

not. Using a quasi difference-in-differences approach, they establish that the shutdown of Megaupload did not have any positive impact on box office revenues across all movies in the sample. In contrast, Danaher and Smith (2013) exploit the fact that there exists cultural variation across countries in the degree to which Megaupload was used as a channel for piracy. They show that digital movie revenues for two studios were 6.5–8.5% higher over the 18 weeks following the shutdown (across 12 countries) than they would have been if Megaupload had continued to operate.

Even if further empirical research is called for to refine the analysis of the effectiveness of anti-piracy measures, some of the existing results suggest that digital piracy may also have some positive impacts on the copyright owners' profits, which may balance the negative "business-stealing" effect. We therefore turn to a second set of economic models that present piracy under a more favorable angle.

Further Developments: Digital Piracy May Increase Profits

A number of theoretical studies (see Peitz and Waelbroeck 2006, and the references in Belleflamme and Peitz 2012) have demonstrated the positive effects that piracy may have on the profits of copyright owners. Three mechanisms have been identified. First, illegal copies of a digital product can play a *sampling* role by attracting consumers and driving them to purchase a legitimate copy later. This argument is based on the observation that digital products are complex "experience goods"; that is, consumers do not know the exact value that they attach to particular digital products before consuming them. Buying a legitimate copy may thus appear as risky, which inevitably reduces demand. However, if an illegal copy can be accessed free of charge, consumers may learn their valuation of the product, and if the latter is large, they may want to purchase the legitimate product (which is often, as argued above, of a higher perceived quality).

The empirical results of Zhang (2013) are consistent with this theory. As we noted above, her analysis shows that the removal of DRM had a positive impact on digital music sales; yet this impact was much more pronounced for niche than for hit albums, which suggests that more flexible sharing increased sales because it lowered search costs (which are arguably larger for lower-selling than for top-selling albums).

The second mechanism originates in the fact that many digital products generate *network effects*; that is, the attraction of the product increases with the number of consumers of that product. This is so with software (the wider the community of users, the easier it is to exchange files, and the larger the supply of complementary products) or with cultural products (whose popularity increases with word of mouth). As it is the cumulated number of consumed copies that matters and not whether these copies are legitimate or not, digital piracy contributes to increase the willingness to pay for legitimate copies. An anecdotal evidence of the importance of this mechanism can be found in the reaction of one of the directors of the series "Game of Thrones" (produced by the American premium cable network HBO) when interviewed about the huge illegal downloading of the first episode of the third season (estimated to over one million times in the space of 24 h); he basically stated that the series benefits from piracy because it feeds the "cultural buzz" that allows this kind of program to "survive" (see <http://tinyurl.com/lu93q6j>).

Finally, the third mechanism, called *indirect appropriation*, resembles the second by invoking the fact that piracy can increase the demand for goods that are complementary to the pirated content; the producer is then able to capture indirectly the value that consumers attach to the pirated good. This goes, for example, for increasing ticket sales for the concert of an artist, whose popularity may be partly due to a large base of fans consuming pirated copies of this artist's songs. Mortimer et al. (2012) provide some empirical evidence along these lines; combining detailed album sales data with concert data for a sample of 1,806 artists on the period 1999–2004, they find that digital piracy reduced sales but increased live

performance revenues for small artists (the impact for large, well-known artists being negligible).

Perspectives

The presence of these potential positive impacts of piracy and the inability to preserve the existing business models drove the content industries to experiment with new solutions. Because it had been the first to be hit by digital piracy, the music industry also took the lead in terms of innovative business models. The first answer to falling CD sales was to move the distribution of music online. At the forefront was the iTunes Music Store operated by Apple, which opened in 2003. These legal online channels for digital music allowed consumers not only to find and download music as easily as via illegal channels but also to start buying individual tracks instead of being forced to buy albums. Koh et al. (2013) suggest that the latter possibility induced a new way of consuming music, which contributed to weaken the negative effect of online music piracy on physical music sales; according to their empirical assessment, it is the legal sales of online music and not digital piracy that displaced physical music sales after 2003.

In the same vein, Aguiar and Martens (2013) conclude that the online legal sales of digital music (through online stores such as iTunes or via streaming services such as Spotify) do not seem to be displaced by illegal downloading; the opposite may even occur. To establish this result, they analyze the behavior of digital music consumers on the Internet. They use direct observations of the online behavior of more than 16,000 Europeans. The main result of their analysis is that illegal downloading has no effect on legal consumption. At best, this effect is positive: a 10% increase in clicks on illegal download websites leads to an increase of 0.2% in clicks on legal purchase websites. Piracy does not induce any displacement of the legal music purchase in digital format; it might even slightly boost sales. (People in the sample have willingly accepted to be observed. This introduces two potential biases: on the one hand, it is quite likely that the “heavy

downloaders” have refused to be part of the sample; on the other hand, individuals in the sample may have changed their behavior knowing that they were observed. We must also keep in mind that in the relevant time period, while increasing, online music sales accounted for only a small fraction of the overall revenues of the music industry and that physical sales have been shown to suffer from piracy (5% in 2010 and 8% in 2011 according to IFPI).)

New business models in the music industry also offer market solutions to increase revenues from the segment of consumers with a low willingness to pay for music and with, therefore, a high disposition to digital piracy. As Waelbroeck (2013) describes it, the streaming services (such as Spotify or Deezer) are based on a “freemium” model, which combines free and premium (i.e., paying) services. The objective is to attract users with the free offering and, later, “convert” them to paying subscribers. This objective can be reached through different ways: the premium offering can include additional “mobility” (e.g., the possibility to access playlists on various devices, such as a computer, a tablet, or a smartphone), better sound quality, a wider library of titles, or the removal of ads.

Markets for information products are undergoing major changes due to technological innovations, which triggered digital piracy and, partly as a response, new business models. As exemplified above, in this changing landscape, some research suggests that consumer behavior exhibits several interesting features. Whether these features are stable over time and space is an interesting area for future research. Such an understanding is necessary to evaluate the impact of digital piracy on markets for information products and to develop successful new business models. It is also necessary to propose appropriate public policy responses.

References

- Aguiar L, Martens B (2013) Digital music consumption on the Internet: evidence from Clickstream data. Institute for Prospective Technological Studies Digital

- Economy Working Paper, 2013/04. European Commission Joint Research Centre. Seville, Spain
- Arnold MA, Darmon E, Dejean S, Pénard T (2014) Graduated response policy and the behavior of digital pirates: evidence from the French three-strike (Hadopi) law. Mimeo. University of Delaware, Newark DE
- Belleflamme P, Peitz M (2012) Digital piracy: theory. In: Peitz M, Waldfogel J (eds) *The Oxford handbook of the digital economy*. Oxford University Press, New York
- Danaher B, Smith MD (2013) Gone in 60 seconds: the impact of the Megaupload shutdown on movie sales. Mimeo. Wellesley College, Wellesley, MA
- Danaher B, Smith MD, Telang R, Chen S (2014) The effect of graduated response anti-piracy laws on music sales: evidence from an event study in France. *J Ind Econ*
- Johnson WR (1985) The economics of copying. *J Polit Econ* 93:158–174
- Koh B, Murthi BPS, Raghunathan S (2013) Shifting demand: online music piracy, physical music sales, and digital music sales. *J Organ Comput Electron Commer*
- Mortimer JH, Nosko C, Sorenson A (2012) Supply responses to digital distribution: recorded music and live performances. *Inf Econ Policy* 24:3–14
- Novos I, Waldman M (1984) The effects of increased copyright protection: an analytic approach. *J Polit Econ* 92:236–246
- Novos I, Waldman M (2013) Piracy of intellectual property: past, present, and future. *Rev Econ Res Copyr Issues* 10:1–26
- Peitz M, Waelbroeck P (2006) Why the music industry may gain from free downloading – the role of sampling. *Int J Ind Organ* 24:907–913
- Peukert C, Claussen J, Kretschmer T (2013) Piracy and movie revenues: evidence from Megaupload: a tale of the long tail? Mimeo. LMU Munich, Munich, Germany
- Waelbroeck P (2013) Digital music: economic perspectives. In: Towse R, Handke C (eds) *Handbook of the digital creative economy*. Edward Elgar, Cheltenham
- Waldfogel J (2012a) Digital piracy: empirics. In: Peitz M, Waldfogel J (eds) *The Oxford handbook of the digital economy*. Oxford University Press, New York
- Waldfogel J (2012b) Copyright protection, technological change, and the quality of new products: evidence from recorded music since Napster. *J Law Econ* 55:715–740
- Zhang L (2013) Intellectual property strategy and the long tail: evidence from the recorded music industry. Mimeo. University of Toronto, Toronto, Canada

Direct Selling: Terminology in Business

► Distance Selling and Doorstep Contracts

Director, Aaron

Robert Van Horn

Economics Department, University of Rhode Island, Kingston, RI, USA

Abstract

Aaron Director (1901–2004) is often recognized as the founder of Chicago law and economics and a leader in establishing the postwar Chicago School. This biographical essay explores Director's early life, that is, his high school and college years, and his principal contributions to the postwar Chicago School.

Biography

Aaron Director (1901–2004) is often recognized as the founder of Chicago law and economics, the pioneer “in reorienting antitrust policy along free-market lines,” and a leader in establishing the postwar Chicago School (see Bork (2004), Posner quoted in Bernstein (2004), and Samuelson (1998)). Through his work at the University of Chicago, Director had a profound influence on colleagues of his own generation, such as Edward Levi and George Stigler, and, on later luminaries, such as Lester Telser, Richard Posner, Ward Bowman, John McGee, Robert Bork, and Reuben Kessel.

Director's colleagues lauded his analytical abilities and joshed about his lack of publications. His long-time colleague, George Stigler, said, “[Most] of Aaron's articles have been published under the names of his colleagues,” (<http://chronicle.uchicago.edu/040923/obit-director.shtml>). Access Date 8/20/09) and Sam Peltzman, a colleague and

Robert Van Horn is an assistant professor of economics at University of Rhode Island. His research has primarily focused on the history of the postwar Chicago School. He is a coeditor, along with Philip Mirowski and Thomas Stapleford, of *Building Chicago Economics* (Cambridge University Press, 2011). *History of Political Economy* and *Journal of the History of the Behavioral Sciences*, among others, have published his work. More information can be found at: <http://www.uri.edu/faculty/vanhorn/index.htm>.

student of Director, observed, “His life was long, his vita was short” (2005, p. 313). Edward Levi, Director’s long-time Chicago Law School colleague, wrote: “[Director is] a self-effacing but determined scholar who has never lost the integrity of his own discipline [of economics] as he has brought this discipline to bear on the problems of [the field of law]” (1966, p. 3).

Shortly after Director’s death, prominent newspapers paid tribute to his life. *The New York Times* described Director as “a theoretician who broadly influenced scholars’ thinking about anti-trust law” (Douglas 2004). *The Washington Post* referred to Director as a “celebrated free-market economist who helped unite the fields of law and economics and mentored several generations of scholars” (Berstein 2004). The University of Chicago also celebrated his life. Its law school held a retrospective. Stephen Stigler and Sam Peltzman each gave memorial lectures, which were published in the October 2005 issue of the *Journal of Law and Economics* – a journal Director helped to found (see Stigler (2005); Peltzman (2005)).

The following pages on Director illuminate two periods of his life. First, I describe Director’s high school and college years (1918–1924) (this section mainly draws from Van Horn (2010)). This period of Director’s life helps us better appreciate his later contributions to Chicago law and economics. Next, I examine the time period during which Director made his most significant contributions to what would become Chicago law and economics (1946–1964) (this section mainly draws from Van Horn (2009, 2011) and Van Horn and Klaes (2011)). While not covered in detail below, over the next couple of years, I plan to research the interim period). I primarily focus on the evolution of Director’s own views during two projects he headed, the Free Market Study (1946–1952) and the Antitrust Project (1953–1957), and then very briefly explore how Director influenced some of the later principals of the Chicago law and economics movement through his work during the Antitrust Project.

High School and College Years

At the age of 12 or 13, Director along with his family emigrated from Charterisk, Russia, to

Portland, Oregon. (Most of what will probably ever be known of Director’s life in Russia is found in Rose Friedman’s (his sister’s) autobiography. See Friedman and Friedman (1998, pp. 2–6).) They were Jewish. Director would have quickly learned that political and social freedoms had their limits in the United States (the description of Portland in the following paragraphs draws heavily from MacColl (1979) as well as from James Breslin (1993), Mark Rothko’s biographer). Although Portland had an established reputation as a progressive city, reactionary politics dominated the city in the 1910s. In 1917, about 85% of Oregon’s residents were native born and overwhelmingly White Anglo-Saxon Protestants (WASPs) (according to MacColl, “the ‘covered wagon complex’ was still prevalent” at this time, partly because the pioneers’ recent descendents had “fought to preserve [their] purity in the face of an influx of foreign immigrants” (1979, p. 139)).

During WWI, to espouse a radical position in Portland was “quite dangerous” (Breslin 1993, p. 39). Across the nation, assimilation – or “Americanization” – became more coercive. The 1917 Espionage Act “effectively made political dissension a crime,” and the 1918 Sedition Act stated that anyone who “spoke disparagingly of the U.S. government, its Constitution, or its flag” would serve 20 years in prison (Breslin 1993, p. 39). The jingoism of WWI reached an unrivaled level in Portland, which “became the patriotic center of the Northwest,” according to Kimbark MacColl. Because of their international ties, Portland Jews, many of whom were recent immigrants, became the target of “super-patriots” and thus needed to exercise uncommon circumspection. (This is not to say that all Eastern European Jews were suspect. For example, Ben Selling, an Eastern European Jew and successful businessmen, was a community leader. During WWI, he demonstrated his patriotism by buying \$400,000 worth of Liberty Bonds (MacColl 1979, pp. 50–51).)

After WWI ended, Portland’s Jewish immigrants faced discrimination of a different nature. Rumors spread across the United States that a “diabolical, radical conspiracy” against the US

government was brewing (MacColl 1979, p. 156). In Portland, without the demonized groups of WWI (e.g., the “Huns”), the “reds” became the new demons, and “many of the ‘reds’ happened to be Jewish” (Breslin 1993, p. 40). According to MacColl, “there was actual fear, even in Oregon, that a Red revolution might begin,” patterned after the recent Bolshevik revolution. Consequently, Oregon passed strong antiradical legislation in January 1921, the month Director graduated from high school. Supporting the law, *The Oregon Journal* wrote, “We must hereafter have an Americanized America” (MacColl 1979, p. 157).

In this environment of ongoing coerced conformity, Director received his education. Director went to Lincoln High School in the spring of 1918. Here Director formed a close bond with several Jewish friends, including Mark Rothko (the famous painter).

Even though Director came from a family that was somewhat better off than the poorest Jewish families, Director, at Lincoln High School, faced tensions arising from ethnic and class differences. Of the 900 students at Lincoln, probably no more than 10% were Jewish (Breslin 1993, p. 36). The majority, WASPs, exerted considerable control at Lincoln. Overseeing the membership in social clubs and athletic teams, WASPs excluded Jews. Director’s friend, MacCoby, complained: “Anyone who has a name ending in ‘off’ or ‘ski’ is taboo and branded a Bolshevik” (quoted in Breslin 1993, p. 36). Slurs were not uncommon: “Jewish clannishness inspired bad jokes about the close friendships...between Director and MacCoby” (Breslin 1993, p. 36).

In Portland, both inside and outside of high school, Director faced a milieu fraught with discrimination. Director had views contrary to those of the establishment of Portland and many of the WASPs at Lincoln. During his senior year, when he served as editor of the January 1921 issue of *The Cardinal*, Lincoln High’s school magazine, Director responded to the majority views with caution and prudence. As editor, Director compiled an editorial section of anonymous editorials. (In the following paragraphs, unless otherwise indicated, all quotations come from the January 1921 issue of *The Cardinal*. See (CARD 1921).)

Since Director was the head editor and since the editorials express a consistent worldview sympathetic to the hardships of immigrants, it seems reasonable to assume that the worldview expressed in the editorials was consistent with Director’s own.

The editorials portrayed a world of not only overt racial prejudice and zealous patriotism but also ubiquitous political corruption and economic evils. The editorials detested the myopic view of the immigrants presented in the newspapers. They claimed that the newspapers only concentrated on “the faults” of immigrants and offered hasty generalizations. The editorials dismissed the effort to determine who was American and who was not as arbitrary classification, and they suggested that the economic evils stemmed in part from a “maladjusted industrial system” (p. 62).

According to the editorials, hope for a cure rested not with material growth and technological progress but with education. They stated: “It is the teacher who is largely responsible for the type of man and of woman that will represent America in the future” (p. 62). Education, they suggested, would alleviate social and economic maladies and remedy the ills of jingoism and bigotry. They challenged their classmates of Lincoln: “The Hungarian, the Russian, the Jew, the Pole; all can teach us something. We must derive benefit from the good that is in them, and show them the good that is in us. Thus a finer and nobler civilization may be evolved” (p. 63). Moreover, they disputed the idea that obeying the law, fighting for one’s country, and being unquestioningly loyal to one’s government were among the necessary conditions to be American. Instead, the editorials maintained that “Americanism” included “progress, reform, and the enlightenment of the human race” (p. 63).

The editorials suggested that social and economic problems could be identified by a vigilant and inquiring mind and that significant reform would only be possible if unrestrained questioning of not only creed and tradition but also public policy and social norms was allowed. The editorials extolled a vision of a heroic reformer – a broad-minded and liberal reformer who possessed the courage to confront both

political corruption and economic evils and had the ability to see the good in all races and understand the true meaning of what it means to be an American (the editorials use the term “liberal educators” to imply free-thinking educators who are without prejudice and who are in favor of progress and reform).

The editorials expressed an idealistic attitude, but not a serious program for reform. Although Director was an iconoclast and not afraid to challenge the prejudice and problems of Portland, he was no social rebel. When Director criticized the establishment, it tended to be with cautious circumspection or with harmless sarcasm; he never directly attacked an individual or an identified group of individuals.

Upon graduation in January 1921, Director left Portland for Yale University. According to Oren, “The public high school mythology that held that America’s great universities were temples of learning attracted the Eastern [Europeans Jews] to ‘worship’ at Yale” (1985, pp. 27–28). The spires of the great university beckoned Director. Ironically, however, the university system that inspired hope in the Eastern Jewish community was slowly excluding them. Director and Rothko entered Yale at a time of heightened anti-Semitism in the nation. The Ivies proved to be one of the primary anti-Semitic battlegrounds in the country, and Yale was one of the least friendly places to Jews.

In the 1921–1922 academic year, Yale reconsidered its admissions policy in order to decrease the percentage of Jews enrolled. This most likely resulted in Director losing his tuition scholarship after his first academic year. Around Yale’s campus, Director also faced discrimination. The Protestant upper classes controlled the social clubs, athletic teams, fraternities, and senior societies, all of which tended to exclude Jews. Many of Yale’s social groups favored graduates of prep schools and students from wealthy families.

Because Director received very little financial help from his family and needed to work his way through, Director, like Rothko, most probably waited tables in the dining hall when he started at Yale. According to Oren, “Students identified as poor, students who had to perform a ‘low class’

job such as waiting on tables. . . landed in almost inescapable social damage” (1985, p. 68).

Director decided to pursue a Ph.B. in “progressive politics.” Putting his major to use before leaving Yale and fulfilling his high school ambitions to be a newspaper editor, Director attacked Yale’s establishment in an underground newspaper. In the spring of 1923, Director along with Mark Rothko and Simon Whitney produced *The Yale Saturday Evening Pest* – an underground newspaper. The *Pest* was a stinging reaction to life at Yale and society at large, and, according to Oren, its “depictions of Yale life” were by and large “fair representations of reality” (p. 90). The *Pest* sheds light on Director’s philosophy of life at this time.

Disillusioned by the reality of Yale, Director and the other editors claimed that they saw the empty lives of Yale’s undergraduates for what they were. In an issue entitled, “False Idols,” the *Pest* observed, “The Yale undergraduate is an idolater. He is as senseless as [anyone] who prays to a totem pole, or [anyone] who mumbles in fear before a meteorite. At least the meteorite has come down from the sky. . .” (SEP, March 17, 1923). The idols of Yale included: athletics, extra-curricular success, social success, the opinion of the majority, and grades. The first received by far the most criticism in all the issues of the *Pest*. The “god” of athletics was “low-browed, but husky, with muscular arms and long legs whose pedal extremities carry a powerful kick. We talk of erecting a statue of him, in the shape, appropriately, of a bulldog.” The worship of this god, which demanded greater veneration at Yale than education, involved participation for some (SEP, February 23, 1923), but, according to the *Pest*, it meant “lung athletics” for the majority in the “bleacher seats.” In the spirit of Veblen’s *Higher Learning in America*, the *Pest* railed, “The present athletic system injures our bodies and narrows our minds, making us insufferable bores to any intelligent man” (SEP, March 17, 1923). About the idols at Yale, the editors of the *Pest*, like Old Testament prophets, proclaimed: “False gods! Idols of clay!”

The *Pest* identified the cause of the “fundamental evil” that afflicted Yale to be “. . . the

rusty condition in which the mass of undergraduates have allowed their minds to molder” (SEP, March 17, 1923). Hope for a cure could not be found by turning to the powers that be at Yale. The *Pest* maintained that universities were ultimately run by merchants, and this merely contributed to the large population of unthinking undergraduates. For the *Pest*, because Yale failed to educate thinking men who could see through bigotry and jingoism, Yale was partly to blame for anti-Semitism and other forms of prejudice and racism in the United States.

Despite Yale being part of the cause of larger social problems, the *Pest* did not prescribe reforms for Yale. Instead, it sought to disillusion Yale undergraduates, causing them to see their own unthinking for what it was. For the *Pest*, “destructive criticism” was the key to freedom from unthinking – hence its masthead with the provocative slogan: “The Beginning of Doubt is the Beginning of Wisdom.”

Destructive criticism promised to show unthinking undergraduates their “uselessness in the world” and the “emptiness of their ambitions” and thereby be life changing (SEP, March 3, 1923). It also promised to be informative. Destructive criticism – if it was thoughtful, sincere, and purposeful – helped others to see the root of social and economic problems and see how bigotry and racism prevented progress. Per the *Pest*, change came from the ground up, starting with the individual. In sum, the *Pest* represented Director’s high school vision of the heroic reformer; Director and the other editors saw themselves as broad-minded educators who challenged the root of economic evils and racism and thereby awakened in others the capacity and vision to remedy these serious social and economic problems.

Yale was a formative place in Director’s life. The entrenched elitism to which Director reacted and the “unthinking” life at Yale and in society at large that Director lambasted shaped his outlook on life. Director had come of age as a skeptic and an individualist. He rejected governing structures as instruments of the established business class that could not be trusted to implement meaningful reform. In doing so, he foreshadowed his distrust

of government intervention into the economy that informed his work in economics in the 1950s. While at Yale, rather than supporting administrative bodies and student organizations, Director championed individuals like himself or an elite group of individuals like the editors of the *Pest* who could transcend the “pestilence of unthinking” through relentless questioning of “creeds and traditions.” Director praised those who, like the heroic educator, could rise to the occasion against the powers that be and challenge prejudice and social injustice.

Director remained a reformer for decades to come. His vision of the heroic educator led him to head the Portland Labor College (1925–1927). Later, it arguably led him to the University of Chicago in 1946.

The Roots of Chicago Law and Economics

Although Director made his seminal contributions to Chicago law and economics while at Chicago from 1946 to 1964, he studied at Chicago from 1927 to 1934, pursuing a Ph.D. Initially, Director worked with Paul Douglas; they authored *The Problem of Unemployment* in 1931. However, by 1932, Director, according to Douglas, fell under Knight’s influence (VPML, Douglas to Frank H. Knight, 5 January 1935, Box 79, folder “Chicago Dept. of Econ., Douglas & Knight”). Thereafter, Director gravitated toward Henry Simons, who became, according to Ronald Coase (1998, p. 602), his “best friend” and “considerably influenced Director’s views.” In 1933, Director published a pamphlet, *Economics of Technocracy*, demonstrating his affinity for price theory for the first time.

In 1934, when the University of Chicago refused to renew his teaching contract, Director went to work in the Treasury Department in Washington, DC (VPML, H. A. Millis to Viner, 31 January 1934, Box 79, folder “Chicago University Department of Economics, Millis”). Then, in 1937, Director traveled to England to conduct research for his dissertation on the quantitative history of the Bank of England. However, the Bank unexpectedly thwarted his efforts, and Director never completed his thesis. While in England, Director became associated with Arnold

Plant and Lionel Robbins and befriended Friedrich Hayek. After attending one of Hayek's seminars, Director considered Hayek his teacher. Once WWII commenced, Director returned to Washington D.C. and worked for many different agencies. For example, he joined the Brookings Institution, where he wrote a book with C. O. Hardy in 1940 entitled *Wartime Control of Prices*.

While in Washington, Director became one of Hayek's political allies and supported Hayek's intellectual crusade to countervail collectivism (i.e., Keynesianism, institutional reformism, and socialism). He helped to persuade the University of Chicago Press to publish Hayek's *The Road to Serfdom* and wrote a laudatory review of it (Director 1945).

Near the end of the war in 1945, Hayek and Simons encouraged Director to return to Chicago to lead a project that would be called the Free Market Study (FMS). The FMS was primarily the product of the efforts of Hayek. In April 1945, when on tour in the United States promoting his recently published *The Road to Serfdom*, Hayek met with Harold Luhnow, head of the Volker Fund – a Kansas City corporation heavily involved in right wing funding in the postwar period. Luhnow wanted Hayek to write an American version of *Road* and offered him money to do so. The two men agreed that the Volker Fund would finance an investigation of the legal foundations of capitalism and that a product of this investigation would be *The American Road to Serfdom*. The two also agreed that Hayek could outsource this investigation.

Hayek convinced the Volker Fund to allow him to subcontract the project to Simons and Director. Since Simons viewed the liberal doctrine as withering and the collectivist doctrine (i.e., socialism and institutional reformism) as burgeoning, he envisioned the project as a way to reinvigorate the liberal doctrine in order to countervail collectivist doctrine. Simons drew up two memoranda, Memorandum I and Memorandum II, the latter being a concise, executive version of the former (SPRL, box 8, file 9). Simons wholeheartedly endorsed Aaron Director as the leader of this project. Simons feared that without an organized

effort to revive liberalism, it would “be lost,” and he believed that Director's leadership would help to engender a liberal stronghold at the University of Chicago in the immediate post-WWII period (SPRL, Memorandum I, undated, box 8, file 9).

Bringing Director back to the University of Chicago meant a great deal to Simons. Indeed, in 1939, Simons had written: “[I]n spite of my efforts and good intentions of other people, I have been, *qua* economist, alone since Aaron left. Certainly I am worth more to the University with Aaron around than without him” (quoted in Van Horn (forthcoming)). Director responded favorably to the proposed project. He also drafted a proposal for the project, which he called “the Free Market Study.” Director's plan delineated the benefits and limitations of the free market and enumerated the departures from the free market at the close of WWII – including: barriers to entry (such as patents and tariffs) and government controls (such as price controls). In keeping with Hayek and Luhnow's agreement, Director also listed numerous policies that needed to be examined to return to a free-market economy, including antitrust policy and corporate policy. In many ways, Director's list echoed Simons' *Positive Program*; for example, Director called for limitations on corporate size and for federal incorporation to be required.

After many trials that have been detailed elsewhere, by July 1946, Director agreed to head the FMS, which would be housed at the Chicago Law School. Director, however, would have to return to Chicago and lead the project without his dear friend Simons; Simons committed suicide on June 19, 1946 (see Van Horn (forthcoming)). Indeed, part of the reason Director returned to Chicago was to carry on the legacy of Simons.

In October 1946, Director assumed leadership of the FMS (for background and more information about the study, see Van Horn and Mirowski (2009), and for more information on the Free Market Study and the Antitrust Project, see Van Horn (2009) and Van Horn and Klaes (2011)). Director and the study's members (Milton Friedman, Frank Knight, Edward Levi, Garfield Cox, and Wilbur Katz) convened regularly in order to discuss and debate. The FMS's task had a sense of urgency because of the perceived strength of

collectivist forces. Hayek conveyed this urgency when he wrote: “The intellectual revival of liberalism is already under way. . . . Will it be in time?” (1949, p. 433).

At the second meeting of the study, Director distributed a research proposal entitled: “A Program of Factual Research into Questions Basic to the Formulation of a Liberal Economic Policy.” As Director’s title suggests, the topics the study decided to investigate were not chosen purely because of theoretical concerns: political necessity was a factor. By empirically investigating the facts taken for granted by both liberals and their opponents, Director believed it would be possible to develop a more robust liberal policy to counter collectivism and thereby bring about policy changes in the United States.

The FMS decided to mainly concentrate its efforts on issues concerning industrial monopoly and corporations. It hired researchers, for example, Warren Nutter (1951), to do empirical work on the issue of industrial monopoly and brought like-minded visiting scholars to Chicago. During the early years of the FMS (1946–49), Director, Friedman, and others were uncertain how to reconstitute liberalism in order to best combat socialism and other forms of collectivism. Like Director during his 1947 Mont Pelerin address, they, in many ways, echoed the beliefs of classical liberals, expressing concerns about concentrations of power, including industrial monopoly.

One year after the FMS commenced, Director agreed to be a charter member of the Mont Pelerin Society in 1947. Led by Friedrich Hayek, liberals turned to organizing an intellectual movement, the Mont Pelerin Society, a transnational institutional project that sought to reinvent a liberalism that had some prospect of challenging collectivist doctrines ascendant in the immediate postwar period. The society enabled its members – liberals from America, many who represented the Chicago School, and Europe – to debate and offer each other mutual support. A crucial objective of the society was to understand the legal foundations necessary for effective competition – that is, how to create a competitive order. The society and the FMS were joined at the hip at birth (Van Horn and Mirowski 2009). The fact that both sought to

investigate a number of legal and policy areas in order to move toward effective competition is just one indication of their conjoined birth. Director’s involvement in the society indicates his determination to see the liberal doctrine reconstituted. In the opening session of the 1947 inaugural meeting, which was on the competitive order, Director gave one of the addresses (at the first meeting, its members, besides debating the issue of “‘Free’ Enterprise or Competitive Order,” debated “The Future of Germany,” “The Problems and Chances of European Federation,” “Liberalism and Christianity,” and “Modern Historiography and Political Education”). His address sheds further light on his views regarding concentrations of business power at this time.

In his address, Director claimed that authority had either supplanted individualism or ominously threatened to do so. Director maintained that state intervention had nearly destroyed the competitive order because liberals lacked solutions to resolving conflicts between social interests and the results of free enterprise. As a remedy Director advocated for a reconstituted liberalism. (The following paragraphs draw from the records of the 1947 Mont Pelerin meeting, Liberal Archives, Ghent, Belgium. See MPS1947LA.)

In keeping with Simons, Director steadfastly believed that the liberal doctrine needed, above all else, to champion freedom by promoting the dispersion of power necessary for a competitive order. Notably, Director observed that a substantial amount of monopoly power existed in the economy. To create a viable competitive order, Director, like Simons, advocated state action on three fronts: (1) preventing private monopoly, (2) controlling combinations among workers and businesses, and (3) providing monetary stability. Given the focus of the FMS and given Director’s contributions to law and economics partly stemmed from his work on antitrust law, we shall restrict ourselves to addressing only (1) and (2).

Regarding industrial monopoly, although Director maintained that international trade normally provided a check on industrial monopoly, he admonished that this was an insufficient check. Indeed, Director blamed England’s

overconfidence in the ability of international trade to eliminate business monopoly as a significant reason for the relatively large number of business monopolies in England. Director expressed qualified praise for the enforcement of American antitrust law and suggested that more vigorous antitrust was necessary to address the substantial amount of monopoly power in America. Additionally, for Director, policy reform needed to target patent law and policy measures needed to address the inequality of income and inequality of wealth that stemmed from monopoly power.

Director asserted that radical corporate reform also needed to be undertaken. He maintained:

The unlimited power of corporations must be removed. Excessive size can be challenged through the prohibition of corporate ownership of other corporations, through the elimination of interlocking directorates, through a limitation of the scope of activity of corporations, through increased control of enterprise by property owners and perhaps too through a direct limitation of the size of corporate enterprise. (Quoted in Van Horn (2009))

Like Director during his 1947 Mont Pelerin address, from 1946 to 1949, the members of the FMS echoed in many ways the classical liberal tradition by expressing concerns about concentrations of power. However, during the latter half of the project (1950–1952), a reconstituted liberalism emerged. This marked a crucial shift in attitude toward concentrations of business power, which dramatically changed the way both corporations and patents would be understood by the postwar Chicago School. By 1950, business monopoly was no longer viewed as a relatively ubiquitous and powerful phenomenon in the United States; rather it was seen as relatively un-pervasive and benign because the “corrosive effects” of competition would always and eventually undermine it (Director 1950). By 1951, large corporations were no longer considered harmful to competition because of their market power, but rather another aspect of a competitive market (Director 1951). Consequently, for Director, concentrated markets tended to be efficient, regardless of the size of business.

As the FMS ended, the Antitrust Project began. Director headed the project and Edward Levi

assisted. The members included John McGee, William Letwin, Robert Bork, and Ward Bowman. The Antitrust Project focused on issues of monopoly, select areas of antitrust law, and the history of the Sherman Act (for a list of the articles and books that the Antitrust Project published and that it caused to be published, see Priest (2005, pp. 353–54)). Under Director, the Antitrust Project produced a prodigious amount of scholarship; the topics included: tying arrangements (Bowman 1957), predatory pricing (McGee 1958), trade regulation (Director and Levi 1956), and the Sherman Act (Bork 1954). The Antitrust Project investigated these topics in the light of the conclusions of the FMS. Moreover, in the spirit of the FMS’s attempt to influence policy, it investigated these topics with a critical eye toward United States antitrust law precedent, and many of the conclusions of the Antitrust Project contravened the conclusions of the courts. In 1954, for instance, Bork, in contrast to Director’s classical liberal concern in 1947 about the excessive size of corporations and their concentrated economic power, maintained that “[vertical mergers added] nothing to monopoly power” (p. 195). This Chicago reconstituted-liberal position suggested, therefore, that vertical mergers should always be legal. Consequently, Bork suggested that one aspect of antitrust law precedent, which required an investigation of motives of a vertical merger in order to make a determination of its legality, was not only extraneous but also erroneous.

Bork’s article fell under the Antitrust Project’s umbrella article and manifesto, “Trade Regulation,” by Director and Levi, which they published in 1956. In this article, Director and Levi demonstrated skepticism about the extension of monopoly power through the use of exclusionary practices, such as tying arrangements, and a concomitant disdain for adjudication or legislation that regarded these practices as *per se* deleterious or as *per se* illegal (Packard 1963, p. 56). Director and Levi suggested that exclusionary practices were no worse than harmless price discrimination and served as either competitive tactics equally available to all businesses or means of maximizing returns on an established market position. Thus, Director and Levi maintained

that when the courts deemed it necessary to consider the legality of an exclusionary practice, they should utilize a rule of reason analysis, not a per se approach.

It is important to appreciate that even though Director published little during the course of the Antitrust Project, he substantially influenced its members through his role as an educator. Most of the members of the Antitrust Project later acknowledged Director's substantial influence on their views and his import for the development of their scholarship. For example, in his book *Patent and Antitrust Law*, Bowman reported: "The analysis on which the conclusions of this work are based is derived from years of association with colleagues, from whose oral and written contributions I have long borrowed heavily. . . . I am especially indebted to professors Aaron Director, Robert H. Bork, [and] John S. McGee" (1973, p. vii).

It is also important to acknowledge that the Antitrust Project served to train and educate members of the generation that would help to lead the law and economics movement in the United States. Director educated lawyers, including Ward Bowman and Robert Bork, who would be at the vanguard of the Chicago law and economics movement in the 1960s and 1970s. Many of these lawyers acquired jobs in other law schools. In the case of Bork and Bowman, they both found positions at the Yale Law School. While at Yale, Bork and Bowman played a crucial role in the Chicago law and economics movement in the 1960s and 1970s. Along with Richard Posner and other Chicago lawyer-economists, they helped Chicago law and economics become one of the dominant schools of jurisprudence in the 1980s (for a detailed look at the rise of Chicago law and economics, see Duxbury (1995) and Teles (2008)).

Until he left the Chicago Law School in 1965, Director taught antitrust law and price theory in the law school, founded the *Journal of Law and Economics*, engaged law school faculty – especially Edward Levi and Walter Blum – and mentored graduate students of economics. In 1965, Director moved to Los Altos, California. He worked at the Hoover Institute and Stanford University, from

where he would retire. Director died September 11, 2004, at the age of 102.

Impact and Legacy

In leading the FMS, Director, alongside a small group of like-minded liberals, critically questioned a number of the fundamental tenets of the classical liberal tradition in order to create a more robust liberal doctrine to counter the intellectual influence of collectivism. In leading the Antitrust Project in the 1950s, Director sought to reorient antitrust policy along free-market lines. To do so, he questioned and analyzed some of the fundamental tenets of status quo antitrust policy in the light of the reconstituted liberalism that emerged from the efforts of the FMS. In doing so, Director oversaw the emergence of a prodigious amount of scholarship in the 1950s and mentored many of the later leaders of the Chicago law and economics movement.

The perspectives Director developed during his youth – his belief in the individual or an elite group of individuals as the catalyst for social change, his distrust of governing structures, his faith in the heroic educator, and his adherence to the skeptical philosophy of H. L. Mencken – were not abandoned after he studied Chicago price theory and came to champion liberalism. Highlighting the importance that "destructive criticism" played in Director's work at the Chicago Law School, one Chicago Law School graduate commented that Director "washed all preexisting" antitrust doctrine in "cynical acid" (Liebmann 2005, p. 18). Through his leadership of the Free Market Study, the Antitrust Project, and the Law and Economics Program, Director was a principal of the postwar Chicago School and the founder of Chicago law and economics, not because of his ability to publish and propagate ideas like his brother-in-law Milton Friedman, but because he played an indispensable role educating the next generation of luminaries in the Chicago School. According to one former law student, "To me [Director's] most important contribution... is much less tangible... He developed and reinforced in his students a state of mind without

which much of what they have done would not have been done or would have been done less well” (quoted in Kitch 1983, p. 184). The perspectives of Director’s youth motivated and empowered his postwar efforts at Chicago.

References

Archival Sources

- CARD *The Cardinal*, January 1921
 MPS1947LA Records of the 1947 meeting, Mont Pèlerin Society, Liberaal Archief, Ghent Belgium
 SEP *Yale Saturday Evening Post*
 SPRL Henry Simons Papers, Regenstein Library, University of Chicago
 VPML Jacob Viner Papers, Mudd Library, Princeton University

Other Sources

- Berstein A (2004) Aaron director dies at 102. Washington Post (13 Sept), B04
 Bork R (1954) Vertical integration and the Sherman Act: the legal history of an economic misconception. *Univ Chicago Law Rev* 22:157–201
 Bork RH (2004) Chicago’s true godfather of law and economics. *Wall St J* (May 3):A17
 Bowman WS (1957) Tying arrangements and the leverage problems. *Yale Law Rev* 67:19
 Bowman WS (1973) Patent and antitrust law. University of Chicago Press, Chicago
 Breslin JEB (1993) Mark Rothko: a biography. The University of Chicago Press, Chicago
 Coase R (1998) Aaron Director. In: Newman P (ed) *The new palgrave dictionary of economics and the law*. Macmillan, New York
 Director A (1945) Review of “the road to serfdom” by Friedrich A. Hayek. *Am Econ Rev* 35:173–75
 Director A (1950) Review of Charles E. Lindblom, *Unions and Capitalism*. *Univ Chicago Law Rev* 18:164–167
 Director A (1951) Conference on corporation law and finance, vol 8. University of Chicago Law School, Chicago, New York
 Director A, Levi E (1956) Trade regulation. *Northwest Univ Law Rev* 51:281–296
 Douglas M (2004) Aaron Director, economist, dies at 102. *New York Times*. 16 Sept 2004, p. B10
 Duxbury N (1995) *Patterns of American jurisprudence*. Oxford University Press, New York
 Friedman M, Friedman R (1998) *Two lucky people*. University of Chicago Press, Chicago
 Hayek FA (1949) The intellectuals and socialism. *Univ Chicago Law Rev* 16(3):417–433
 Karabel J (2006) *The chosen: the hidden history of admission and exclusion at Harvard, Yale, and Princeton*. Houghton Mifflin, New York
 Kitch EW (ed) (1983) *The fire of truth: a remembrance of law and economics at Chicago, 1932–1970*. *J Law Econ* 26:163–234
 Levi EH (1966) Aaron director and the study of law and economics. *J Law Econ* 9(1):3
 Liebmann GW (2005) *The common law tradition*. Transaction Publishers, London
 MacColl EK (1979) *The growth of a city: power and politics in Portland, Oregon 1915 to 1950*. The Georgian Press, Portland
 McGee JS (1958) Predatory price cutting: the standard oil (N.J.) case. *J Law Econ* 9:135
 Nutter GW (1951) *The extent of enterprise monopoly in the United States, 1899–1939*. University of Chicago Press, Chicago
 Oren DA (1985) *Joining the club: a history of Jews and Yale*. Yale University Press, New Haven
 Packard HL (1963) *The state of research in antitrust law*. Walter E. Meyer Research Institute of Law, New Haven
 Peltzman S (2005) Aaron Director’s influence on antitrust policy. *J Law Econ* 48(2):313–330
 Pierson GW (1955) *Yale: the University College, 1921–1937*. Yale University Press, New Haven
 Priest GL (2005) The rise of law and economics: a memoir of the early years. In: Parisi F, Rowley CK (eds) *The origins of law and economics*. Locke Institute, Northampton
 Samuelson PA (1998) How foundations came to be. *J Econ Lit* 36(3):1375–86
 Stigler S (2005) Aaron director remembered. *J Law Econ* 48(2):307–311
 Teles S (2008) *The rise of the conservative legal movement*. Princeton University Press, Princeton
 Van Horn R (2009) Reinventing monopoly and the role of corporations. In: Mirowski P, Plehwe D (eds) *The road from Mont Pelerin*. Harvard University Press, Cambridge, MA
 Van Horn R (2010) Harry Aaron Director: the coming of age of a reformer skeptic (1914–1924). *Hist Polit Econ* 42(4):601–630
 Van Horn R (2011) Jacob Viner’s critique of Chicago neoliberalism. In: Van Horn R, Mirowski P, Stapleford T (eds) *Building Chicago economics*. Cambridge University Press, Cambridge
 Van Horn R Forthcoming (2014) A note on Henry Simon’s Death. *Hist Polit Econ* 46(3)
 Van Horn R, Klaes M (2011) Chicago neoliberalism versus cowles planning. *J Hist Behav Sci* 47(3):302–321
 Van Horn R, Mirowski P (2009) The rise of the Chicago school of economics and the birth of neoliberalism. In: Mirowski P, Plehwe D (eds) *The road from Mont Pelerin*. Harvard University Press, Cambridge, MA

Direct correspondence to Robert Van Horn, University of Rhode Island, Economics Department: rvanhorn@mail.uri.edu. I am indebted to University of Rhode Island Seed Grant for research

support. I would like to thank Monica Van Horn for helpful editorial comments. Note that portions of this manuscript have been adapted and reprinted with permission from “Harry Aaron Director: The Coming of Age of a Reformer Skeptic (1914–1924).” *History of Political Economy*. 42(4): 601–630, Duke University Press, Copyright © 2010

Discrete Choice Models

Patrice Bougette

Department of Economics, Université Côte d’Azur, CNRS, GREDEG, Nice, France

Abstract

Discrete choice models (DCM) have been essential in modeling agents' decision-making behavior. Empirical analysis in law and economics uses therefore such a method. This essay summarizes the definition and the different types of DCMs.

Synonyms

[Qualitative choice models](#)

Definition

Discrete choice models (DCM) describe the behavior of individuals' choices among discrete available alternatives. Decision makers can be consumers, firms, authorities, and any other decision-making unit, and the alternatives represent competing products, courses of action, or any other options over which choices must be made (Train 2009). Examples are decisions about buying a new automobile (individual), allowing a merger (competition authority), entering a new market (a firm), marital status, family size, transport choice, and so on (e.g., Henscher et al., 2005).

Random Utility Models (RUM)

Economics and psychology models often explain observed choices by using a random utility function. The utility of a specific choice can be interpreted as the relative expression of her preferences with respect to the other alternatives available within a finite choice set. The individual is assumed to choose the alternative for which the associated utility is the highest. However, as certain components of utilities are not known to the researcher with certainty, they are therefore treated as random variables. When the utility function contains a random component, the individual choice behavior becomes a probabilistic process (for a historical perspective, e.g., McFadden, 2001).

The random utility function of individual i for choice j can be decomposed into deterministic and stochastic components:

$$U_{ij} = V_{ij} + \varepsilon_{ij},$$

where V_{ij} is a deterministic utility function (i.e., contains all the measured characteristics), generally assumed linear in the explanatory variables, and ε_{ij} is an unobserved random variable that captures the factors that affect utility that are not included in V_{ij} (the error term).

Different Classes of DCMs

Different assumptions on the distribution of the errors – ε_{ij} – give birth to different classes of DCMs. First, logit and nested logit are derived under the assumption that the unobserved portion of utility ε_{ij} is independently and identically distributed (iid) with the extreme value distribution and with a type of generalized extreme value, respectively (McFadden, 1974). Nested logits include different levels of choice.

One important property of the general logit model is known as the *Independence from Irrelevant Alternatives* (IIA). The ratio of any two probabilities depends exclusively on the attributes of the two alternatives concerned and is therefore independent of the number and nature of all other alternatives that are simultaneously considered. For instance, the introduction of a new alternative alters the probabilities of all outcomes in

the same proportion, leaving the ratios unchanged. Thus, the IIA property implies proportional substitution. One limitation is that in reality substitution is rarely proportional. The Hausman-McFadden and the Small-Hsiao tests allow checking whether the IIA property is violated.

Second, probit models are derived under the assumption that ε_{ij} follows a normal distribution. Third, mixed logit models are based on the assumption that the unobserved portion of utility consists of a part that follows any distribution specified by the researcher ε_{ij} plus a part that is iid extreme value.

The dependent variable assumes discrete values. The simplest form is when the dependent variable is binary. In this case, DCMs will be called binary logit or probit models. When the dependent variable has more than two alternative choices, one refers to multinomial type of DCMs (e.g., multinomial logit or probit models). Depending on the nature of the dependent variable, multinomial models can be ordered (e.g., product categories), nonordered (e.g., transport mode choice), or sequential. Last but not least, Tobit models are variants of DCMs in which the dependent variable is censored. In other words, the variable is observed in only some of the ranges (Maddala 1983).

References

- Maddala GS (1983) Limited-dependent and qualitative variables in econometrics. Econometric society monographs No 3. Cambridge University Press, Cambridge, MA
- Hensher D, Rose J, Greene W (2005) Applied Choice Analysis: A Primer. Cambridge University Press, Cambridge, MA
- McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers of econometrics. Academic, New York, pp 105–142
- McFadden D (2001) Economic choices. Am Econ Rev 91:351–378. Nobel Prize Lecture
- Train K (2009) Discrete choice methods with simulation, 2nd edn. Cambridge University Press, Cambridge, MA

Further Reading

- Anderson SP, de Palma A, Thisse JF (1992) Discrete choice theory of product differentiation. The MIT Press, Cambridge, MA

Distance Contract

► Distance Selling and Doorstep Contracts

Distance Selling and Doorstep Contracts

Sven Hoeppe

Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium

Guest Researcher, Max Planck Institute for Research on Collective Goods, Bonn, Germany

Abstract

Distance-selling and off-premises contracts are two major ways in which consumers and sellers interact. Law and economics research has established that these interactions potentially suffer from market power of sellers, from both ex-ante and ex-post information asymmetries, and from consumer bounded rationality. The most promising tool analysed and advocated by law and economics scholars is a cooling-off period coupled with a right of the consumer to withdraw from the contract. This entry surveys law and economics research on these concerns. Interestingly, relevant questions to this line of research remain, which have been brought to attention mainly by insights from behavioral economics. To exemplify and inspire further research along these lines, this entry discusses potentially perverse incentives created by withdrawal rights and the impact of fairness concerns on the consumer choice to withdraw.

Synonyms

Direct selling; terminology in business; Distance contract; Distance selling contract; Doorstep contract; Off-premises contract

Definitions

A **distance selling** contract is a sales contract concluded between a consumer acquiring some good or service and a business partner selling it without the simultaneous physical presence of either the consumer or the professional seller and with exclusive use of one or more means of distance communication until the contract is concluded. Usually the law also requires the professional seller to have employed an organized distance sales mechanism.

A **doorstep contract** is any contract between a consumer acquiring some good or service and a business partner selling it, either:

- (1) Concluded in the simultaneous physical presence of a professional seller and consumer but at a location that is not the business premises of the professional seller or
- (2) Concluded on the business premises of the professional seller (or through any means of distance communication) immediately after the consumer was personally and individually addressed at a location that is not the business premises of the professional seller in the simultaneous physical presence of the professional seller and consumer or
- (3) Concluded during an excursion organized by the professional trader for the purpose or to the effect of promoting and selling the goods or services to the consumer

Introduction

Doctrinal lawyers of consumer law tend to opine that a consumer is in an inferior bargaining position compared to professional seller of a good or service (e.g., Bourgoignie 1992; Weatherill 2005; Loos 2009; Eidenmüller 2011). Is the consumer not less skilled, less knowledgeable, economically much more fragile, and thus equipped with much less bargaining power? The fear that consumers will be exploited if they are not legally protected also resonates in European consumer law (cf. Hoepfner 2012).

This entry surveys, introduces, and reviews law and economics scholarship on two key elements of consumer protection law: distance selling and off-premises – or doorstep – contracts. Law and economics contributions on the topic can be distinguished into two streams. One analyzes the relationship of the contracting parties. The other develops legal responses to specific structures in this relationship. It is clear that the results of the latter depend on insights of the former.

Thus, also the structure of this entry is given. After the economic characteristics of the relationship between consumers and sellers have been introduced, this entry devotes sufficient space to a discussion about the arguably most important tool that is widely used to address the perceived imbalance between consumers and sellers, namely, the right to withdraw from a distance or doorstep contract within a specified amount of time called a “cooling-off period” that, if granted, typically last three full weeks in the USA and two weeks in Europe. Moreover, this entry will also address some concerns about this consumer protection instrument.

The Transaction

Distance selling and doorstep sales have undisputed advantages, mainly a reduction of distribution costs and dissemination of product information across the market, thereby facilitating welfare-increasing transactions. However, these advantages are curbed by specific drawbacks that also result from the nature of distance selling and doorstep transactions (cf. Rekaiti and Van den Bergh 2000; Eidenmüller 2011, Hoepfner 2012).

Information Problems

If the theoretical assumption of complete information of contracting parties may not be satisfied, allocative efficiency is endangered. Therefore, one main problem of distance and doorstep transactions is that consumers are “in the dark” (Dickie 1998, p. 217) regarding both the seller and the quality of the good. The seller may be unreliable or even fraudulent. The good may possess

characteristics that are worse from what the consumer expected.

In other words, there is a multidimensional, *ex ante* information asymmetry between the contracting parties. Without further information – e.g., through inspection of the product – it is very difficult for the buyer to form beliefs about the seller's reliability and product quality. If the product is rich in experience and/or credence (trust) dimensions, even inspection could not ameliorate the information problem because they are (nearly) impossible to observe (Rekaiti and Van den Bergh 2000). *Ex ante* information asymmetries – i.e., hidden characteristics – are especially troublesome for the pre-contractual stage. The economic conjecture is that inferior goods crowd out superior ones through adverse selection. In the limit, only the worst quality goods will remain in the market (Akerlof 1970).

Market Power

Although there are often alternative suppliers or close substitutes readily available for products sold in distance and door-to-door transactions, a specific concern of these transactions is temporary market power. This special twist is caused by so-called situational monopolies (cf. Rekaiti and Van den Bergh 2000, Hoeppepner 2012). "Situational monopolies arise out of particular circumstances surrounding particular exchanges, where this transaction-specific market power is exploited opportunistically" (Trebilcock 1993, p. 101) to extract supracompetitive prices.

How can temporary market power be established although market structure, objectively, is not conducive for concentration? Lele (2007, p. 45) posits that monopolies constitute "an ownable space for a useful period of time." This emphasizes the importance of marketing techniques that temporarily may convince consumers that it will be more costly to engage in alternative search. In door-to-door transactions, crucially important elements are high-pressure sales techniques that lock in the consumer (Hoeppepner 2012). There is a plethora of these techniques. Very familiar to everyone, for instance, may be the buy-on-deadline pattern. This pattern uses the

idea that the consumer will lose out if she does not close the deal right away and is therefore designed to rush the buyer into a decision without proper consideration. It can include anything from first-time-only benefits that accompany the sale, to predictions by the salesperson that the price of the good will drastically increase tomorrow, to the sudden surprise that the good is the last in stock. In regard to distance selling, for example, Rekaiti and Van den Bergh (2000) mention techniques such as an advertisement that claims product uniqueness or vending mechanisms that induce unreflective buying that may induce temporary market power.

Nonrational Behavior

Rekaiti and Van den Bergh (2000) also take issue with the standard assumption of consumer rationality. Consumer preferences may, in contrast to standard economic theory, not be stable over time. To substantiate this idea, the researchers elaborate on possibly inconsistent intertemporal preferences (e.g., Frederick et al. 2002), psychological costs of regret contingencies (cf. Goetz and Scott 1980; Coricelli et al. 2005), and reduced risk perception leading to less deliberative decision-making.

Hoeppepner (2012) delves more deeply into psychological research and the underlying processes that may likely influence consumer decision-making. He adds to the discussion the fact that the ability to predict how the satisfaction of one's decision evolves over time is crucial for decisions to align with the rationality assumptions (cf. Loewenstein and Schkade 1999). It turns out, however, that individual predictions of one's future affect – hence also preferences, decisions, and behavior – are not very accurate. Often these assessments are biased toward initial salient impressions that are more readily available. People who do affective forecasts systematically have it wrong (cf. Gilbert and Wilson 2007). Several sources of error are involved in these predictions. Important for buying decisions in distance and doorstep transactions are the so-called hot/cold empathy gap and the projection bias. Empathy gaps occur when present and future predictions of future affect go hand in hand with different

states of arousal. For instance, a person, who is now excited about a new product and makes a prediction about her future affect concerning the product, is likely to fail to account for satiation effects. Therefore, the person will overestimate the product's impact on his utility. This will result in projection bias: the individual tendency to falsely extrapolate current preferences into the future. This interplay of empathy gap and projection bias often leads to problems of self-control and helps explain impulsive decisions and even self-destructive behavior (cf. e.g., Loewenstein et al. 2003). It is these self-control problems that are worrisome in regard to consumer's decisions about entering a transaction. If the mentioned phenomena undermine utility-maximizing consumer behavior in consumer-seller relationship, this may justify legal intervention (see below).

Remedy: Withdrawal Rights

In response to the specific characteristics of distance selling and doorstep transactions, most jurisdictions have implemented statutory rules that mandate withdrawal rights for the consumer. It allows buyers to inspect the purchased goods and – without any reasons – return the goods to the seller within a certain period of time (cooling-off period), whereas the seller has to reimburse all payments received from the buyer. In the EU, based on the Directive on Consumer Rights (Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011, OJ 2011 L 304/64), withdrawal rights feature prominently in contract law, especially in distance selling and doorstep transactions (cf. Loos 2009, Eidenmüller 2011).

Unless contractually agreed upon, in the USA, sellers usually do not have an obligation to take back goods that simply do not satisfy consumers. Consumers do not have a generic right to withdraw. Even in the USA, however, withdrawal rights are sometimes provided for certain goods, for instance, by an FTC regulation for door-to-door transactions and by some state statutes for special kinds of distance selling, such as telemarketing (cf. Ben-Shahar and Posner 2011).

As a matter of fact, before mandatory rules were introduced that granted withdrawal rights (cf. Borges and Irlenbusch 2007) or where such rules do still not exist (cf. Ben-Shahar and Posner 2011), a surprisingly large majority of sellers already voluntarily offered, or offer, a return option. This observation leads to different views among law and economics scholars on withdrawal rights. Some do understand the right to withdraw as part of the optimal contract between consumer and seller (Ben-Shahar and Posner 2011). The majority view, however, is that these and similar rights are remedies for potential market failure (e.g., Rekaiti and Van den Bergh 2000; Borges and Irlenbusch 2007; Hoepfner, 2012; Stremitzer 2012) and is, therefore, somewhat close to the traditional legal perspective. This antagonism deserves attention in future law and economics scholarship.

Another interesting piece of information is the extent to which the right to withdraw is exercised. According to questionnaire results reported by Borges and Irlenbusch (2007), return frequency among German mail-order sellers increased from 24.2% in 1998 to 35% in 2004. However, other estimates of the rate of return are much lower (cf. OFT 2004). The different inquiries do not amount to consistent evidence. One has likely to distinguish between the specific kind of transaction and also between different kinds of goods. Drawing conclusions based only on these numbers appears to be inappropriate.

A Cure for Market Failure?

Withdrawal rights in distance selling and doorstep contracts are supposed to restore the balance of interest in consumer-business transactions by protecting consumers. In consideration of the market power of sellers, withdrawal rights render situational monopolies contestable. Potential withdrawal and cooling-off period facilitate access to the consumer by the competition and, thus, market entry. If rational sellers anticipate this, they may be disciplined by the legal rules (Hoepfner 2012). In this sense, legal rules can control market power (Stremitzer 2012).

Withdrawal rights may also counter some problems of nonrational consumer choice where

utility-maximizing choice is endangered. A cooling-off period facilitates risk perception to adjust and may bring attention to the contrast between long-term preferences and short-term choice (Rekaiti and Van den Bergh 2000; Hoepfner 2012). However, before jumping to hasty conclusions about legal intervention on these grounds, one has to carefully demonstrate that such and similar phenomena are indeed at work. Most of the phenomena discussed are context-dependent, and often, the different causes of anomalies are not well understood. In this regard, Korobkin (2003) has written an excellent article on the intricacies and complex problems that the endowment effect poses for legal analysis. Similar complexities also need to be considered in regard to consumer-seller relations. Affective forecasting errors are both persistent and prevalent (see above). This poses a problem only, however, if the mismatch between *ex ante* projection about utility and actual long-term satisfaction indeed causes an over-investment. More recent research revealed that adaptation processes are very different between simple, material consumption goods and experiential purchases. Specifically, after an initial rush of utility – that biases affective forecasting – people adapt to the consumption of material goods rather fast. By contrast, the people adapt to experiential purchases rather slowly, and sometimes their satisfaction even grows (Van Boven and Gilovich 2003; Frank 2008, Chap. 8; Nicolao et al. 2009). Therefore, the one-size-fits-all solution of withdrawal rights may be drawn into question at least insofar as disadvantageous consumer decision-making in regard to projection bias is concerned. The relation between the different aspects of nonrational consumer decision-making and possible legal interventions, however, is left to detailed future research. Moreover, Korobkin (2003) warns that once legal scholars import into their work concepts and findings from other disciplines, the sophistication of translating these findings into policy-relevant analyses differs a lot.

Therefore, one should require a higher burden of proof to justify legal intervention on these and similar grounds, which are foreign to the legal scholarship. In fact, this concern emphasizes the

importance of more carefully testing implications and policy solutions – possibly in the laboratory (cf. Engel 2013) – before implementing them in practice. Such testing also has the potential to reveal unanticipated forces at work (see below).

Withdrawal rights also seem to be an appropriate remedy for the multidimensional, *ex ante* information asymmetry. In fact, this has been suggested to be “the most fundamental aspect of building consumer confidence” (Dickie 1998, p. 223). On the one hand, withdrawal rights and cooling-off periods enable a discovery process for testing mainly experience dimensions of a product, but also other product characteristics that could not be validated *ex ante*. Therefore, withdrawal rights can be understood as information technology (cf. Hoepfner 2012).

On the other hand, however, it is also noteworthy that mandatory withdrawal rights delete important information signals. Signaling credible commitment is not possible under a mandatory regime for those sellers that are reliable and offer high-quality products and want to distinguish themselves from their inferior competitors. The information signal of voluntarily offering a withdrawal right gets lost (cf. Hoepfner 2012). As is the case with warranties (cf. Emons 1989), the incentive function remains insofar as better quality products, and higher seller reliability leads to lower withdrawal costs (Rekaiti and Van den Bergh 2000).

Moreover, although a withdrawal right may ameliorate the *ex ante* information problem, this mechanism also introduces *ex post* information asymmetries (hidden action). The danger of *ex post* opportunism, specifically consumer moral hazard, looms large (Stremitzler 2012). The consumer may, e.g., use the product to an excessive degree and just ship it back after making use of his withdrawal right. To align post-contractual incentives, distance and door-to-door selling regimes often implement liability rules for the consumer in case of excessive or even normal usage of the good that causes deterioration of the good.

Finally, the potential to unilaterally exercise a withdrawal right burdens sellers with additional risk. This risk leads to relatively increased costs for the seller because only a share of transactions

is completed. This translates to higher prices (Rekaiti and Van den Bergh 2000; Hoepfner 2012).

To conclude, it appears that there is an efficiency rationale to justify withdrawal rights. Ben-Shahar and Posner (2011) argue – on the basis of their model – that there is reason to recognize a generic right to withdraw but that the rule should be a default rule, not a mandatory rule. However, their model hinges on information asymmetries and the perverse ex post consumer incentives. The other aspects discussed here do not enter. Therefore, the potential of withdrawal rights to address causes of market failure is not so clear-cut after all. At best, they are second-best solutions (critically toward withdrawal rights also: Eidenmüller 2011; Hoepfner 2012). Moreover, there is more room for future research than one may think. The following passages exemplify that relevant research on withdrawal rights is far from exhausted.

Perverse Incentives?

If one wants to draw a conclusion so far, it will likely be that withdrawal rights can lead to efficient incentives ex ante and ex post as long as the right is implemented correctly. In fact, a consumer right to withdraw coupled with the obligation to pay depreciation costs is analytically comparable to breach of contract when coupled with reliance damages (Ben-Shahar and Posner 2011).

One unintended consequence of such a simple way to terminate the contract has been brought to attention by Hoepfner (2012). He emphasizes that consumers face two decisions: (1) to contract for the good and (2), if contracted, to withdraw from the contract. In light of the second decision, which can be compared to an opt-out opportunity (not withdrawing is the default), sellers have an incentive to increase consumer compliance with the contract. In other words, sellers have an incentive to manipulate downward the probability that consumers withdraw from the contract. This would put a question mark behind the presumed effectiveness of withdrawal rights.

In the context of doorstep sales, Hoepfner (2012) further elaborates that sellers can use specific bargaining techniques that exploit behavioral

quirks that relate to the two important driving forces in individual decision-making: reciprocity and consistency. These mechanisms, however, can be easily translated to distance selling. Once the withdrawal stage is reached, consumers may face a status quo dilemma. In light of economic (e.g., transaction cost, uncertainty, specific investments, switching cost) and psychological variables (e.g., risk aversion, loss aversion, regret aversion), consumers may disproportionately often decide for the status quo. They may not withdraw although withdrawal would be the optimal choice.

Hoepfner (2012) suggests that, if these mechanisms work out as in his analysis, there will be an inefficiently high number of contracts entered into and an inefficiently low number of withdrawals. Consequently, as a first step he suggests changing the default from the withdrawal option from presumed consent to presumed denial – from opt out to opt in. However, these conclusions should be properly tested before jumping to policy conclusions by making use of empirical methods.

Fairness Considerations?

As mentioned above, withdrawal rights are thought of as a mechanism to restore the balance of interest, or to promote fairness, where individual skills, information, and/or bargaining power are very unequally distributed between contracting parties. The idea of fairness is one of the most important normative ideals in legal scholarship (cf. Kaplow and Shavell 1999; Singer 2008). In many contexts, the law requires contracting parties to take into account the legitimate interests of the other parties. Since research in behavioral economics has gained momentum, fairness considerations in exchange relationships also feature prominently in economics and law and economics (e.g., Kahneman et al. 1986; Konow 2003).

However, whereas withdrawal rights are supposed to promote fairness in an imbalanced relationship, they also provide the opportunity for the entitled parties to exploit their legal position. Borges and Irlenbusch (2007) experimentally tested the effect of withdrawal rights in the context of distance selling transactions. The researchers are

investigating two aspects. First, they test whether the exclusive liability of the seller for the return costs increases the withdrawal rate as compared to a situation where return costs are shared. The intuition is that, given no return cost, opportunistic buyers have an incentive to either order the good and reap the short-term value of use or order multiple goods with uncertain characteristics and afterward – through the withdrawal right – return the unfit goods to the seller for a reimbursement of the purchase price. Second, the researchers test whether the shift from a voluntarily granted withdrawal right to a statutorily mandated right increases the withdrawal rate. This prediction is based on fairness considerations. Briefly put, there is substantial evidence that people reciprocate perceived fairness cues (e.g., Fehr and Gächter 2000; Falk and Fischbacher 2006). If a seller voluntarily offers a withdrawal right, sellers may perceive this as fairness-induced kindness and reciprocate by exercising their withdrawal right less opportunistically. However, legally imposing a right to withdraw may provide an entitlement to the buyer to exercise this right and, moreover, signal that sellers are legally presumed not trustworthy. Mandating a right to withdraw may render buyers less fairness oriented. The researchers manipulate these variables across different distance selling scenarios.

The results indicate, first, that jointly bearing the return cost does not significantly change buyers' decision-making; in general, buyers do not behave friendlier, i.e., less opportunistically, if this does not also maximize their own payoff. Put differently, the mere existence of return costs does not distract participants in the laboratory to misuse their withdrawal right.

Moreover, although individual payoff maximization appears to be a central driver of buyers' behavior observed in the lab, Borges and Irlenbusch (2007) find clear indication that fairness considerations also play a systematic role in general. More interestingly, the researchers find an observable difference in how buyers interact with sellers depending on whether the withdrawal option was voluntarily provided or mandated. The number of choices that are unfavorable for the seller is significantly higher when the withdrawal

right is legally provided. In fact, from the experimental data, it is estimated that an unfriendly choice is 7.4% more likely under a mandatory regime (Borges and Irlenbusch 2007, p. 97). Ironically, implementing a statutory withdrawal right to protect buyers from unfair behavior crowds out their fairness considerations. This is clearly of concern for consumer contract law.

Summary

This entry quickly surveyed law and economics research important to analyze the consumer-seller relationship. An understanding of this relationship is fundamental for distance and door-to-door contracts, specifically, and consumer contract/protection law, in general. On this basis, the entry offers a discussion on research in law and economics about the major remedy available to consumers in these relationships, namely, withdrawal rights.

A provisional result of law and economics research is that withdrawal rights are able to address the imbalance in situational market power of sellers, ex ante information asymmetries that threaten the lemon market process, and non-rational consumer choice. However, withdrawal rights are second-best solutions only since they merely shift risk from consumers to sellers and facilitate consumer ex post opportunism. At least, withdrawal rights ought to be coupled with an obligation of the consumer to pay depreciation costs. This is analytically comparable to breach of contract coupled with a liability for reliance cost that is well known from seminal law and economics research. This solution is more suitable to establish efficient incentives.

However, despite this efficiency rationale research on distance and doorstep contracts is far from conclusive. Some information signals inevitably get lost in a mandatory regime. Although defaults tend to be sticky, a default rule on withdrawal rights may lead to efficiency gains compared to a one-size-fits-all solution because it facilitates sorting and signaling of heterogeneous buyers and sellers. Moreover, more research is needed concerning the perverse incentives to manipulate downward the withdrawal rate and

concerning the dynamics between the voluntary and the mandatory provision of withdrawal rights as well as the interplay with social norms and preferences.

Cross-References

- [Choice Under Risk and Uncertainty](#)
- [Signalling](#)
- [Transaction Costs](#)

References

- Akerlof G (1970) The market for “lemons”: quality uncertainty and the market mechanism. *Q J Econ* 84:488–500
- Ben-Shahar O, Posner R (2011) The right to withdraw in contract law. *J Legal Stud* 40:115–148
- Borges G, Irlenbusch B (2007) Fairness crowded out by law: an experimental study on withdrawal rights. *J Inst Theor Econ* 163:84–101
- Bourgoignie T (1992) Characteristics of consumer law. *J Consum Policy* 14(3):293–315
- Coricelli G, Critchley HD, Joffily M, O’Doherty JP, Sirigu A, Dolan RJ (2005) Regret and its avoidance: a neuroimaging study of choice behavior. *Nat Neurosci* 8:1255–1262
- Eidenmüller H (2011) Why withdrawal rights. *Eur Rev Contract Law* 7:1–24
- Emons W (1989) The theory of warranty contracts. *J Econ Surv* 3:43–57
- Engel C (2013) Legal experiments: mission impossible? Eleven International Publishing, The Hague
- Dickie J (1998) Consumer confidence and the EC directive on distance contracts. *J Consum Policy* 21:217–229
- Falk A, Fischbacher U (2006) A theory of reciprocity. *Game Econ Behav* 54:293–315
- Fehr E, Gächter S (2000) Fairness and retaliation: the economics of reciprocity. *J Econ Perspect* 14: 159–181
- Frank RH (2008) *Microeconomics and behavior*, 7th edn. McGraw-Hill, New York
- Frederick S, Loewenstein GF, O’Donoghue T (2002) Time discounting and time preference: a critical review. *J Econ Literat* 40:351–401
- Gilbert DT, Wilson TD (2007) Prospection: experiencing the future. *Science* 317:1351–1354
- Goetz CJ, Scott RE (1980) Enforcing promises: an examination of the basis of contract. *Yale Law J* 89:1261–1322
- Hoepfner S (2012) The unintended consequence of doorstep consumer protection: surprise, reciprocation, and consistency. *Eur J Law Econ*. <https://doi.org/10.1007/s10657-012-9336-1>
- Kahneman D, Knetsch JL, Thaler R (1986) Fairness as a constraint on profit seeking. *Am Econ Rev* 76:728–741
- Kaplow L, Shavell S (1999) The conflict between notions of fairness and the pareto principle. *Am Law Econ Rev* 1:63–77
- Konow J (2003) Which is the fairest one of all? A positive analysis of justice theories. *J Econ Literat* 41:1188–1239
- Korobkin R (2003) The endowment effect and legal analysis. *Northwest Univ Law Rev* 96:1227–1293
- Lele MM (2007) *Monopoly rules: how to find, capture, and control the world’s most lucrative markets in any business*. Kogan Page, London
- Loewenstein GF, O’Donoghue T, Rabin M (2003) Projection bias in predicting future utility. *Q J Econ* 118:1209–1248
- Loewenstein GF, Schkade D (1999) Wouldn’t it be nice: predicting future feelings. In: Kahneman D, Diener E, Schwartz N (eds) *Well-being: the foundations of hedonic psychology*. Russel Sage, New York, pp 85–105
- Loos M (2009) Rights of withdrawal. In: Howells G, Schulze R (eds) *Modernising and harmonising consumer contract law*. Sellier, Munich, pp 237–278
- Nicolao L, Irvin JR, Goodman JK (2009) Happiness for sale: do experiential purchases make consumers happier than material purchases? *J Consum Res* 36:188–198
- OFT Market Study on Doorstep Selling (2004) Available at: <http://www.of.gov.uk/OFTwork/markets-work/completed/doorstep-selling>. Accessed 16 July 2014
- Rekai P, Van den Bergh R (2000) Cooling-off periods in the consumer laws of the EC member states: a comparative law and economics approach. *J Consum Policy* 23:371–407
- Singer J (2008) Normative methods for lawyers. Harvard Law School Public Law Research Paper No. 08–05
- Stremitz A (2012) Opportunistic termination. *J Law Econ Org* 28:381–406
- Trebilcock MJ (1993) *The limits of freedom of contract*. Harvard University Press, Cambridge
- Van Boven L, Gilovich T (2003) To do or to have? That is the question. *J Pers Soc Psychol* 85:1193–1202
- Weatherill S (2005) *EU Consumer Law and Policy*. Cheltenham: Edward Elgar

Distance Selling Contract

- [Distance Selling and Doorstep Contracts](#)

Doorstep Contract

- [Distance Selling and Doorstep Contracts](#)

Double Tax Agreements

► Double Tax Conventions

Double Tax Conventions

Pasquale Pistone^{1,2,4} and Martin Zagler^{3,4}

¹University of Salerno, Salerno, Italy

²IBFD, Amsterdam, The Netherlands

³UPO University of Eastern Piedmont, Novara, Italy

⁴WU Vienna University of Economics, Vienna, Austria

Synonyms

Double Tax Agreements; Double Tax Treaties

Definition

Double tax conventions are international treaties between sovereign states to assign taxing rights between them in order to avoid double taxation. They tend to follow model conventions, the most prominent being the OECD Model Convention. Double tax conventions based on the OECD Model Convention consist of seven chapters. Chapter 1 regulates the scope. Chapter 2 contains the rules of interpretation and the definitions. Chapters 3 and 4 include the rules on the allocation of taxing powers. Chapter 5 provides the two methods for relieving international double taxation. Chapter 6 addresses special provisions and chapter 7 the final provisions. After introducing the concept of international double taxation, this entry defines double tax conventions and argues why countries might wish to sign such a treaty, and why not. The last chapter before the conclusion discusses special issues with double tax conventions, in particular its effect on profit shifting and foreign direct investment.

Introduction

The exclusive right to levy taxes is a defining principle of the public sector. As long as we remain firmly within the limits of the nation state (Bodin 1579), this principle remains largely uncontested (McLure 2001). Once we leave these narrow confines, the issue at hand gets more difficult. Every single nation state may be inclined to impose taxes elsewhere, however small the link is. Without a supranational authority, nation states will dispute over the tax base. This is the dilemma of international taxation, which double tax conventions (DTC henceforth) aim to overcome. There are over 3.000 DTCs worldwide (of about 17.000 potential treaties if every country in the world would have a treaty with every other country) (IBFD 2017). A major project for their coordination is undergoing through the so-called multilateral instrument, which is in essence an international legal instrument that steers bilateral treaties towards cross-border consistency at the worldwide level without requiring the renegotiation of such treaties.

This contribution will answer the following questions:

- What are double tax conventions?
- Why do countries form DTCs? (And related, why do countries decide not to form a DTC? And why might countries terminate DTCs?)
- Finally, what are the consequences of DTCs?

We will answer these questions as follows. In the following entry, we will present the problem of double taxation, which is the historically dominant argument for DTCs. Next, we will define DTCs and explain their main elements and historical evolution. Thereafter, we will discuss several issues within DTCs, before discussing international tax coordination as a potential solution.

International Double Taxation

International double taxation arises when two governments claim taxing rights on the same revenue or wealth in respect of the same taxable year.

Depending on whether it affects the same taxpayer in a legal or economic sense, it can be characterized as juridical or economic double taxation. Tax law literature differentiates between juridical and economic double taxation, according to whether the duplication of the tax burden affects one and the same individual or entity (juridical double taxation) or two different ones (economic double taxation).

In principle, the absolute nature of tax sovereignty could lead any government to claim unilaterally complete worldwide taxing rights on everything. However, in practice this does not happen, since national tax sovereignty finds limits in the exercise of taxing powers at the international level. Therefore, the concept of tax jurisdiction is narrowed down to the situations in which a country can establish a reasonable genuine link, based on the so-called connecting factors, with its right to exercise tax sovereignty.

International double taxation is the result of the exercise in parallel of more than one tax jurisdiction and can arise as a consequence of similar or different connecting factors. The connecting factors to tax jurisdiction are essentially of two types, namely personal and objective. In the case of income taxation, objective connecting factors tax income as such, i.e., for the mere fact of being sourced on the territory on which the state exercises its jurisdiction. The state that exercises its taxing jurisdiction on a territorial basis through objective connecting factors is often referred to as the state of source. Because of the objective connection with the taxing jurisdiction, such state does not take into account any personal circumstance of the taxpayer who derives this income.

Personal connecting factors have developed more recently in the twentieth century for allowing one state to exercise its taxing jurisdiction on the overall ability to pay of a taxpayer, even when s/he derived income outside the territory of the state. For this reason, it was originally called extraterritorial taxation. In fact, it is not, since personal connecting factors allow countries to exercise their jurisdiction on the taxpayers established on their territory. The need for a factual link with the territory of a country has steered the development of personal connecting factors in

line with criteria that expressed a sufficient presence of the taxpayer on such territory. For this reason, most countries have adopted residence or domicile to establish the personal link with their tax jurisdiction. Therefore, a country exercising its tax jurisdiction based on a personal connecting factor is often described as the country of residence (Connecting factors may vary whether we are discussing natural or legal persons.). However, some countries, most notably the United States, adopt citizenship as its main personal connecting factor, thus allowing a more direct link with the public services that it provides to persons holding this status.

Countries have stretched both types of connecting factors over the years, with a view to extending their tax jurisdiction and preventing possible attempts to circumvent it. Accordingly, legal fictions have deemed income to be within the national territory for source-based taxation purposes and the number of personal connecting factors has increased in order to exercise residence-based taxation in the presence of any minimum personal link with the tax jurisdiction.

The need to counter tax avoidance and evasion has brought personal connecting factors to operate in a country even when a taxpayer with separate legal personality is established on the territory of another country and is controlled by a resident person of the former country, such as in the case of controlled-foreign-company (or CFC) legislation (OECD 2015). From the perspective of the country of source, the stretching of objective connecting factors was caused by the need for countering base erosion of value created on the territory and shifted to that of another tax jurisdiction. Accordingly, measures like the diverted profits tax in Australia and the UK (Nguyen 2017), and the Indian equalization tax on services (Lahiri et al. 2017) have been introduced in order to safeguard the exercise of taxing powers on income sourced within the national territory.

This situation broadens the potential for the exercise in parallel of tax jurisdiction and magnifies the exposure of cross-border situations to international double taxation, which can arise in three main groups of conflicts of tax jurisdictions.

The most frequent and typical case of international double taxation is the one caused by residence-source conflicts. This occurs for instance when income is generated in one country (source) and the benefit accrues to a resident in a different country. However, international double taxation can also be the outcome of residence-residence conflicts (or similar phenomena involving a conflict with another personal connecting factor, such as citizenship) and of source-source conflicts.

With few exceptions, taxation exhibits negative welfare implications due to its distortionary effect on prices. Double taxation even more so. Consider the simple case of a 2 country, 2 goods world, where for the sake of simplicity supply is infinitely elastic (standardized at $p = 1$) and linear demand is given by $q = k - p$. Both markets create total welfare of k^2 . The introduction of a tax t on one market reduces welfare by half the square of the tax rate, $t^2/2$. Now suppose country 1 taxes market 1 at a tax rate t , and country 2 taxes market 2 coincidentally at the same rate. The global deadweight loss of taxation would be t^2 . If instead both countries insist on taxing the same market, the deadweight loss on this market would amount to $(2t)^2/2 = 2t^2$. Even if neither country touches the other market (double nontaxation), the deadweight loss of double taxation exceeds by a large margin single taxation. Government revenues, too, suffer from double taxation. In the case of single taxation on both markets, global governments revenues amount to $2t(1 - t)$, whereas in the case of double taxation on a single market, global tax revenues would equal $2t(1 - 2t)$, which is less than the above due to the distortionary effect of taxes (Nowotny and Zagler 2009, 286 ff).

In principle, there is no natural order of priority when two states exercise in parallel their taxing jurisdiction. Historically, it has even occurred that more than 100% of the tax base has been taxed due to double taxation (Herndon 1932). For this reason, the introduction of specific legal remedies is needed in order to counter international double taxation.

One simple remedy to international double taxation is obviously unilateral measures. Each

of the two countries in the above example can unilaterally forgo the right to tax a particular market, thus avoiding double taxation. While unilateral measures may help to avoid double taxation, they may generate unwanted side effects in cross-border activities. Forgoing taxing rights will put domestic exporters on par with their foreign competition (export neutrality), but it will at the same time treat differently foreign importers, who will not need to pay taxes in the source state as opposed to their domestic competition (import non neutrality) (Lockwood 2001).

Another remedy is to address international double taxation of income by means of bi- and multilateral measures framed into international tax treaties, specifically designed to counter this phenomenon. Those treaties, which we shall from now on indicate as double tax conventions, essentially coordinate the exercise of taxing powers with a view to preventing, reducing, and relieving international double taxation.

Double tax conventions are most frequently bilateral and drafted along a common international pattern, based on different model conventions, most notably the OECD Model Convention. Besides the OECD Model, other important model conventions are the UN model and several national models, such as the US model. They regulate the exercise of tax jurisdiction on cross-border situations by establishing additional conditions to either Contracting State. In some cases, they attribute exclusive taxing powers, thus preventing international double taxation. In other cases, they keep the right of each Contracting State to exercise its taxing jurisdiction, sometimes establishing limits for either of them (such as in the case of passive income, which the source state can tax only up to the maximum amount agreed in the treaty), and provide for common rules that relieve international double taxation.

Treaty measures for relieving international double taxation are essentially similar to the ones that may apply based on domestic law. Credit and exemption are the two main methods for relieving international double taxation in the state of residence of the taxpayer. However, their effects are different from a legal (CJEU, FII Group Litigation I 2006, paras. 45–48, FII Group

Litigation II 2012) and economic perspective. For this reason, in principle one State may not apply different methods to relieve (economic) double taxation in purely domestic and cross-border situations.

When credit applies, the state of residence exercises its taxing jurisdiction on a worldwide basis, thus also on foreign sourced income or wealth, and allows for a deduction of taxes paid in the other state. This mechanism turns lower foreign taxes paid in the state of source into a lower deduction against taxes due in the residence state, thus pursuing capital export neutrality. However, this mechanism operates in a way that may not affect the right of the latter country to exercise its taxing jurisdiction on domestic sourced income or wealth. This limitation, also known as ordinary credit, gives no relief for foreign taxes when higher than those applicable in the residence state.

When exemption applies, the state of residence does not exercise its taxing jurisdiction on foreign-sourced income and wealth, thus turning foreign taxes into final and allowing taxpayer to bear the same tax burden of residents of the source state. This method pursues capital import neutrality and approximates the exercise of taxing jurisdiction to what it would normally be in a pure territorial system. However, it makes the exercise of taxing powers in cross-border situations vulnerable to phenomena of double nontaxation, which occurs when the source state does not tax income or wealth that it would be entitled to under the double tax convention. For this reason, double tax conventions generally allow for a switchover to credit in such circumstances.

A variation of the exemption method – known as exemption with progression – preserves tax progression, in order to allow domestic-sourced income of individuals to be taxed at the rate that would correspond if all income had been sourced in the state of residence.

A variation of the credit method – known as tax sparing – where the residence State is bound to give credit for the taxes that should have been paid in the source State, whether the source State chooses to exercise taxing rights or not; therefore, allows the source State to pass on the benefits of

any incentives to the taxpayer. This preserves the right of the source state to pursue its international tax policy goals without interferences by the residence state. This method may also be used for allowing the source state to keep the effects of its tax incentives at the international level (Brooks 2009).

Although remedies against international double taxation can also operate under domestic law, their functioning at treaty level provides qualitatively better results, since the latter may coordinate the exercise of taxing powers between the Contracting States and reduce the extent to which international double taxation may arise. This is one of the reasons for double tax conventions to become so common at the international level.

Double Tax Conventions (Why Countries Conclude or Not a DTC)

The Worldwide Network of Bilateral Double Tax Conventions

The first modern double tax convention goes back to 1899 when Prussia and Austria-Hungary signed such a treaty (Easson 2000). Since then, the number of treaties has been rising steadily; at the beginning, mostly industrialized countries entered into such treaties with each other. During the last two decades, developing economies have increasingly been integrated into the global treaty network. After 1990, the number of DTT signatures has been surging, so that around 60% of today's DTTs have been signed in the last 20 years (Baker 2014).

The historical development of double tax conventions is the outcome of the activism of international organizations. Their technical work in this field started with the economic studies on international taxation carried out under the auspices of the League of Nations (LoN) from the 1920s onwards and continued with the proposals and drafts of the Organization for European Economic Cooperation (OEEC) in the 1950s. Since the preliminary studies, it became clear that the establishment of a global multilateral framework was a too ambitious goal to achieve, due to the numerous differences across the various

countries' tax systems. For this reason, double tax conventions have essentially developed along the paths of bilateral negotiations.

However, the current international framework for double tax conventions presents a significant degree of coordination across bilateral treaties, which is largely the outcome of the constant efforts by the Organization for Economic Cooperation and Development. The Model Tax Conventions released by the OECD since 1963 have come to constitute the main reference for the clauses contained in most of the existing bilateral tax treaties around the world. The OECD Model pursues tax policy goals of capital exporting countries in the allocation of taxing powers. It owes its success to the very high standards of consistency in promoting this standard and to its Commentaries, which extensively elaborate on the interpretation of clauses of double tax conventions. For this reason, it quickly turned into the main source for negotiating treaties between OECD countries.

Since 1980, there is also a UN Model Convention. The main goal pursued by this Model Convention is to provide for allocation rules that better pursue the international tax policy goals of developing countries. However, the diffusion of this UN Model Convention in related tax treaties has been rather limited in relations between developed and developing countries and, more recently, also in treaties between developing countries.

The Policy Goals of the OECD Model Convention and Its Diffusion in Relations with Developing Countries

The diffusion of the OECD Model (OECD 2017) in double tax conventions with – and, sometimes even between – developing countries is a potential source of concern, since this model reflects the policy goals of capital exporting countries and developing countries have not contributed to develop its clauses. Yet, such countries agree to sign double tax conventions shaped along the OECD Model. The existence of asymmetries in flows of income and capital with capital exporting countries turns the conclusion of double tax conventions for such countries to cause more losses in their exercise of taxing powers than advantages.

Yet, such countries sign double tax conventions, sometimes for reasons connected with the perception of investors, some other times as a package deal with the signature of other international treaties or for other reasons.

Industrialized (typically capital exporters) and developing countries (typically capital importers) may indeed have different motives when signing a DTC. Fostering outbound investment and thus encouraging the international expansion of domestic companies may arguably be more relevant for capital-exporting countries. For capital importers, encouraging inbound investment may be more in the focus, with policy makers wishing to attract foreign direct investment entailing the transfer of skills and technologies and thus fostering economic growth (Lang and Owens 2014). Finally, the function of DTCs as a signaling device indicating that the signatory states play by the internationally accepted tax standards may be more relevant for developing countries (Dagan 2000).

The Structure of Double Tax Conventions Following the OECD Model Convention

Since most bilateral treaties in fact follow the OECD Model Convention, it makes sense to outline their structure by referring to such Model and its clauses. However, we shall include some reference to the Model Convention drafted by the United Nations when the clauses contained in such Model contains some significant deviations from the pattern provided by the OECD Model. For the purpose of simplicity, we shall focus our analysis of international double taxation by referring to the OECD Model Tax Convention on income and capital. Yet, the OECD has also produced in 1982 a Model Tax Convention on Inheritance and Gift Taxes.

Double tax conventions based on the OECD Model Convention consist of seven chapters. Chapter 1 regulates the scope. Chapter 2 contains the rules of interpretation and the definitions. Chapters 3 and 4 include the rules on the allocation of taxing powers. Chapter 5 provides with the two methods for relieving international double taxation. Chapter 6 addresses special provisions and chapter 7 the final provisions.

The application of the core part of double tax conventions follows in substance four main steps, to which we shall now briefly refer, followed by some brief reference to the special and final provisions. The first step is the entitlement to treaty benefits, also known as the subjective scope of tax treaties. It covers cross-border situations involving persons who are resident of either or both Contracting States. This clause is contained in Article 1, but the concept of residence is to be interpreted in the light of Article 4, which primarily relies on domestic law of the Contracting States, but also includes tie-breaker rules to address cases of dual residence and the residence-residence conflict connected thereto.

The second step is the objective scope of tax treaties, which includes the existing taxes specifically indicated by the Contracting States and the ones that may replace them with similar features. This clause is contained in Article 2.

The third step is the most important one. It contains the 16 clauses that limit the exercise of taxing powers on cross-border income and the one clause on capital covered by the double tax convention. Such clauses are better known as allocation rules and are included in Articles 6–22 (excluding Article 9).

The ones on income essentially regulate six basic types of income (from immovable property, business income, passive income, capital gains, income from employment, and the residual category of other income).

Allocation rules include specific provisions whenever the specific features of income require so in order to achieve a balanced allocation of taxing powers. This can be the case of Article 16 on directors' fees, or of Article 18 on pensions from past private employment, in respect of which the basic rule of Article 15 cannot operate due to the uncertain determination of an actual place of exercise of the activity or its absence. In case no specific allocation rules are deemed applicable, treaties contain general rules as a fallback option.

The existence of different limits to the exercise of taxing powers across the allocation rules can be the unintended source of cross-border tax biases and gives rise to a potential for tax avoidance. For instance, the clauses on passive income in the

OECD Model Convention may allow for higher withholding taxes on dividend than on interest. Furthermore, the OECD Model Convention prevents double taxation on royalties and capital gains by allocating exclusive taxing powers to the state of residence. This context may induce taxpayers to alter their behavior and prefer capitalization over distribution of dividends. However, it may also create a potential for circumvention of the clauses in order to obtain unintended savings of taxes in the state of source through rule shopping (which is a form of tax avoidance), such as for instance, when taxpayers seek characterization as interest over that as dividends.

From the perspective of the allocation of taxing powers, we can group the main bulk of double tax convention clauses into three categories. The first category includes clauses that allocate taxing powers to one Contracting State and thus prevent international double taxation, thus making the fourth step unnecessary. Such clauses are the ones contained in Articles 7 (in the absence of a permanent establishment), 8, 12, 13 (3), 13 (5), 15 (1) (first sentence), 15 (2) (if the three negative conditions are not met), 18, 19 (1) (a), 19 (1) (b), 19 (2) (a), 19 (2) (b), 21 (1), 22 (3), 22 (4). Furthermore, this is also the case of Article 14 of the UN Model Convention in the absence of a fixed base. Interestingly, most clauses of this variety allocate taxing rights to the resident state, except where shipping, aircraft, etc. are covered. However, there is a proposal to change these provisions to favor residence taxation as well in the 2017 update of the OECD Model Convention.

The second category includes clauses that allocate taxing powers to both Contracting States but limit the exercise of powers by the Contracting State of Source up to a maximum amount. When the State of Residence relieves double taxation by the credit method, this approach reduces the risk of unrelieved international double taxation. Such clauses are the ones contained in Articles 10 (2), 11 (2) and, in the UN Model Convention, also Article 12 (2). The third category includes clauses that allow both Contracting States to exercise their taxing powers, thus leaving it up to the State of Residence to relieve double taxation. Such clauses

are contained in Articles 6, 7 (for income attributable to a permanent establishment), 10 (1), 10 (4), 11(1), 11 (4), 12 (when the beneficial ownership requirement is not met), 12 (4), 13 (1), 13(2), 13 (4), 15 (1) (second sentence), 15 (2) (when the three negative conditions are met), 16, 17, 21 (2), 22 (1) and 22 (2), as well as, Article 14 (in the presence of a fixed base), 18 (2) (alternative B) and 21 (3) of the UN Model Convention.

For clauses on allocation of taxing powers falling under the second and third category, the fourth step is necessary. In particular, it obliges the State of Residence to give relief for international double taxation by means of exemption (Article 23A) or credit (Article 23B).

The special provisions apply also beyond the scope of double tax conventions and can be grouped into two main clusters. First, the non-discrimination clauses, contained in Article 24, essentially regulate how the Contracting States may exercise their taxing powers in respect of non-nationals. Second, three provisions regulate the relations between tax authorities of the Contracting States. In particular, Article 25 allows them to interact in order to achieve a common view on the interpretation and application of the convention in the framework of the so-called mutual agreement procedures. Furthermore, Article 26 sets the conditions for mutual assistance concerning cross-border exchange of information in tax matters and Article 27 – not always included in bilateral double tax conventions – the ones related to assistance in the collection of taxes.

The final provisions reiterate the immunities established by international law conventions for members of diplomatic and consular missions (Article 28), define the territorial scope (Article 29), entry into force (Article 30), and termination of the double tax convention (Article 31). Furthermore, the 2017 Draft Update to the OECD Model Convention includes a new treaty clause that limits the benefits of the tax convention in order to counter cases of tax avoidance that may harm the taxing rights of the State of Source. Once approved, this clause will be inserted in Article 29 and will thus imply the renumbering of the last three Articles currently included in the OECD Model Convention.

Although double tax conventions have a bilateral nature and only binding their signatory states, their structure and content are now undergoing an unprecedented process of multilateralization in connection with the implementation of the multilateral instrument that 68 countries have signed on 7 June 2017 in the framework of activities connected with the base erosion and profit shifting (BEPS) project. Essentially, the multilateral instrument constitutes a convention that co-exists with bilateral agreements of the signatory States and steers them towards the agreed goals and minimum standards of the BEPS project, which include the countering of abuse of double tax convention and of harmful tax competition and the solution of cross-border tax disputes.

Why Do Countries Sign DTCs?

The preamble to a DTC is not fixed under the OECD Model Convention and most DTCs contain wording stating that the DTCs have been entered into to prevent double taxation in a cross-border transaction between the two concerned States. However, some DTCs also expressly provide for other reasons such as the “encouragement of mutual trade and investment”. See for example, the preamble and title to the India-Mauritius DTC (1983). The Multilateral Convention to Implement Tax Treaty Related Measures to Prevent BEPS also allows the option to add such language to the preamble.

Over the years, DTCs have come to pursue additional goals to the one of countering double taxation. DTCs may provide certainty in tax matters for international investors, prevent tax discrimination for investments in the other state, and avoid double taxation of income arising in cross-border transactions (Pickering 2013). In particular, DTCs serve to mitigate international tax avoidance and evasion and to protect the domestic tax base. This justifies the conclusion of double tax treaties also with countries that either have lower levels of taxation, or no taxation such as the United Arab Emirates or Hong Kong.

In various ways, DTCs can contribute to achieve these goals. They address cross-border transactions between associated enterprises

(Article 9 of the OECD Model Tax Convention on Income and on Capital (OECD Model)) and they provide for information-sharing between the contracting states (Article 26 of the OECD Model). Furthermore, specific provisions and concepts are inserted such as the limitation of benefits provisions or the beneficial ownership concept, which “restrict access to treaty benefits to residents of the contracting states” (Baker 2014). Accordingly, the Multilateral Convention to Implement Tax Treaty Related Measures to Prevent BEPS (2017) seeks to modify the preamble of covered DTCs to expressly provide that while DTCs intend to avoid double taxation, they should not be used to facilitate double non-taxation or treaty shopping, as a “minimum standard” measure under the BEPS project.

Voget and Lighthart (2011) empirically study the motives to conclude DTCs for a large country sample covering both industrialized and developing countries. They conclude that countries sign DTCs primarily to reduce international double taxation. DTCs mitigate double taxation by “harmonizing tax definitions, defining taxable bases, assigning taxation jurisdictions, and indicating the mechanisms to be used to remove double taxation when it arises” (Baker 2014). Yet, many authors argue that double taxation can be – and by most countries is – prevented unilaterally (Braun and Fuentes 2014; Rixen and Schwarz 2009).

In light of the above, some researchers argue that the main benefit role of DTCs lies in the harmonization and the lowering of withholding tax rates on international capital income (Davies 2003). OECD countries are encouraged to conclude a double tax convention for limiting the exercise of taxing powers by the source state. Reciprocity in flows of income and capital is expected to level out any potential loss of taxing powers between such states.

Whereas reduced source taxation may improve investment attractiveness, this depends on the method used for the avoidance of double taxation. Specifically, the exemption method will typically allow the investor to gain benefits from the lower withholding taxes. However, under the credit method, the lowering of withholding tax rates may also entail “distributional implications.”

Unless tax sparing is provided for, the reduced source taxation results in higher residence taxation, and therefore providing no benefits to the investor. In particular, with asymmetric investment positions, the lowering of withholding tax rates in treaties using the ordinary credit method “involves a revenue transfer from the net capital importer to the net capital exporter” (Rixen and Schwarz 2009). Therefore, the benefits of DTCs are sometimes being described as more on the side of capital exporters (Dougherty 1978).

Chisik and Davies (2004) discuss these distributional implications in the framework of tax-treaty bargaining. Based on a theoretical model, they predict that it may be more difficult for highly asymmetric countries to negotiate a tax treaty. They then show that highly asymmetric countries tend to conclude tax treaties with higher withholding tax rates.

Despite different preferences, having more economical strength and more bargaining power, capital exporters may have the power to pressure capital importers to enter into treaty negotiations (Pickering 2013). Nevertheless, Paolini et al. (2016) point out that DTCs may be perceived by capital-importing countries as means to partly regain their sovereignty with respect to the taxation of income, which non-residents generate on their territories. Due to the treaty, the allocation of taxing rights will be stated clearly and the taxes for business income paid in the capital-importing country will become final and thus relevant for firms. As a result, capital-importing countries will be in a position to use tax policy instruments in order to attract international investment flows.

The above discussion shows the complexity of the treaty formation process. Numerous empirical studies investigate the main drivers behind this process. Generally, we observe that countries have tax conventions in place with countries with which they have close economic ties. Egger et al. (2006) find that the bilateral country size and host government expenditure as a percentage of GDP significantly increase the likelihood that a country-pair signs a DTC. Lejour (2014) finds that sharing a common colonial past and the same language have a significantly positive impact on the probability of two countries concluding a

DTC, while distance is found to have a significantly negative effect. Finally, Barthel and Neumayer (2012) analyze DTC formation patterns focusing on spatial dependence. They find evidence that country-pairs are more likely to sign a DTC the more DTCs have previously been concluded in the region.

Issues with DTCs

DTCs and FDI

DTCs cover a large majority of foreign direct investment (Radaelli 1997). Several studies have examined whether and to what extent FDI responds to different tax treatment, finding that firms do indeed respond to a variety of tax policies and that this can result in an inefficient allocation of investment across countries. This can allocate investment away from its most productive use. One potential method of eliminating this inefficiency is a double tax convention. These treaties adjust the tax environment for investment between treaty partners by specifying the applicable tax base, the withholding taxes that can be applied, and other measures affecting the taxation of FDI.

Double taxation occurs if a multinational company (henceforth MNC) pays tax on the same corporate income earned from economic activity in a foreign country twice: once to the tax authorities of the foreign country, which is host to the economic activity, and once to the tax authorities of the home country, in which the company is domiciled. Double taxation could represent an obstacle or barrier to foreign investment, thus distorting the efficient allocation of scarce financial resources across countries of the world. Yet, DTCs can also reduce FDI in as much as they reduce tax avoidance, tax evasion and other more or less legal tax-saving strategies such as transfer pricing by multinational companies (Blonigen and Davies 2002).

DTCs can increase FDI as they standardize tax definitions and jurisdictions. Janeba (1996) theoretically shows that such coordination can reduce the double taxation of affiliate income. Tax

treaties affect the taxation of multinational enterprises by lowering withholding taxes and increasing tax certainty. In particular, Edmiston et al. (2003) find that uncertainty over tax policy is a significant barrier to FDI. Thus, if a tax treaty reduces the likelihood of a host nation unilaterally changing its tax policy, this added certainty would increase FDI. The combination of these two roles of treaties increases the expected value of after-tax returns from FDI. By contrast, the increased enforcement of transfer pricing regulation may actually exhibit a negative impact on FDI. DTCs can impact FDI also through transfer pricing enforcement regulations. Blonigen et al. (2014) argues that DTCs have a positive effect on FDI, which is larger for firms that use differentiated inputs. These (multinational) firms benefit from treaty provisions establishing guidelines for resolving disputes between taxation authorities. In contrast, firms that use more homogenous inputs are on average less likely to see any significant effect. This difference can be attributed to the additional regulations on the calculation of internal prices and encouraging the exchange of information between authorities. The establishment of anti-treaty shopping provisions inhibits the ability to direct profits through low-tax treaty partners in order to minimize tax payments. Since these increase the taxation of affiliate income in a given host, they would lead one to anticipate that a tax treaty might reduce FDI.

The empirical literature generally finds little evidence for the impact of DTCs on FDI (Louie and Rousslang 2008; Millimet and Kumas 2017). This result is often interpreted suggesting that the FDI increasing aspects of treaties, such as tax certainty or withholding tax reductions are balanced with negative effects as mentioned above, yielding a zero net effect of treaties on multinational enterprises.

Blonigen and Davies (2002) represent the first attempt to estimate the impact of DTCs on FDI. Respectively using panel data on OECD FDI (where FDI is measured as stocks) and US FDI (where FDI is measured as stocks or sales), they find that after controlling for country fixed effects there is either a small negative or insignificant

effect of treaty formation on FDI. The authors suggest that one possible reason for the non-promotion effect of treaties on FDI activity is that treaties reduce firms' abilities to evade taxes through transfer pricing or treaty shopping. An additional possibility for nonpromotion of FDI activity by new treaties is that treaties may increase investment uncertainty, at least in the short run.

Egger et al. (2006), who control for the endogenous selection of which treaties are actually formed, find that treaties significantly reduce FDI stocks. Davies et al. (2007) expand the research on this by utilizing affiliate-level data from Swedish-owned multinationals from 1965 to 1998. In line with earlier studies, they find no significant effect from treaty formation on the level of affiliate sales.

An important study from Neumayer (2006) finds, against all the results so far mentioned, robust empirical evidence that DTCs increase FDI to developing countries. However when the author splits developing countries into low-income and middle-income countries, he found that DTCs are effective in the group of middle income countries.

Just like trade diversion (Viner 1950), DTCs can create FDI diversion. Think of a simple case of a negatively sloped domestic demand schedule for capital, $K = f(r)$. Assume further that the domestic supply of capital is fixed at K^* . This implicitly determines the domestic autarky interest rate, $r^* = f^{-1}(K^*)$. Now suppose there are two foreign countries that can supply infinite amounts of capital, with $r_A + t > r^* > r_B + t > r_A$. Clearly, in the absence of any DTC that would eliminate the double taxation t , there would be capital imports ΔK from country B until $r^* = r_B + t$, with $\Delta K = K - K^* = f(r_B + t) - K^*$. The conclusion of a DTC with country A would change matters dramatically. Capital would no longer be imported from country B , but now from country A until $r^* = r_A$. There would be additional capital flowing into the economy (FDI creation) as $f(r_A) < f(r_B + t)$, but all the capital that previously arrived from country B will now be provided from country A (FDI diversion).

Selected Issues Concerning Double Tax Conventions

The application of double tax conventions raises numerous technical issues concerning the tax treatment applicable to cross-border situations. Such issues include the application of double tax conventions to dual resident companies and transparent entities, the concrete functioning of the arm's length method in transactions between associated enterprises, the inadequacy of the permanent establishment concept to the context of the digital economy and the protection of taxing rights of developing countries.

Our focus in this section is nevertheless on issues connected with base erosion and profit shifting, which have raised a serious concern on the effectiveness of the application of double tax conventions in connection with the exercise of national taxing sovereignty. From a structural perspective, insofar as double tax conventions are bilateral and negotiated in each case between two states, different conditions resulting from the negotiation unavoidably turn into an uneven set of rules. This creates the potential for cross-border tax disparities that in turn may give rise to unintended tax advantages, especially when taxpayers plan their affairs across the borders having that objective in mind.

A clear example of unintended tax advantages arises when the taxpayer pursues the application of a double tax convention in circumstances to which it was not meant to apply, thus abusing the tax convention, such as in the case of treaty shopping. Treaty shopping is a tax avoidance scheme targeting the reduction of source state taxation. For such purpose, cross-border investment is channeled through an intermediate company established in a treaty partner of the target state. Such a double tax convention limits the exercise of the taxing powers in the state of source more than what would otherwise apply under the domestic legislation of such country, or its double tax convention with the country of residence of the investor. This practice has significantly increased over the years and is countered through anti-avoidance rules contained in domestic legislation of the source state (general anti-avoidance

clauses usually drafted in the form of the principal purpose test, as well as various types of targeted and specific anti-avoidance clauses) and in double tax conventions (limitation-on-benefits clauses), including through the new Article 29 to be inserted in the 2017 Update to the OECD Model Convention. Such clauses essentially look at the function of the intermediate company and may limit the entitlement to the benefits of the double tax convention when such company lacks substance.

In principle, taxpayers have the right to arrange their affairs in a way that minimizes the tax burden, thus without any obligation to choose the most burdensome option from a tax perspective. However, international tax planning over the past decades has stretched this right to its extreme boundaries. Legal uncertainty concerning the tax advantages arising from the exploitation of cross-border tax disparities has contributed to prevent an effective solution to such problem.

Because of the size of this problem and its implications on the tax treatment of cross-border situations, the G20 has embarked on a global campaign against base erosion and profit shifting by multinational enterprises in the framework of the BEPS project. The implementation of this project is now steering international taxation and double tax conventions out of the traditional bilateral dynamics into a form of coordinated exercise of tax jurisdictions that secures consistency in tax treatment across borders.

The effects of the BEPS project on international taxation are manifold, including the obligation to counter international tax avoidance and aggressive tax planning through general, targeted and specific clauses also to be included in double tax conventions (Dourado et al. 2017; Krever 2016). This is steering double tax conventions towards a dimension in which they not only counter double taxation, but also its opposite phenomenon, i.e., double nontaxation. However, this conclusion should only apply to cases in which the Contracting States have not intended to produce this phenomenon. Therefore, even if the rules originating in the BEPS project support a fight against harmful tax competition, they should

not apply to cases when the Contracting States have intended to accept the existence of double nontaxation as one of the possible consequences of the allocation of taxing powers under the treaty. This is clearly the case of tax sparing, which leaves it up to the country of source to decide whether, when and at what conditions to exercise the taxing powers that the treaty has reserved to it. We consider this as particularly important in order to preserve consistency with the goals of the BEPS project. In particular, if the goal of the BEPS project is to allow the country of value creation to preserve the effectiveness of its taxing jurisdiction from erosion and profit shifting, this project should not prevent the source country from deciding not to exercise its taxing jurisdiction. This situation may often occur in developing countries in order to attract foreign direct investment and the measures of the BEPS project should not be used to allow the capital exporting country to tax income voluntarily forgone by the country of value creation. If that were the case, the BEPS project would in fact lead to opposite goals from the one that it officially pursues, harming the overall balance in the allocation of taxing powers under the double tax convention.

Another important point concerning the impact of the BEPS Project on double tax conventions with developing countries arises as to the settlement of cross-border disputes. The mutual agreement procedure has now turned into a minimum standard for double tax conventions, whose implementation is being secured through the multilateral instrument. Mutual agreement procedures are in essence a common forum for competent authorities of the Contracting States to reach a common view on technical issues concerning the interpretation and application of the double tax convention (and additional issues). However, developing countries lack capacity for running such procedures, especially when technical issues arise in relations with OECD countries, whose knowledge of technical issues is far more advanced. This type of problems in relations with developing countries would be even more difficult to address in the presence of arbitration clauses, which the BEPS multilateral

instrument only includes as a part of the optional content.

Why Do Countries Terminate a DTC?

The reasons for terminating a double tax convention may differ according to the country and context. Since OECD countries mainly conclude double tax conventions for enhancing the consistency in the exercise of taxing powers on cross-border situations, they also terminate such treaties for replacing them with new rules, which improve this situation. Whilst the conclusion of protocols fine-tunes the treaty to specific needs or problems, the termination of a treaty between OECD countries is an *extrema ratio*. Accordingly, it may occur when the two countries need to completely rediscuss overall conditions agreed (which for instance happened in the case of some economies in transition over the past decade), or when either country fails to execute the treaty in good faith.

Furthermore, in the relations with non-OECD countries the termination of treaties also occurs when either Contracting State feels that the conditions agreed are no longer suitable to achieve a balanced allocation of taxing powers. This occurred for instance when Germany felt that tax sparing clauses were no longer justified in the relations with Brazil, due to the considerable economic development of the latter country as compared to the moment in which those clauses had been negotiated in the framework of a package to promote economic development (Schoueri 2015). We may add here that a developing country should terminate a double tax convention when such country feels that the convention contributes to deprive it of the exercise of its tax sovereignty.

International Tax Coordination

Under the political mandate of the G20, the OECD has produced a dramatic change in the exercise of tax jurisdictions in order to counter base erosion and profit shifting. This project, better known as the BEPS project, was completed with the signature of the BEPS Multilateral Instrument

on 7 June 2017 and plays an important role for international tax coordination (Lang et al. 2017).

This development is important to achieve consistency in cross-border tax treatment, thus preventing the unintended tax advantages that result from the exploitation of tax disparities across the borders and countering tax avoidance more effectively.

The BEPS multilateral instrument should steer double tax conventions towards a much closer coordination of their content. In such context, bilateralism is not meant to disappear but will be exercised in a way that secures the effective countering of undesirable phenomena, such as tax avoidance and aggressive tax planning, along common schemes that constitute global minimum standards, from which states may not deviate. Furthermore, clauses constituting the BEPS minimum standards will no longer require negotiation by the Contracting States but simply apply as a direct consequence of the multilateral instrument. This development may bring double tax conventions back to their original function, which is to counter international double taxation by coordinating the exercise of taxing powers on cross-border situations, leaving it up to multilateral conventions to counter tax avoidance and aggressive tax planning with a single global instrument. There are already signs of a possible development in this direction.

The first sign is connected with tax conventions that regulate mutual assistance between tax authorities. This aspect has been long the object of one ancillary clause (Article 26) within double tax conventions for its instrumental function to securing the correct interpretation of tax treaties and domestic. However, after the establishment of a global standard on tax transparency, a rather large number of states have been showing an inclination to join the multilateral convention on mutual assistance in tax matters. This convention was drafted by the Council of Europe and the OECD, and currently allows for all methods of exchange of information (i.e., upon request, automatic, and spontaneous) and for additional procedures, including assistance in the recovery of taxes.

The second sign arises in the European Union, which is already moving in the direction of multilateralism on specific matters related to the implementation of the BEPS project with the EU Anti-Tax Avoidance Directive, issued on 12 July 2016 (1164/2016). The object and purpose of this Directive is to implement some aspects of the BEPS minimum standards in a potentially homogeneous way within the EU Internal Market. A possible future expansion of the directive can reduce the scope for tax treaties to regulate this aspect within the European Union, thus completing the process outlined above. A global multilateral convention against tax avoidance and aggressive tax planning can achieve a similar effect elsewhere, based on the understanding that one of the typical features of bilateralism is to have states negotiating the content of tax conventions. Therefore, insofar as they may no longer negotiate some aspects, there is also no need for a bilateral tax convention.

However, there are things that international tax coordination cannot and should not change. We refer hereby to the circumstance that the elimination of international double taxation is a structural goal of double tax conventions and this aspect should be tailored to the needs of bilateral relations.

Our view is that this conclusion should also apply and be adapted to the specific needs arising in relations with developing countries. In such context, double tax conventions have traditionally combined the elimination of double taxation with additional goals, in order to strike a fair balance for developing and developed countries when concluding a double tax convention.

For the reasons that we have indicated earlier, capital-exporting countries have a clear convenience in concluding double tax conventions. For capital-importing countries, concluding a double tax convention generally implies a loss of taxing powers and additional expenses connected with the obligation to supply more information to the one that they are interested in receiving. Until now, the application of mechanisms, such as tax sparing, has allowed turning double tax conventions into an instrument for those countries to remain the masters of their own international tax

policy decisions. Since the BEPS project was designed in order to protect the right of the country of value creation from base erosion and profit shifting, it should also be interpreted and adapted to the needs of countries that have the right to pursue their economic development without external interferences in their policy disguised under the need to counter international double taxation. Accordingly, the multilateral framework for combating tax avoidance and aggressive tax planning should make sure that any phenomenon of harmful tax competition is effectively countered. However, double tax conventions with developing countries should regulate all other aspects in line with what can be desirable for such countries and the developed country that is from time to time involved and they should do so without shifting taxing powers from the country of value creation to that from which capital originates.

Cross-References

- [Avoidance](#)
- [Economic Development](#)
- [Fiscal Federalism](#)
- [Tax Evasion by Firms](#)
- [TRIPS Agreement](#)

References

- Baker B (2014) An analysis of double tax treaties and their effect on foreign direct investment. *Int J Econ Bus* 21(3):341–377
- Barthel F, Neumayer E (2012) Competing for scarce foreign capital: spatial dependence in the diffusion of double tax treaties. *Int Stud Q* 56: 645–660
- Blonigen BA, Davies RB (2002) Do bilateral tax treaties promote foreign direct investment? Working paper Nr. 8834. National Bureau of Economic Research, Boston
- Blonigen BA, Oldenski L, Sly N (2014) The differential effects of bilateral tax treaties. *Am Econ J Econ Pol* 6(2):1–18
- Bodin J (1579) *Les six livres de la République*, Jean de Tournes, livre I, chapter 10
- Braun J, Fuentes D (2014) A legal and economic analysis of the Austrian tax treaty network. VIDC, Vienna

- Brooks K (2009) Tax sparing: a needed incentive for foreign investment in low-income countries or an unnecessary revenue sacrifice? *Queens Law J* 34: 505–564
- Chisik R, Davies RB (2004) Asymmetric FDI and tax-treaty bargaining: theory and evidence. *J Public Econ* 88(6):1119–1148
- Court of Justice of the European Union (CJEU) (2006) FII group litigation I, 12 Dec 2006, case C-446/04
- Court of Justice of the European Union (CJEU) (2012) FII group litigation II, 13 Nov 2012, case C-35/11
- Dagan T (2000) The tax treaties myth. *NYU J Int Law Polit* 32:939–996
- Davies R (2003) Tax treaties, renegotiations, and foreign direct investment. *Econ Anal Policy* 33(2): 251–273
- Davies RB, Norbäck PJ, Tekin-Koru A (2007) The effect of tax treaties on multinational firms: new evidence from microdata. Oxford University Centre for Business Taxation, WP 07/21
- Dougherty JC (1978) Tax credits under tax treaties with developing countries. *Int Bus Lawyer* 6(i):28–46
- Dourado AP et al (2017) Tax avoidance revisited in the EU BEPS context, in EATLP international tax series, vol 15 (Munich Congress)
- Easson A (2000) Do we still need tax treaties? *Bull Int Fisc Doc* 54(12):619–625
- Edmiston K, Mudd S, Valev N (2003) Tax structures and FDI: the deterrent effects of complexity and uncertainty. In: *Fiscal studies*, vol 24, pp 341–359
- Egger P, Larch M, Pfaffermayer M, Winner H (2006) The impact of endogenous tax treaties on foreign direct investment: theory and evidence. *Can J Econ* 39(3):901–931
- Herndon JG (1932) Relief from international income taxation: the development of international reciprocity for the prevention of double income taxation. Callaghan and Co., Chicago
- IBFD (2017) Tax research platform. <https://www.ibfd.org/IBFD-Tax-Portal/About-Tax-Research-Platform>
- Janeba E (1996) Foreign direct investment under oligopoly: profit shifting or profit capturing? *J Public Econ* 60:423–445
- Krever R (2016) Chapter 1: general report: GAARs in GAARs – a key element of tax systems in the post-BEPS tax world (Lang M et al (eds), IBFD 2016), Online Books IBFD
- Lahiri AK, Ray G, Sengupta DP (2017), Equalisation levy. In: *Brookings India working paper* 02
- Lang M, Owens J (2014) The role of tax treaties in facilitating development and protecting the tax base, WU international taxation research paper series no 2014–03. Available at <http://ssrn.com/abstract=2398438>
- Lang M et al (2017) The OECD multilateral instrument for tax treaties: analysis and effects. Kluwer Law International, Wien
- Lejour A (2014) The foreign investment effects of tax treaties, CPB discussion paper 265, Netherlands Bureau for Economic Policy Analysis
- Lockwood B (2001) Tax competition and tax co-ordination under destination and origin principles: a synthesis. *J Public Econ* 81:279–319
- Louie H, Rousslang DJ (2008) Host country governance, tax treaties, and US direct investment abroad. *Int Tax Public Financ* 15(3):256–273
- McLure C (2001) Globalization, tax rules and national sovereignty. *Bull Int Tax* 55:328–341
- Millimet D, Kumas A (2017) Reassessing the effects of bilateral tax treaties on US FDI activity. *J Econ Financ*, available online <https://doi.org/10.1007/s12197-017-9400-3>
- Neumayer E (2006) Do double taxation treaties increase foreign direct investment to developing countries? *J Dev Stud* 43(8):1495–1513
- Nguyen HK (2017) Australia's new diverted profits tax: the rationale, the expectations and the unknowns. *Bull Int Tax* 71:2017, no. 9 (accessed 23 Aug 2017)
- Nowotny E, Zagler M (2009) *Der Öffentliche Sektor, Einführung in die Finanzwissenschaft* (The public sector: introduction to public finance), 5th edn. Springer, Berlin
- OECD (2015) Designing effective controlled foreign company rules, action 3–2015 final report, OECD/G20 base erosion and profit shifting project. OECD Publishing, Paris. <https://doi.org/10.1787/9789264241152-en>
- OECD (2017) Model tax convention on income and on capital, 2017 update
- OECD (2017) Multilateral convention to implement tax treaty related measures to prevent BEPS, Paris 7/7/2017
- Paolini D, Pistone P, Pulina G, Zagler M (2016) Tax treaties with developing countries and the allocation of taxing rights. *Eur J Law Econ* 42:383–404
- Pickering A (2013) Why negotiate tax treaties? Papers on selected topics in negotiation of tax treaties for developing countries, paper no.1-N. United Nations, New York/Geneva
- Radaelli CM (1997) The politics of corporate taxation in the European Union, knowledge and international policy agendas. Psychology Press, London
- Rixen T, Schwarz P (2009) Bargaining over the avoidance of double taxation: evidence from German tax treaties. *Public Financ Anal* 65(4):442–471
- Schoueri LE (2015) Arm's length: beyond the guidelines of the OECD: "It is better to be roughly right than precisely wrong." (John Maynard Keynes), in 69 *Bull Int Tax*. 12, Journals IBFD
- Viner J (1950) The customs union issue. Carnegie Endowment for International Peace, Washington, DC
- Vogel J, Ligthart J (2011) The determinants of double tax treaty formation. Tilburg University, Mimeo

Double Tax Treaties

► Double Tax Conventions

Droit de Suite

Nathalie Moureau

Université Paul-Valéry MONTPELLIER 3,
Montpellier, France

An Economic Perspective for a Recurrent Issue: The Legitimacy of the Resale Right “I’ve been working my ass off for you to make all this profit. The least you could do is send every artist in this auction free taxis for a week.”

-Robert Rauschenberg to Robert Schull

NB: Schull originally bought the artwork \$900 in 1958 and resold it for \$85,000 in 1973 (quoted by Wu 1999, p. 531)

Abstract

Resale right consists of a small percentage of the resale price that art market professionals pay to artists at each resale of their works with the involvement of an auction house, gallery, or dealer. Until the new millennium, the resale right was implemented in a small number of countries. In 2014, more than 70 countries have resale rights. The United States, which has been very reluctant toward the adoption of the resale rights, seems to have changed its mind very recently. The debate about the opportunity to implement a resale right is commonly structured around two main axes. The first discusses whether or not visual artists profit from the resale right. The second deals with distortions of trade and competition within different countries that this right could create. While numerous governmental reports and academic research studies concern these two axes, focusing on the effects and consequences of the implementation of a resale right, fewer works deal with its economic rationale.

Synonyms

Droit de suite; Follow-up right; Resale right

Definition

Resale right consists of a small percentage of the resale price that art market professionals pay to

artists at each resale of their works with the involvement of an auction house, gallery, or dealer.

According to the legend, the story began in France with an engraving by Forain titled, “Un tableau de Papa,” depicting two ragged children observing a painting through a window. This scene, which is said to have inspired the resale right, referred to the sale of the *Angelus* by Millet at a record price. Millet originally sold this painting in 1860 for 1,000 francs to the Belgian painter Victor de Papeleu; in 1889, the copper merchant Secretan sold it for 553,000 francs (Fratello 2003), whereas his granddaughter lived in the greatest poverty, selling flowers in the street (<http://bibliotheque-numerique.inha.fr/collecton/12406-un-tableau-de-papa-lere-planche/>) (Fig. 1).

The resale right was at first established in France by the law of the May 20, 1920, and then reaffirmed in 1957 with the law on the literary and artistic property (article L122-8). This right was settled to recognize the particular situation of visual artists who sell their original works and therefore cannot make profit from copies as



Droit de Suite, Fig. 1 Jean Louis Forain (1852–1931) ‘Un tableau de Papa.’ Lithography

other artists usually do. According to the law, visual artists and their beneficiaries receive a small percentage of the resale price of their creation, for a limited period of time; each time their art work is resold through an art market professional. In the beginning, just auction houses were concerned; today gallerists, art dealers, or auctioneers are concerned. Moreover, this right is non-transferable and inalienable.

Belgium (1921) and Czechoslovakia (1926) soon implemented the French legislation adopting similar rules. Internationally, the Berne Convention for the Protection of Literary and Artistic Works included this right in 1948 (Article 14 bis and today article 14 ter because of different minor modifications) after the French proposed to add it to the convention in 1928 (Revision conference in Rome about the Berne convention). Nevertheless, its implementation remains optional, and reciprocity between countries is required for the right to be claimed: “[the right] may be claimed in a country of the Union only if legislation in the country to which the author belongs so permits, and to the extent permitted by the country where this protection is claimed.” Moreover, the convention pointed out “the procedure for collection and the amounts shall be matters for determination by national legislation.”

Practically, until the new millennium, the resale right was implemented in a small number of countries. In Europe, the resale right was enforced in nine countries of the 15 European Union (EU) Member States: Belgium, Denmark, Finland, France, Germany, Greece, Portugal, Spain, and Sweden enforced the right. In Italy and Luxembourg it was not applied because of the lack of precisions for an implementation. Four countries did not apply the resale right: Austria, Ireland, Netherlands, and the United Kingdom. Moreover, practices differed greatly regarding the minimum threshold, the rate in force, the sales concerned (only public auctions in Belgium and France), and even the management of the rights (mandatory, collective, or individual) (Raymond and Kancel 2004). Outside of Europe, some countries had introduced the right in their law, but without an effective implementation. This

was the case of Brazil, Paraguay, Uruguay, Asia, Mongolia, and the Philippines. This right was not recognized in leading places for the art market, notably in the United States, except in California. Mexico and Venezuela were the rare countries outside of Europe that implemented resale rights.

At the turn of the millennium, different events reactivated the debate. In 1996, the European commission proposed a new directive to harmonize the practices in Europe and the Council adopted it in July 2001 (article 48 of the DAVSI Law implementing the directive 2001/84/EC). It plans the payment of royalties on the basis of a scale beginning at 4% for works of art over 3,000 euros to 0.25% for works worth over 500,000 euros and up. All professional resales are affected auction and gallery sales. Moreover, the right is transferred to the heirs for a period up to 70 years after the artist's death. The total amount of the right payable to the artist or his family cannot exceed 12,500 euros. Some adaptations had been allowed in some countries that did not recognize the right previously, notably in the United Kingdom where its adoption had been controversial; during initial implementation, the right only applied to living artists. It has been extended to heirs of deceased artists from the beginning of 2012.

In Europe, the new law brought about many discussions in countries that had previously supported the law, such as France, because of its extension to art galleries. Obviously, the debates had been even stronger in countries such as the United Kingdom, a crucial area for the art market, where the right was not recognized prior to that time (Dallas-Conte and Mc Andrew 2002; Ginsburgh 2005, 2008; Kirstein and Schmidtchen 2001; Pfeffer 2004).

Despite these disputes, a growing number of countries have followed Europe. Today, more than 70 countries have resale rights. The right has been in effect in Australia since June 9, 2010 (George et al. 2009). In China, a specific clause is included in the draft of a new copyright law soon to be submitted to China's State Council, the country's cabinet. China's first copyright law took effect in 1991; the latest draft brings the country closer into line with prevailing European

and American standards. And the United States, a major player in the contemporary art market, which has been very reluctant toward the adoption of the resale rights (Landes 2001), has changed its mind very recently. Indeed, a report published in December 2013 by the US Copyright Office recommends Congress to consider enacting a resale royalty for visual artists. The same organization had declared in 1992 (Claggett et al. 2013), when it had last considered the subject that it was “not persuaded that sufficient economic and copyright policy justification exists to establish resale right in the United States” (Register of Copyright 1992, p. 149). Up to now in the United States, California has been the only state to apply this right, in a soft version, i.e., when the release price records an increase exceeding \$1,000. Currently, the situation could evolve favorably.

The debate about the opportunity to implement a resale right is commonly structured around two main axes presented below. The first discusses whether or not visual artists profit from the resale right. The second deals with distortions of trade and competition within different countries that this right could create. While numerous governmental reports and academic research studies concern these two axes, focusing on the effects and consequences of the implementation of a resale right, fewer works deal with its economic rationale as it is shown in the last section.

Resale Rights: Significant or Lackluster Profits for Visual Artists?

Discounting Effect

Resale rights are introduced in order to increase the artist earnings. Nevertheless, such an introduction tends to lower the market price of first sale. Under a hypothesis of rational expectation, research shows that the buyer takes into account the resale royalty he will pay in the future and then deducts its discounted value from the initial price he would have accepted to pay without such a right. Thus, the wealth an artist can expect from his initial sale is lowered (Filer 1984; Karp and Perloff 1993; Mantell 1995; Perloff 1998).

In the long term, profitability depends on the artist's tolerance of risk. If he is risk adverse, then the introduction of a resale right can induce two negative consequences. Firstly, the artist has no choice but to accept a risky lottery instead of a sure income. And usually, it is easier for collectors compared to artists to bear the risk, because they are often wealthier and more able to diversify their portfolio (Filer 1984; Karp and Perloff 1993; Mac Cain 1994). Secondly, there is what Kirstein and Schmidtchen call a paradox of “risk aversion.” That is to say the artist's lifetime utility may be lowered even if the resale royalty and the incentive effect had a positive net effect on his monetary lifetime income. This result appears when the income of the artist increases over time. Due to risk aversion, the utility function is concave; an additional euro when the income is low can bring more utility compared to when the income is already high (Kirstein and Schmidtchen 2001).

Nevertheless, the hypothesis of risk adversity is controversial. Many studies show that a growing number of artists enter the occupation even if the income distribution is strongly biased toward the lower end of the range. An explanation could be that artists are true risk lovers or that there is a probabilistic bias (Menger 2006), meaning that artists overestimate their chance like lottery players. The other explanations are as follows: artists are “committed to a lobar or love” or “rational fools” (Menger 2006, p. 776). More recently, Wang (2010) showed that the introduction of resale rights increases the artist profit, but lowers the consumer surplus, the whole effect on the social welfare being negative.

Collection Costs

The costs of the implementation of the system are usually deducted before the distribution of royalties. Then, the benefit for the artists might be lowered by important collection costs. According to some authors, these costs are quite high (Ginsburgh 2008; Graddy et al. 2008), whereas others underline the equivalence with perception costs for other intellectual rights, between 12% and 17% in France and 15% in the United Kingdom (DACS 2008; Farchy 2011). The European

Commission came to a similarly ambiguous conclusion in its last report (2011). Whereas some inefficiency in the administration of the system in some countries is recorded, the conclusion remains optimistic, underlying the necessity for an exchange of best practices.

Few Winners

Moreover, as discussed above, cultural markets are structured as stardom markets. Small differences in talent lead to huge differences in earnings, “Sellers of higher talent charge only slightly higher prices than those of lower talent, but sell much larger quantities; their greater earnings come overwhelmingly from selling larger quantities than from charging higher prices” (Rosen 1981). An immediate consequence for the art market is that a large percentage of artists will never benefit from the resale market. Available data about the resale rights distribution among artists support this phenomenon in Australia (Stanford 2003). More recently, a study about the United Kingdom art market in 2006/2007 showed an average payment per work of £693; nevertheless, for 85% of the items, the average payment per work was only £249 versus £3,430 per item for the remaining 15% (Graddy et al. 2008). Other data confirm this disparity. In the United Kingdom, 60% of the artists who received a right earned less than 24£ per artworks, whereas 2% of the artists earned more than 50,000£ (DACS 2008). In France, on the average 68% of the artists earned 1,114 euros, while only 1% earned 15,908 euros. Moreover, the percentage breakdown of sales (in value) submitted to the resale right is 74% for deceased artists and 26% for living artists (Farchy 2011).

Unwaivability, Two-Sided Effects

Another ambiguity lies in the unwaivability of the resale right (Hansmann and Santilli 2001). Some people argue that this unwaivability is necessary for protecting the artist against an unbalanced negotiation with gallerists. A limited number of gallerists face the vast population of artists. Due to the asymmetry of bargaining power, gallerists pay the minimum to the artists, who have no choice

but to accept. According to this reasoning, the discounting effect described in the previous section cannot happen; gallerists cannot lower the price on the first market with a resale right because the price is already fixed at its minimum. Consequently, the resale right is finally helpful for artists. The difficulties of the Projansky agreement could illustrate this unbalanced negotiation and the need for unwaivability. According to this agreement, the artist benefited from some moral rights and would receive 15% of the appreciated value each time a work was transferred; nevertheless, despite a large publicity, this agreement did not encounter a large success. There is also a downside of unwaivability. Notably, it deters artists to indicate the quality of their artwork. According to the theory, the more an artist trusts in his production, the higher the resale right he requires (Hansmann and Santilli 2001). Nevertheless, as the authorities fix the later, this indication is no longer relevant.

Visual Artists' Earnings in Relation to Other Cultural Workers

A central claim for the resale rights rationale is that visual artist cannot benefit from usual protection provided by copyright (reproduction, representation, etc.) as other artists do. A comparison is not easy to conduct because the business models of the other cultural areas differ due to the nature of the product. Recent data offer records of the median wages and salaries of fine artists (including painters, sculptors, illustrators, and multimedia artists, but excluding photographers and graphic designers), writers and authors (including advertising writers, magazine writers, novelists, playwrights, film writers, lyricists, and crossword puzzle creators, among others), and musicians. While visual artists appear poor (\$33,982) by comparison to writers and authors (\$44,792), they are better off than musicians \$27,558 (Nichols 2011). Data from the US BLS confirm that visual artists' earnings are not lower than other creative industry professionals; the median annual wage is \$44,850 (\$54,000 for the mean) for visual artists, \$55,940 (\$68,420 for the mean) for writers and authors, and \$47,350 (\$53,420 for the mean) for composers.

Resale Right: Gravel or Sand in Market Mechanisms?

Distortions of Competition on the International Art Market

The implementation of a resale right increases transaction costs and theoretically possibly reduces the competitiveness of a given country if its competitors do not apply such a right. Indeed, for valued artworks, the expected resale right may overstep sometimes transportation fees so that delocalization of sales appears as profitable.

This argument was at the heart of the European community concerns in 2006 when it decided to harmonize the resale right in Europe. Indeed, the resale right was considered as a crucial factor “which contributes to the creation of distortions of competition as well as displacement of sales within the Community” (European Commission 2011, p. 3). The United Kingdom fought this extension. They did not apply the resale right previously due to the risk of losing competitiveness among other European countries, as well as the United States, all of which are leaders in the art market.

In practice, findings suggest that these concerns were ill-founded. No evidence has been found on a weakened position of the United Kingdom in the international art scene. Surprisingly, according to a study conducted by the IPO, just after the introduction of the resale right, the proportion of eligible works to the resale right in the United Kingdom increased, so did their prices (comparison of the period 2006/2007 with 2003/2004). In the short term, it appears that the implementation of the resale right in the United Kingdom did not have a negative impact on the relative position of its market compared to other countries (Banternghansa and Graddy 2011; Graddy et al. 2008). Conclusions of an EU report in 2011 are less optimistic because of the decrease of the UK’s market share on the international scene between 2008 and 2010 from 34% to 20%. Between 2005 and 2010, UK’s market share decreased from 27% to 20%. Nevertheless, in the same period, the US market share also declined from 54% down to 37%, whereas China

increased its share from 8% up to 24% (European Commission 2011).

Distortions of Competition Between Auctions and Galleries?

Resale right also has indirect effects. The international competition among auction houses depends on their ability to attract sellers and valuable items. Then, in 2007 Christie’s France shifted the economic burden of the royalty from seller to buyer. Nevertheless, such a shift was considered as anticompetitive behavior because an auction sale seemed more attractive to sellers than a sale through a French dealer. Indeed, at auction a seller would receive the hammer price without the deduction of the resale royalty, the latter being paid by the buyer. Whereas with a French dealer, he would receive the price less the resale royalty because dealers charged the resale royalty to the seller according with French law. According to this reasoning, the French Association of Antique Dealers took action against the auction house; the French court of Appeal took the view that parties are not allowed to shift the economic burden of the resale right from the seller to the buyer.

From an economic point of view, and according to auction theory, the buyer bids up to his reservation price. Then, if a resale right is introduced, the buyer will reduce his reservation price by an equivalent amount. The situation is equivalent for the seller regardless of the method of sale used.

Distortion of Competition Between the Art Market and Financial Market?

The relative attractiveness of the art market compared to the financial one is reduced because of an increase in transaction costs. Collectors act on a medium- or a long-term basis and do not necessarily plan to resell their artwork, whereas speculators have short-term views and are motivated by the increased value they will obtain when they resell the item (Kakoyiannis 2006). Thus, an indirect effect of the introduction of the resale right could be to “clean prices,” bringing market prices of artworks closer to their fundamental artistic value.

Resale Right: Is There Any Need to Correct a Market Failure?

Consequences of the Physical Embodiment of the Creation for Visual Art

The main economic rationale that sustains the general copyright lies in the necessity to correct a market failure and the public good property of a creation (nonrivalry and nonexclusivity). This is due to the split existing between a work and its material embodiment. Once a creation is disseminated, anyone can appropriate it and reproduce it at a low marginal cost. Without protection, the risk is high for the creator to not recover his initial investment. Curiously, for visual artists and in the case of the resale right, the idea originally put forward is that because of the uniqueness of the creation they produce, visual artists do not benefit from reproduction rights in the same way as other artists do; above all, the aim of the resale right is to “ensure that authors of graphic and plastic works of art share in the economic success of their original works of art” (European Directive). Nevertheless, for visual artists, the market failure does not exist because the public good property of the creation disappears; no one can copy the creation without sustaining a significant marginal cost. As a consequence, it is not necessarily to artificially create a monopoly for the visual artist on his creation because, by nature, the creation and its physical embodiment are intertwined and uniqueness is one of the major characteristics of the art market. Moreover, prices of different artworks produced by an artist are linked and depend on the artist’s reputation. Then, it does not seem necessary to protect an artist; if the value of one of his artworks goes up, he just has to sell another one on the market to increase his profit. It is a well-known law that on the art market, in the very beginning of an artist’s career, supply exceeds demand, whereas rationing can appear on the market with queuing phenomenon as the artist obtains fame. In other words, if the market power of the artist increases along with his reputation, then it will be easy for him to earn money selling another piece.

Externalities of Future Artworks on Current Ones

If there is a need to correct a market failure with the resale right, it could be the externalities of future artworks on the current ones. Indeed, the artistic recognition of an artist depends on his whole production. Depending on the quality of future artworks, the prices of the current ones can evolve in the future, positively or negatively. These externalities are not taken into account. Because of such failure, there can be underproduction in case of positive externalities. The introduction of a resale right could be a means to internalize these externalities. Nevertheless, the law only considers positive externalities and increases in prices, but not negative ones. If in the future the artist produces artworks of lesser quality, these will lower his global reputation and produce a negative externality for the future market of current artwork. The resale right does not take into account such a negative externality; thus, a risk of overproduction appears.

Visual Art: A Durable Good Monopoly Issue?

Some economists studied the issue from a symmetrical point of view, analyzing if resale royalties have incentive effects on the artists’ output of subsequent decisions. The artist produces a durable good, and when managing his market power, he encounters a dynamic consistency problem of “competing with one’s future self” (Solow 1998). Resale right effects depend on the nature of future artworks; if they are substitutes of current artwork, then the resale right will have a negative impact with a decrease in the artist’s production and an increase in prices. Conversely, if future artworks are complementary, there is an incentive for the artist to increase its production (Solow 1998). In Solow’s analysis, the artist is supposed to be “price setter,” i.e., only well-known artists are able to set their price in the first period. Wang extends the analysis considering not only well-known artists but also new artists (price takers). In both cases, the consequence of the introduction of a resale royalty is to lower the global production and to increase the artist’s lifetime profit. Nevertheless, the rise of the artist’s profit remains

questionable as social welfare globally decreases (Wang 2010).

Resale Right, “Much Ado About Nothing?”

Not only do market failures really apply in the visual arts market, but also resale rights create disincentives for a crucial intermediary for artist recognition, the gallerist (Moulin 1994). Since the beginning of the twentieth century, the art dealer has become a crucial intermediary for the artist’s legitimacy in the market. This changes the analysis significantly. Indeed, under the assumption that the promotion of the value of an artist’s work depends both on the efforts of the artist and of the dealer, it is shown that a specific royalty, i.e., a “share cropping” contract, could be positive. However, assuming that the promotion of the artwork’s value depends solely on the dealer, the resale right is totally counterproductive (Kirstein and Schmidtchen 2001).

It is difficult to draw a clear conclusion. Both the benefits and costs are lower than expected, and then a balance between the two parts becomes plausible. But at the same time, why discuss a government intervention on the market when real market failure does not affect the art market? Probably, the record of lower costs than initially expected and the symbolic reward given to the artist through the resale right help explain the general movement for its implementation on an international level. Nevertheless, it is important to take into account the role of imitation; we know that imitating the actions of others can be a rational behavior to improve one’s own information in case of uncertainty. The last US Copyright Office report seems to adopt such a rule when writing “at the same time recent developments – including in particular the adoption of resale right laws by more than thirty additional countries since the Office’s prior report – would seem to warrant renewed consideration of the issue.” Nevertheless, one must be careful and keep in mind that imitation can also lead to a misinformed cascade of followers, causing the vast majority of the population to make bad decisions (Bikhchandani et al. 1992).

References

- Banternghansa C, Graddy K (2011) The impact of the Droit de Suite in the UK: an empirical analysis. *J Cult Econ* 35(2):81–100
- Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom and cultural change as informational cascades. *J Polit Econ* 100(5):992–1026
- Claggett K et al. (2013) Resale royalties: an updated analysis. United States Copyright Office
- Dallas-Conte L, Mc Andrew C (2002) Implementing Droit de Suite (artists’ resale right) in England. The Arts Council of England Report 28-
- Design and Artist Copyright Society (2008) The artist’s resale right in the UK DACS. Available at <http://www.parliament.nz/resource/0000131776>
- European Commission (2011) Report on the implementation and effects of the Resale Right Directive. European Commission Report (2001/84/EC)
- Farchy J (2011) Le droit de suite est-il soluble dans l’analyse économique? ADAGP report. Available at <https://circabc.europa.eu/sd/d/c8fa35dc-59eb-4c57-bc7f-ec9a26a02f4d/ADAGP.pdf>
- Filer R (1984) A theoretical analysis of the economics impact of artists’ resale royalties legislation. *J Cult Econ* 8:1–28
- Fratello B (2003) France embraces Millet: the intertwined fates of “The Gleaners” and “The Angelus”. *Art Bulletin* 85(4):685–701
- George J et al (2009) Resale royalty right for visual artists Bill 2008, House of representatives standing committee on climate change, water, environment and the arts. The Parliament of the Commonwealth of Australia, Canberra, A.C.T
- Ginsburgh V (2005) Droit de suite. an economic viewpoint, The modern and contemporary art market. The European Fine Art Foundation, Maastricht
- Ginsburgh V (2008) The economic consequences of the droit de suite in the European Union. In: Towse R (ed) Recent developments in cultural economics. Edward Elgar, Cheltenham/Northampton, pp 384–393
- Graddy K, Horowitz N, Szymanski S (2008) A study into the effect on the UK art market of the introduction of the artist’s resale right. IP Institute
- Hansmann H, Santilli M (2001) Royalties for artists versus royalties for authors and composers. *J Cult Econ* 25(4):259–281
- Kakoyiannis J (2006) Resale royalty rights and the context and practice of art bargains. <http://www.jequ.org/index.php?/links>
- Karp LS, Perloff JM (1993) Legal requirements that artists receive resale royalties. *Int Rev Law Econ* 13:163–177
- Kirstein R, Schmidtchen D (2001) Do artists benefit from resale royalties? An economic analysis of a new EU directive. *Law Econom Civil Law Countr* 6:257–274
- Landes W (2001) What has the visual artist’s rights Act of 1990 accomplished? *J Cult Econ* 24(4):283–306
- Mac Cain R (1994) Bargaining power and artist’ resale dividends. *J Cult Econ* 18:108–112

- Mantell E (1995) If art is resold, should the artist profit? *Am Econ* 39(1):23–31
- Menger P-M (2006) Artistic labor market: contingent work, excess supply and contingent risk management. In: Ginsburgh V, Throsby D (eds) *Handbook of the economics of art and culture*, vol 1. North Holland/Amsterdam, pp 766–812
- Moulin R (1994) The construction of art values. *Int Sociol* 9(1):5–12
- Nichols B (2011) Artists and art workers in the United States, Research note 105. National Endowments for the Arts, Washington, DC
- Perloff JM (1998) Droit de suite. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*. Macmillan, New York, pp 645–648
- Pfeffer J (2004) The costs and legal impracticalities facing implementation of the European Union's droit de suite directive in the United Kingdom. *NWJILB* 24(2): 533–561
- Raymond M, Kancel S (2004) Le droit de suite et la protection des artistes plasticiens. Inspection des affaires sociales (rapport 2004/039), inspection générale de l'administration des affaires culturelles (rapport 2004/12)
- Register of Copyright (1992) Droit de suite: The artist's resale royalty. US Copyright Office
- Rosen S (1981) The economics of superstars. *Am Econ Rev* 71(5):845–858
- Solow J (1998) An economic analysis of the droit de suite. *J Cult Econ* 22:209–226
- Stanford JD (2003) Economic analysis of the droit de suite the artist's resale royalty. *Aust Econ Pap* 42(4): 387–398
- Wang G (2010) The resale royalty right and its economics effects. *J Econ Res* 15:171–182
- Wu JC (1999) Art resale rights and the art resale market: a follow-up study. *J Copy Soc USA* 46:531–552

Drug Price Regulation

Jean-Michel Josselin¹, Laurie Rachet Jacquet²,
Véronique Raimond³ and Lise Rochaix²

¹University of Rennes 1 and CREM UMR-CNRS, Rennes, France

²Hospinnomics, Paris School of Economics, Paris, France

³Haute Autorité de santé, Saint-Denis, France

Abstract

Drug prices are regulated in a legal framework that organizes the negotiation between pharmaceutical firms and a third-party payer

responsible for healthcare reimbursement. This regulation aims at compensating for market failures associated with drug specificities. Explicit economic reasoning through the so-called health technology assessment framework is increasingly embedded in the institutional and administrative process of the evaluation procedure leading to market access, pricing and reimbursement for new drugs.

Definition

Drug prices are regulated in a legal framework that organizes the negotiation between pharmaceutical firms and a third-party payer responsible for healthcare reimbursement. This regulation intends to optimize the use of limited public resources in the provision of healthcare. Private market pricing would fail to adequately take account of the specificities of medicines as vectors of health improvement.

Regulatory tools are thus necessary to ensure that the allocation of resources to medical interventions is welfare enhancing. An increasing number of countries have adopted regulation laws for the reimbursement of drugs that rest not only on medical prerequisites but also increasingly on cost-effectiveness requirements and budget impact analyses. Explicit economic reasoning through the so-called health technology assessment framework is increasingly embedded in the institutional and administrative process of the evaluation procedure leading to market access, pricing, and reimbursement for new drugs.

Why Should Drug Prices Be Regulated?

This question is rooted in the broader setting of healthcare provision. Healthcare comprehends “those goods and services whose primary purpose is to improve -or prevent deterioration in- health” (Hurley 2000, p. 67). As a medical intervention, healthcare consists of procedures (e.g., surgery), care (e.g., nursing and care follow-up), programs (e.g., screening or vaccination), drugs and medical devices, etc. It is also an economic good as it is

provided in a context of scarce resources compared to their potential uses. There is thus always an opportunity cost to allocating resources to a specific use and giving up the advantages of waived alternatives. Healthcare is a combination of rival and exclusive private goods: for instance, individual treatment is rival, and drug pricing may prevent access to it. In this context where drugs are taken as economic and private goods, and at first glance, private market pricing could appear as a natural candidate for an efficient allocation of resources.

However, characteristics pertaining to the specific nature of healthcare and pharmaceuticals – these are not standard market goods – preclude private market pricing. Market mechanisms for resource allocation rest, among other things, on individual preferences, as the consumption of private goods directly impacts individuals' utility. Healthcare and more specifically medicines do not as such provide utility, as in the case of painful treatments, because the desired commodity is not medical consumption but health improvement. Painkillers or adverse side effects constitute an exception as they directly affect utility while they also contribute, if the treatment is eventually more effective than harmful, to health improvement. The demand for healthcare and drugs is fundamentally a derived demand for health (Grossman 1972). Since health is neither tradable nor transferable, the demand-supply framework for valuing goods, in this instance, pharmaceuticals, through market prices is not suited. Health is thus a primary commodity that requires the consumption of goods and services, among which are drugs, along with other determinants with individual and collective dimensions such as lifestyle, genetic endowment, education, and safe environment, among the most important.

An optimal market equilibrium is such that “if a competitive equilibrium exists at all, and if all commodities relevant to cost and utilities are in fact priced by the market, then the equilibrium is necessarily [Pareto] optimal in the following precise sense that there is no other allocation of resources to services which will make all participants in the market better off” (Arrow 1963, p. 942). Drug and healthcare specificities prevent

competitive market pricing from optimally allocating healthcare resources. However, non-competitive mechanisms for resource allocation, such as drug price regulation, can be welfare improving, which thus call for nonmarket interventions.

Not only healthcare needs cannot be anticipated, but the consequences of healthcare consumptions in terms of life expectancy, future consumption and utility, labor supply and productivity, healthcare provision level, and scope are uncertain (Meltzer 1997). The immense variety of health risks and informational asymmetries (adverse selection into insurance schemes, uncertainty about the individuals' risk profiles, moral hazard in health-related behavior) impede the establishment of a comprehensive competitive insurance market system. In addition to these aspects, the system may face unsustainable risk-bearing markets as in the case of orphan diseases or pandemic noncommunicable diseases like type 2 diabetes. As a consequence, the need for regulation of insurance access is often typically addressed through risk pooling by a third-party payer (Morris et al. 2007): individual insurance premiums feed a common fund managed by that third-party responsible for reimbursing healthcare providers once they have provided treatment to patients.

The healthcare sector is also characterized by substantial externalities, as, for instance, in immunization against communicable infectious diseases or through the productivity improvement associated with enhanced health status in the general population. In addition, current demand does not fully reflect potential demand: the availability of healthcare (e.g., through the permanent presence of hospitals) is valued as such and separately from its actual usage given the infrequent nature of medical interventions. The internalization of those external effects requires nonmarket institutions. Finally, asymmetric information between producers and patients-consumers mostly occurs during the physician-patient encounter. The need for drug licensing and control of prescription practices arises from the fact that patients' interest may be balanced with manufacturers' profit and physicians' own welfare, thus creating a risk of

supplier-induced demand and biased medical prescription.

As a consequence, drug provision calls for price regulation and public intervention. How are these aspects grounded in economic theory? More specifically, how is drug regulation related to health technology assessment standards, namely, cost-effectiveness and budget impact analyses? What are the institutional translations of such regulatory requirements?

How to Regulate Drug Prices?

One should distinguish between the control of the provision of prescription drugs to patients and the regulation of their price. The regulation of drug supply aims at several purposes. From a public health perspective, it is meant to guarantee a minimal level of quality and safety of the product in itself as well as in its usage by the various potential subgroups of patients. In this respect, the main tool of regulation is the marketing authorization based on clinical trials, but production is controlled and safety still monitored while the drug is marketed. The distribution of prescription drugs is usually authorized for a monopoly of chartered pharmacists, advertisement is controlled, and indications can be limited through guidelines. Mandatory and optional insurance schemes define the perimeter of reimbursed medicines. Early access to innovation is increasingly facilitated through compassionate use programs or dedicated patient access schemes, while in the meantime, following sanitary scandals in the past decades, the strain on product safety agencies has increased toward maximal safety guarantee. From a macro-economic perspective, the supervision of drug provision intends to ensure the sustainability of total health expenditures while encouraging innovation through the patents system, antitrust regulation, and public funding for fundamental clinical research. Transnational regulations in healthcare remain limited (examples are the European Marketing Authorization or the international protocols for clinical trials), and countries still differ in their institutional choices. The regulation of prescription drugs remains largely national as it reflects

differing social insurance choices and is, to some extent, embedded in cultural preferences.

Drug price regulation is meant to ensure that once a new drug has been assessed with a significant degree of medical efficacy and improvement compared to the existing drugs with similar therapeutic indications, the outcomes of its use are worth the cost incurred by the third-party payer. Health technology assessment provides methods for the economic evaluation of disease treatment and follow-up or of prevention by screening or vaccination (Drummond et al. 2015). The two main methods are cost-effectiveness analysis and budget impact analysis.

Cost-effectiveness analysis compares the relative costs and outcomes of two or more strategies (or treatment options) competing for the implementation of a health program. Contrary to cost-benefit analysis, it does not use an estimation of the equivalent money value of the outcomes. The standard measure of effectiveness is the number of life years gained by the patients, which can be adjusted by the quality of life in order to produce a cost-utility analysis. Cost-effectiveness analysis is both comparative (selecting a strategy implies that the net advantages of the waived ones are given up) and consequentialist (the focus is on maximizing the outcome from the available resources). This opportunity cost approach was initially conveyed through the incremental cost-effectiveness ratio. Its numerator expresses the cost difference when moving from one strategy to another, and the denominator is the difference in effectiveness. The ratio is then compared to the collective or decision-maker's marginal willingness to pay for an additional unit of effectiveness. One strategy will be preferred to another if its incremental cost-effectiveness ratio is lower than the collective willingness to pay or efficiency threshold. If no such value is available or if the decision-maker wishes to consider a range of thresholds then a second and more general indicator, the incremental net benefit provides a linear rearrangement of the incremental cost-effectiveness ratio. It is the subtraction of, on the one hand, the difference in effectiveness valued by a predefined collective willingness to pay from, on the other hand, the difference in costs.

If the incremental net benefit is positive, then the switch to the new strategy is accepted. When comparing several strategies for a given threshold, the one with the highest benefit should be selected.

The simultaneous comparison of all the competing strategies can be summarized in an efficiency frontier that plots differential effectiveness against differential cost with the least effective treatment option as the reference and origin of the graph. The method allows discriminating between efficient (cost-effective) strategies located on the frontier and those out of the frontier. The latter are said to be “strictly dominated” if they yield higher cost and lower effectiveness than another strategy. They are subject to “extended dominance” if their incremental cost-effectiveness ratio is greater than that of the next more effective strategy. The efficiency frontier allows sorting treatment options independently of any specific value of collective marginal willingness to pay.

The second method in health technology assessment is budget impact analysis (Mauskopf 1998; Mauskopf 2014; ISPOR 2014). It examines the extent to which the introduction of a new treatment option in addition to the existing ones affects the third-party payer’s budget. The new strategy may contribute to reshuffle the supply shares in the set of treatment options, change outcome achievements, as well as the expected mid-run budget burden. The budget impact is calculated as the difference between the expenses borne by the third-party payer before and after the introduction of the new drug and the associated treatment option.

How Is Health Technology Assessment Used in Drug Price Regulation?

Health economic evaluation for efficient resource allocation is steadily promoted by international joint actions like the European Network for Health Technology Assessment (EUnetHTA) including 33 countries among which are Russia and northeastern nations. Twenty-five of them have issued one or several methodological

guidelines (Heintz et al. 2016). They are directed to pharmaceuticals specifically or have a more general purpose including all types of healthcare technologies. For the vast majority of them, their purpose pertains to reimbursement or price negotiation with only a handful of guides with a scope limited to information. A majority of guidelines has a mandatory status as opposed to the less stringent “recommendation” status. In what follows, we review a number of country case studies as representative, but not exhaustive, of the contrasted pathways that can be followed by drug price regulation institutional processes.

England is the historically leading proponent of the inclusion of health technology assessment into the process of healthcare resource allocation. After attempts at efficiency analysis as far back as the early 1970s, the creation in 1999 of the National Institute for Clinical Excellence (NICE) constitutes a landmark and the institutional recognition that the admission of drugs to reimbursement should be grounded on both medical and economic appraisal. The Health and Social Care Act of 2012 has extended and reinforced NICE (incidentally, the same acronym now stands for National Institute for Health and Care Excellence), with an emphasis on recommendations on the uncertainty surrounding the efficiency of the evaluated drug or medical device. Assessments are conducted by the pharmaceutical companies and appraised by NICE in reference with guidelines periodically published. The evaluation rests on the calculation of an incremental cost-utility ratio expressed in cost per QALY. Health technology assessments admittedly have little leverage in the financial negotiations between the National Health Service (NHS) and the manufacturers in the framework of the 5-year Pharmaceutical Price Regulation Scheme. However, they are crucial as they determine the set of healthcare services included in the NHS reimbursement design (Chalkidou et al. 2009; Sorenson and Chalkidou 2012). In contrast with almost all other countries using cost-effectiveness evaluation, NICE explicitly refers to a decision threshold set between £20,000 and £30,000 (McCabe et al. 2008). Any submitted new drug or medical device is either integrated into that scheme, rejected, or

subjected to additional clinical or economic data collection.

In France, cost-effectiveness analysis plays an increasing role in the price negotiation process, compared with England, in a progressive attempt to reconcile healthcare quality and sustainability. The National Health Insurance Reform law of 2004 created the French National Authority for Health (*Haute Autorité de Santé*, HAS). The process was later finalized by the 2012 Social Security financing law and its application decree relative to the health-economic mandate of HAS. Since October 2013, the economic evaluation and public health committee of HAS provides the interministerial Healthcare Products Pricing Committee (*Comité Economique des Produits de Santé*, CEPS) with a cost-effectiveness opinion on those drugs and medical devices which are expected to have a significant medical benefit and impact on the health insurance budget. Reimbursement decisions by the national health insurance scheme are based on the actual clinical benefit and the added clinical benefit with regard to existing treatment options, as appraised by the HAS medical committees. Cost-effectiveness opinions are used by the CEPS, together with other criteria (added clinical benefit, price of comparators, expected sales volume, and European reference price) in the price negotiation process. A feature of economic appraisal that France shares with England and several other countries is the growing interest in the exploration and documentation of uncertainty surrounding the cost-effectiveness outcomes of the manufacturers' submissions. Indeed the greater that uncertainty, the weaker should the price claim be.

The German regulatory framework for pharmaceutical prices currently rests on a strict assessment of clinical benefit (through the so-called early benefit assessment procedure) and a posteriori cost-containment measures. It is characterized by a quasi-absence of health economics criterion, despite several attempts by the regulatory power to include cost-effectiveness assessment for medicines (Klingler et al. 2013). The Institute for the Quality and Efficiency of Healthcare Services (*Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen*,

IQWiG) was created in 2004 as an independent scientific body for health technology assessment. In the 2007 Statutory Health Insurance Act to Promote Competition (§35b), IQWiG was explicitly given the objective to set maximum reimbursement prices for pharmaceutical and non-pharmaceutical products. The landmark 2010 Pharmaceutical Market Reorganization Act (AMNOG) however substantially reduced the role of health economics evaluation and set forth the importance of drugs clinical assessment as a criterion for the price negotiation between the pharmaceutical firms and the Federal Joint Committee (G-BA). Health economics evaluations conducted by IQWiG now amount to interventions of the last resort, to which recourse may be sought by the Federal Association of Sickness Funds or by a drug manufacturer, following a failure in price negotiations and a decision by an arbitration committee.

Beyond the current sound financial situation of the sickness funds, several historical and ethical factors have been invoked in the literature (Klingler et al. 2013; Gerber-Grote et al. 2014) to explain the reluctance to give economic evaluation a more substantial role in the German drug regulation system. Incidentally, despite the absence of a formal role for efficiency evaluation, medicines, even patented medicines, that fail to demonstrate additional clinical benefit cannot obtain higher prices than their comparators, by virtue of the reference pricing procedure.

Health coverage is fragmented in the United States of America, with a historically high level of private activity and a shared power between federal and state governments regarding healthcare regulation. The use of cost-effectiveness analysis exemplifies the decentralized organization of healthcare provision and coverage. Drugs are mainly regulated at the federal level to guarantee effectiveness and safety (Rice et al. 2013). A fast track process designed by the federal Food and Drug Administration facilitates early access for patients to drugs addressing unmet needs. In contrast, the regulation of prices varies widely among health insurance systems. Health coverage is divided between private employment-based or direct-purchase insurance and welfare programs,

mainly Medicare (covering people over 64 years), Medicaid (covering people with low income), and the State Children's Health Insurance Program. Prescription drug prices are not directly regulated. The price and patient-co-pay for a given drug vary and depend on the insurance plan that covers the expense. Importation of drugs from a country in which drugs are sold at a lower price is prohibited in the United States.

Regarding Medicare, private plans compete on costs and coverage and separately negotiate drug prices with pharmaceutical companies (Walton et al. 2017). The noninterference clause stipulates that the federal administration for health may not interfere with the negotiations between drug manufacturers, pharmacies, and plan sponsors. The Federal Medicaid Drug Rebate Program, for which pharmaceutical firms must apply to have their drugs reimbursed, guarantees that Medicaid, as well as the Department of Veterans Affairs, will not be charged more than the lowest price available to private payers; most states negotiate further rebates for Medicaid drugs plans (Rice et al. 2013).

Health Technology Assessment is remarkably referred as "comparative effectiveness research" in the United States and rarely includes cost-effectiveness evaluation. It is mostly performed by insurers, pharmacy benefit managers, and non-profit organizations including the Patient-Centered Outcomes Research Institute (PCORI) created by the Patient Protection and Affordable Care Act (Chalkidou et al. 2009). The 2010 Patient Protection and Affordable Care Act states that cost-benefit analyses are not allowed for healthcare practice or reimbursement in the Medicare program. However, cost-effectiveness has been successfully established in other areas of health administration: Veterans Health Administration and the Military Health System conduct health technology assessments on pharmaceuticals, operated by the Pharmacy Benefits Management Strategic Healthcare Group and the Department of Defense Pharmacoeconomic Center. States as well can perform or commission cost-effectiveness evaluations to support Medicaid programs administration. The development of the private Institute for Clinical and Economic

Review reveals the demand for cost-effectiveness evaluation in a context of high-price therapeutic innovation (Walton et al. 2017). Eventually, cost-effectiveness evaluation is also performed by the pharmaceutical industry itself, partly to address foreign national regulatory requirements.

Budget impact analysis is currently mandatory in less than 20 countries including Germany, Malaysia, Mexico, Norway, Poland, Thailand, etc. It is still often viewed as a complement to cost-effectiveness analysis and, as such, substantially varies in the extent to which it follows international methodological recommendations (ISPOR 2014). Examples of such flaws in the case of the USA are provided by Mauskopf and Earnshaw (2016). However, when adequately performed, budget impact analysis does provide an outline of the financial burden that is likely to be faced if the evaluated product is to be added to the existing set of interventions. Admittedly, market shares after the introduction of the new drug are subject to structural uncertainty; it remains true that financial projections provide indispensable estimations of the impact on annual healthcare budgets and population health.

The total budget effect of introducing a new treatment option among existing ones naturally questions the affordability and sustainability of the corresponding financial effort. In this respect, budget impact analysis is descriptive rather than prescriptive: it provides information (undeniably surrounded by uncertainty) to the third-party payer, but does not infer any decision from it. The decision is left to the health insurance agency to accept or not the new product into the reimbursement scheme and at what price. In the case of highly expensive innovative drugs, even a relatively small number of patients can nevertheless bear a significant budget burden. For non-communicable and especially chronic diseases with high prevalence and growing incidence, comparatively small drug price variations can have a huge budgetary impact.

Questions about the affordability and sustainability of healthcare innovations are also present in the practical implementation of cost-effectiveness analysis. Incremental cost-effectiveness ratios consider the collective willingness to pay as a

parameter against which decision is made to switch to the new treatment strategy if the ratio is below the efficiency threshold. The incremental net-benefit approach considers the threshold as a variable, but a decision cannot be made without defining a precise value or range of values of collective willingness to pay. The efficiency threshold is thus, in theory, crucial for decision-making. In practice, however, a majority of countries does not provide a value for this threshold, or if it does, it is susceptible to many exceptions (innovative cancer treatments, orphan diseases, etc.). Methodological ambiguities are here reflected in institutional (non)-choices. More than the expression of citizens' willingness to pay, represented by the third-party payer protecting them as patients, the efficiency threshold is a matter of marginal productivity of healthcare provision and opportunity cost in a given macroeconomic context. In times of expansion, the threshold should increase; a lower value would accompany periods of budget contraction. There should also be a trend toward lower values as the productivity of the healthcare system increases: efficiency requirements for new treatments should be more stringent; otherwise the opportunity cost of their adoption would increase. In practice, there is a general trend in most countries to accept the reimbursement of innovative drugs with increasingly higher incremental cost-effectiveness ratios compared to existing treatments. This acceptance could reflect the choice of citizens or at least of decision-makers to give more weight to other legitimate criteria than the efficiency.

Conclusion

Since health is a primary commodity whose characteristics cannot be reduced to those of a private good, provision of healthcare cannot be left to private markets only. The ensuing regulation of healthcare prices and specifically of drug prices increasingly involves assessments by national evaluation agencies more or less dependent on the government. The methodological framework for national regulations usually abides by international recommendations, even though the national political and administrative context remains

significant in the shaping of the evaluation process. The legal framework for the assessment protocol explicitly includes economic reasoning, which is quite innovative in the broader field of public intervention where concerns about efficient resource allocation are often absent from the decision-making process. The rationale behind this explicit and legally binding inclusion of both medical and economic criteria has sometimes stemmed from budgetary pressures due to the contraction of available resources as well as from political pressure to rationalize the use of scarce collective resources.

Nevertheless, the progressive inclusion of economic criteria in national healthcare policies faces a number of hurdles. Price regulation, combining medical and economic assessment tools, takes place in a changing legal process, with the creation of agencies, usually independent from governments, and granted with varying mandates. The feedback from two or three decades of contrasted countries' experience shows that health technology assessment still does not play a full role in the allocation of healthcare resources. The inclusion of economic evaluation in drug price regulation has been part of a response to the challenge of increasingly high-cost treatments, but economic rationale through health technology assessment still has limited leverage on pricing decision (Franken et al. 2016). Yet, there is fast learning by doing, both for the healthcare institutions themselves and for their evaluation methods, due to the great challenges faced. In methodological terms, the quality of cost-effectiveness data appears as one of the most important stakes, along with the role of uncertainty surrounding efficiency outcomes in the price negotiation process. As for institutions, national regulatory agencies will increasingly have to find ways to conciliate country specificities with the necessity to provide joint and fast responses to the price claims of international firms.

Cross-References

- [Administrative Law](#)
- [Cost-Benefit Analysis](#)

- [Efficiency, Types of](#)
- [Market Failure: Analysis](#)

References

- Arrow K (1963) Uncertainty and the welfare economics of medical care. *Am Econ Rev* 53:941–973
- Chalkidou K, Tunis S, Lopert R, Rochaix L, Sawicki P, Nasser M, Xerri B (2009) Comparative effectiveness research and evidence-based health policy: experience from four countries. *Milbank Q* 87:339–367
- Drummond M, Sculpher M, Claxton K, Stoddart G, Torrance G (2015) *Methods for the economic evaluation of health care programmes*. Oxford University Press, New York
- Franken M, Heintz E, Gerber-Grote A, Raftery J (2016) Health economics as rhetoric: the limited impact of health economics on funding decisions in four European countries. *Value Health* 19:951–956
- Gerber-Grote A, Sandmann G, Zhou M, Thoren T, Schwalm A, Weigel C, Balg C, Mensch A, Mostardt S, Seidl A, Lhachimi K (2014) Decision making in Germany: is health economic evaluation as a supporting tool a sleeping beauty? *Z Evid Fortbild Qual Gesundhwes* 108:390–396
- Grossman M (1972) On the concept of health capital and the demand for health. *J Polit Econ* 80:223–255
- Heintz E, Gerber-Grote A, Ghabri S, Hamers F, Prevolnik Rupel V, Slabe-Erker R, Davidson T (2016) Is there a European view on health economic evaluation? Results from a synopsis of methodological guidelines used in the EUnetHTA partner countries. *PharmacoEconomics* 34:59–76
- Hurley J (2000) An overview of the normative economics of the health sector. In: Culyer A, Newhouse J (eds) *Handbook of health economics*. Elsevier, Amsterdam, pp 55–118
- ISPOR (International Society for Pharmacoeconomics and Outcomes Research), Sullivan S, Mauskopf J et al (2014) Budget impact analysis-principles of good practice: report of the ISPOR 2012 budget impact analysis good practice II task force. *Value Health* 17:5–14
- Klingler C, Shah M, Barron J, Wright S (2013) Regulatory space and the contextual mediation of common functional pressures: analyzing the factors that led to the German efficiency frontier approach. *Health Policy* 109:270–280
- Mauskopf J (1998) Prevalence-based economic evaluation. *Value Health* 1:251–259
- Mauskopf J (2014) Budget-impact analysis. In: Culyer A (ed) *Encyclopedia of health economics*, vol 1. Elsevier, Amsterdam, pp 98–107
- Mauskopf J, Earnshaw S (2016) A methodological review of US budget-impact models for new drugs. *PharmacoEconomics* 34:1111–1131
- McCabe C, Claxton K, Culyer A (2008) The NICE cost-effectiveness threshold: what it is and what that means. *PharmacoEconomics* 26:733–744
- Meltzer D (1997) Accounting for future costs in medical cost-effectiveness analysis. *J Health Econ* 16:33–64
- Morris S, Devlin N, Parkin D (2007) *Economic analysis in health care*. Wiley, New York
- Rice T, Rosenau P, Unruh L, Barnes A, Saltman R, van Ginneken E (2013) United States of America: health system review. *Health Syst Transit* 15:70–80
- Sorenson C, Chalkidou K (2012) Reflections on the evolution of health technology assessment in Europe. *Health Econ Policy Law* 7:25–45
- Walton S, Basu A, Mullahy J, Hong S, Schumock G (2017) Measuring the value of pharmaceuticals in the US health system. *PharmacoEconomics* 35:1–4

Ecolables: Are they Environmental-Friendly?

Lisette Ibanez

LAMETA, CNRS, INRA, Montpellier SupAgro,
Université de Montpellier, Montpellier, France

Abstract

This article provides a general overview of the technical, economical, regulatory and environmental aspects of ecolabeling. An ecolabel is a market-based policy instrument that can be either voluntarily adopted or mandated by law. Ecolabels are applied to services and products in order to inform consumers of their environmental-friendliness and to avoid market failures. In reality, however, ecolabels do not always succeed in achieving environmental improvements. The mis-use of environmental standards, the practice of strategic manipulations that create trade-distortions, the excessive use of claims, and behavioural biases are some of the factors that can prevent an ecolabel from reaching its initial objective to reduce or even eliminate environmental externalities.

History

The phenomenon of ecolabeling had its beginning in the 1970s, driven by an increasing consumer willingness to pay for sustainable consumption.

Until that time, the management of environmental externalities relied mainly on the use of regulatory mechanisms that imposed requirements or restrictions on the operations of potentially polluting firms and entities.

The first official ecolabeling program, Blue Angel, was launched in Germany in 1977 and was followed by many other ecolabeling schemes in various countries and areas. Today, its development has permeated many sectors such as the food industry, cosmetics, home building, and even the automobile industry. After the 1992 UN Earth Summit conference in Rio, in which no world-wide agreement was reached for environmental protection, various initiatives (by governments, NGO's, private stakeholders, etc.) aimed to integrate environmental issues into manufacturing procedures. The idea was to influence consumption patterns in order to achieve sustainable development.

Since then, a burst of interest emerged for self-regulation and new environmental policy instruments, including ecolabeling, and an important change in manufacturing was observed in the United States: The percentage of new products claiming to be environmentally friendly rose from 1.1% in 1986 to 9.5% in 1999. Today, ecolabels are found everywhere: on a large variety of products and services, appearing on packaging, manufacturer's websites, or related to countries, or geographical areas. In total, 458 official ecolabeling schemes concerning 25 industry sectors have been reported in 195 countries

(Cf. <http://www.ecolabelindex.com/consulted> on October 2, 2014.). What's more, a great number of unofficial environmental claims are currently in use, as well.

Definition

Broadly speaking, ecolabels are claims related to environmental friendliness. They inform consumers about the environmental quality of products and services, enabling consumers to make better choices (Thøgersen et al. 2010). It is widely understood that people rely heavily on information provided by labels in making their consumption decisions. Nevertheless, the word ecolabel remains a fuzzy and ill-defined term that can encompass a variety of different meanings. It includes a diversity of environmental claims ranging from third-party certification schemes to self-declaratory statements. Thus, at one extreme, ecolabels are issued by independent organizations (and can be voluntarily adopted by firms), as for example, Forest or Marine Stewardship Council Certifications (FSC and MSC). At the other extreme, ecolabels can be simple marketing claims or logos related to some aspect of environmental friendliness and are often vague, undefined, unverified, and/or unverifiable. Examples of these types of ecolabels include statements on detergent boxes indicating that the product contains no phosphates, or a label on a washing machine claiming that it is energy efficient.

In general, the conceptual design of ecolabeling schemes includes four stages (Grolleau et al. 2007). The first stage involves the selection of product categories or services to which the ecolabel will apply. The second stage defines the criteria that the product, services, or production process must respect in order to be allowed to use the ecolabel. Both of these stages, i.e., choice of product categories and ecolabel criteria, may be established by governments, environmental organizations, consumer groups, industries, or some combination of these parties. The third stage concerns the inspection of the products and services according to the defined criteria. This stage establishes the methods by which eligible

products and services are to be tested and/or audited. The final stage develops the communication strategies or the way in which the environmental attributes will be signaled to the public.

Of course, the seriousness and accuracy of each of these stages vary highly according to types of environmental claims and labels and can therefore result in a wide range of compliance costs. In general, ecolabeling costs include the opportunity costs associated with efforts to conform to relevant standards (and thus testing and labeling) as well as the costs associated with efforts to create new markets. This means that firms must bear the costs of adjusting production processes in order to assure that products will earn the ecolabel, as well as the expenses of subscribing to and maintaining participation in these ecolabeling programs. Fees for these programs can be significant. Thus, under the assumption that firms are rational profit maximizers, their voluntary adoption of an ecolabel is conditional on the prospect of sufficient sales, higher prices, and due recognition of their environmental status by stakeholders and users.

Classification of Ecolabels

There are various ways of classifying ecolabels: they may be classified according to their legal status (mandatory or voluntary), the certifying entity (governmental, nongovernmental, or private rule-setting), the certification criteria (single or multiple criteria, product or process oriented, fixed or evolutionary mechanism), and even the geographical dimension of their production (regionally, nationally, or internationally defined).

In the interest of increasing the clarity and transparency of the ecolabeling process, the International Standard Organization has undertaken efforts to standardize ecolabeling principles. Specifically, it distinguishes between three types of environmental product labels, classified as ISO Type I, II, and III (Lathrop and Centner 1998). These three types of labels can be described as follows:

Type I (ISO 14024) labels are based on a set of criteria defined by private or public environmental

labeling programs and are issued and controlled by third-party certifiers. The awarding body may be either a governmental organization or a private noncommercial entity. Examples of such types of labels include the EU Ecolabel, the German Blue Angel, and the Program for the Endorsement of Forest Certification (PEFC). Type I labels are often referred to as ecolabels, despite the fact that they may coexist with other ecolabeling schemes (like the FSC program for example).

Type II labels are self-declared claims adopted by manufacturers, retailers, or anyone else likely to benefit from such claims. These claims can be factual (e.g., indicating the percentage of a product that is composed of recycled materials) or more unsubstantiated (e.g., simply “eco-friendly”).

Type III labels consist of quantitative product information based on life cycle impacts. In general, these types of labels consist of reports that are presented in a form that facilitates comparison between products, often using a common set of parameters. However, relative product performance based on these parameters is not provided in the label itself, as there are no comparative statements made with respect to other products. An example of Type III claims is the “carbon footprint.”

Despite their widespread use, relatively few regulations exist regarding the use of ecolabels on national levels. From a legal point of view, environmental claims are primarily seen as advertisements and less as market-based policy tools. As such, they are mainly regulated through antifraud and truth-in-advertising laws. However, the World Trade Organization (WTO) monitors the practice of ecolabeling in order to ensure that it doesn’t lead to discrimination between trading partners or between domestically produced goods and services and imports. In the case of disputed environmental claims, the WTO relies on rules such as the General Agreement on Tariffs and Trade (GATT) or the Agreement on Technical Barriers to Trade (TBT). More specifically, the WTO issued a Code of Good Practice for Preparation, Adoption, and Application of Standards to TBT Agreement in order to facilitate ecolabeling and standard setting in a manner that does not conflict with agreed upon

international trade frameworks. Often, it is very difficult to establish objective and scientifically defensible criteria that can justify or dispute claims of environmental superiority.

In cases where the voluntary involvement of an industry in an ecolabeling scheme is considered to be too low, governments may decide to oblige firms to undertake measures to reduce environmental impacts, adopt specific technology changes, or instate mandatory (national) labeling standards. For instance, in order to promote health, food, and environmental safety, governments may impose warning labels (e.g., for flammable materials, potentially dangerous household chemicals, alcoholic beverages, cigarettes). In these cases, producers who do not comply with the national standards cannot obtain access to markets within the country. For example, since 2010, EU Timber regulations prohibit the selling of illegally harvested timber and products derived from illegally harvested timber on the EU market. Trade barriers are not only enforced by laws, however, and can also result from consumer boycotts. Nevertheless, labeling requirements and practices must respect all the WTO regulations, including the provision of the TBT Agreement and the GATT. In other words, ecolabeling schemes should not discriminate either between trading countries and firms or between domestic and foreign products and services. There are several examples of debates regarding the legitimacy of ecolabels, due to the trade distortions arising from the enforcement of their environmental criteria. First, as stated by the OECD (2002) *“The experience of Colombian exporters of cut flowers to Germany provides an example of the impact which a powerful domestic NGO-driven voluntary eco-label can have on a developing country’s trade prospects (OECD 2002).”* This conflict originated with German NGO’s that campaigned against labor and environmental conditions in the flower export industry of developing countries. The Flower Label Program was created in 1996 and disputed by Colombian flower exporters who claimed that it created discrimination and unfair competition.

Another case in which pressure by local consumers and the media effectively excluded foreign

products from the domestic marketplace and created a conflict between international trade and domestic environmental law was the US/Mexico Tuna embargo dispute. In the 1990s, the US banned tuna imports that could not attest to being caught in a dolphin-safe way. The dolphin-safe label has legal status in the United States under the Dolphin Protection Consumer Information Act but was contested by Mexico as creating discrimination and illegal trade barriers (Dyck and Zingales 2002).

Principles of Ecolabeling

The rationale behind ecolabeling is to overcome the market failures that arise in marketing eco-friendly products and services. Markets for environmentally friendly products and services are dysfunctional because most of the promised environmental attributes are both public (nonrival and nonexclusive) and unobservable (credence attributes).

Neoclassical economic theory predicts that consumers will attempt to freeride by enjoying a public good (environmental quality) without incurring the costs associated with its provision. This means that if ecolabeled products are more expensive, there should be no market for them. However, in the last decades, consumers have increasingly considered not only their own well-being but also the environmental and social impacts of their purchasing decisions (Cohen and Viscusi 2012). Motivations for adopting socially appropriate behavior can be explained by preexisting personal norms or by emerging social norms. Nevertheless, even if it has largely been shown that consumers possess other-regarding preferences and are willing to pay for environmental attributes, another requisite for firms to invest in environmentally friendly production technologies and to offer less polluting products and services is the opportunity to provide credible information to users and consumers.

In order to determine the value that consumers attribute to an ecolabel, it is important to distinguish consumer understanding of the ecolabel, consumer confidence in the ecolabel, and their

willingness to pay for environmental amenities (Delmas et al. 2013).

In the literature, it has been shown that simple eco seal type labels are insufficiently convincing. Adding additional information about specific criteria that are defined and verified by a certification scheme can significantly increase willingness to pay (O'Brien and Teisl 2004). Ecolabels have also been found to perform better if the certifier is a familiar organization and/or endorsed by an independent third party (D'Souza et al. 2007), without ignoring that the success of the ecolabeling scheme is also determined by the individual characteristics of the purchaser (Teisl et al. 2008). Moreover, consumers are willing to pay higher premiums for ecolabels where the benefits of environmental improvement are also linked with private benefits. For instance, organic farmers strive to make their food products more appealing by stressing not only the reduced environmental impact of using less pesticides but also by emphasizing the accompanying reduced health risks.

Under the assumption that ecolabeling schemes have the ability to correct informational asymmetry within the producer-consumer relationship, firms are generally motivated to adopt ecolabels in order to differentiate their products. As pointed out by Porter (1991), (environmental) competitive advantages are obtained either by lower costs or by attribute differentiation. In many cases, new technologies with lower environmental impacts are not associated with lower costs; thus, a firm's primary interest in using ecolabels is founded in a search for strategic environmental competition. In developing a strategic advantage based on environmental quality attributes, firms can either focus on their internal organizational process or on their product line. Globally, differentiation will be based on creating a positive image, pursuing positive recognition by stakeholders, or the anticipation that more stringent environmental standards may be implemented in the future.

Voluntary label adoption as a differentiation strategy will affect competition, as it enables products to be perceived as imperfect substitutes (Bonroy and Constantatos 2015). In addition,

ecolabel adoption only partly corrects market failures relating to environmental degradation (Ibanez and Grolleau 2008). The incentives for firms to turn to new environmentally friendly standards and to adopt ecolabeling depend on several factors that can be classified into three general categories: factors relating to market structure and governance, willingness to internalize environmental externalities, and the ability of ecolabels to accurately communicate the certification criteria.

Firstly, there are factors related to the market structure and to the involvement, number, and type of policy makers. Firms' incentives to ecolabel are highly dependent on the costs of investing in green technologies, which concern both fixed (sunk) and relative unit costs, and on the ability of ecolabeling to release price competition. The inherent aim of the investment decisions made by firms is to consider how to best adapt their technology to standard requirements and do not necessarily aim at socially optimal outcomes in terms of environmental improvements (Amacher et al. 2004). Ecolabeling programs differ crucially from standard requirements in their ultimate goal. If an NGO initiates an ecolabeling program, it may likely set standards that maximize environmental improvements, whereas an industry-initiated ecolabeling program may instead be oriented by the pursuit of maximizing industry profits. Consequently, one would expect ENGO's to set more stringent standards than industry-sponsored ecolabeling schemes. It should be noted that more stringent environmental requirements do not necessarily lead to better global outcomes in terms of environmental degradation due to restricted use and consumption (Grolleau et al. 2009). Certifying entities may also compete to capture a firm's willing to adopt an ecolabel, strategically choosing their qualifying criteria in order to maintain their market share of participating firms. The coexistence of both NGO and industry ecolabels should theoretically favor industry profits, since environmental improvement is only feasible if the reduction in "high-environmental quality" labels doesn't outweigh the increase of overall ecolabel users (Fischer and Lyon 2014). Governmental involvement also

plays a role in firms' willingness to adopt ecolabels and thus the overall effectiveness of ecolabels as an environmental policy strategy. In some cases, a market-based ecolabeling mechanism may be reinforced by complementary policy instruments in order to internalize externalities more efficiently (Nunes and Riyanto 2005). These complementary regulatory policies can either be mandatory (e.g., minimum environmental quality standards) or voluntary (e.g., subsidies for green investment).

A second determinant that explains the emergence of ecolabels concerns consumer willingness to pay for the reduced environmental impacts of products and services. The inclusion of social and environmental considerations in consumer decision-making processes has become more common in the last decades, and globally it has been shown that there exists a willingness to pay for social and environmental attributes. However, disagreements remain regarding the price premiums of these attributes (McCluskey and Loureiro 2003). This variation in estimated willingness to pay (WTP) observed in the literature can be attributed to the elicitation methodology used (based on either stated or revealed preferences) but can also, and is most often, structurally based. Indeed, ecolabeling may suit certain products better than others. As an illustration, if consumers value ecolabels for conspicuous reasons, it is important that the ecoseal and the associated product are seen by others. If warm-glow motivations are an important driver of behavior, then consumption of goods with symbolic logos (e.g., dolphin-safe fish, panda bear logo of WWF) may benefit from higher premiums. Thus, the design of an ecolabeling scheme and the associated communication strategy should consider the specific motivations underlying the demand for social and environmental attributes.

Lastly, the salience and efficiency of ecolabeling schemes rely on factors related to the ability of ecolabels to provide information on environmental attributes perfectly and symmetrically. Mason (2011) argues that incentives for firms to apply for ecolabels depend on the ability of third-party certification schemes to accurately distinguish high-environmental quality producers from low-environmental quality ones. Imperfect

label attribution arises either because monitoring is random or because the auditor is incapable of accurately identifying compliance with the required standards at a reasonable cost. This “noise” in testing procedures may even encourage less environmentally sensitive firms to apply for certification, which lowers consumer expectations of environmental quality in all similarly labeled products. Moreover, trust in an ecolabel is reinforced when the label is provided by independent and well-recognized certifying agencies (D’Souza et al. 2007) and if the entity has no financial interest in providing the certification (Harbaugh et al. 2011). In addition, the degree of consumer uncertainty regarding ecolabel claims rises along with increased market proliferation of these types of claims, as consumers will have more difficulty interpreting the absence or presence of an ecolabel, and consequently willingness to pay will decrease.

Efficiency of Ecolabels: Positive and Negative Side Effects

From a theoretical point of view, the introduction of ecolabels should be welfare enhancing as it provides information on the environmental attributes of products or services for which consumers are willing to pay, enabling simultaneous environmental improvements (Ibanez and Grolleau 2008). However, market-based policies and information disclosure, even if they are well known to minimize market distortions, may also result in unintended side effects. Indeed, the misuse of environmental standards, strategic manipulation that creates trade distortions, the excessive use of claims in general, and behavioral biases are some factors that might prevent an ecolabel from reaching its initial objective to reduce or even eliminate environmental externalities.

The motivations for introducing ecolabels and the process of defining and choosing standards are not necessarily the same among environmental organizations, governments, and industrial-oriented entities. While environmentalists focus on criteria that lead to the greatest environmental improvements, industries aim to set standards on

product categories and/or criteria that involve the lowest costs and to take advantage of positive spillover effects such as an improved reputation among consumers and/or an increase in sales of other product categories (Dosi and Moretto 2001). Then, in some cases, the consequence of ecolabel introduction may in fact turn out to be globally harmful for the environment. In these cases, the environmental gains achieved are inferior to accompanying increases in product consumption and resulting pollution.

However, nonprofit labeling schemes do not always eliminate inefficiencies either. Potential examples include technological inertia in the form of reluctance to accept innovative technology developments, imposing excessively stringent standards or costly monitoring processes, and failing to gain sufficient recognition by consumers.

Moreover, the choice of labeling criteria can potentially be used for protectionist purposes. For example, industrialists may select environmental considerations that both meet environmentalists’ concerns and disadvantage rivals. An intuitive example can be found in the environmental impacts of transportation. Domestic firms may overemphasize the role of certain transportation means (e.g., air transport vs. rail transport) in order to discredit foreign rivals. Differences in environmental issues among countries may also be employed in order to disadvantage foreign producers. Furthermore, for variety of reasons – e.g., political support, ideological protectionism – domestic governments may be more sensitive to arguments emanating from domestic producers and environmental activists, regardless of their scientific validity.

Harmonizing ecolabel standards on an international level seems to be a difficult task. Institutional, technological, and environmental heterogeneity between manufacturers is likely to favor strategies that seek to raise rivals’ costs. In other cases, entities may focus on local or regional environmental priorities that lack international relevance and/or create trade barriers.

In general, developing countries are more vulnerable to the discriminatory impacts that can arise with ecolabeling schemes. First, it is more

difficult for these countries to bear the costs of the certification procedures required for the compliance, testing, and verification of the labeling criteria. Second, developing countries often lack information on requirements and do not have access to the necessary skills and technologies that would allow them to conform to domestic standards.

The proliferation of ecolabels also raises important questions. Theoretically, providing information on products and services should enable consumers to update their beliefs, thus having a positive impact on environmentally friendly consumption decisions. However, the proliferation of ecolabels today has, in some cases, led to an excess of information that users must process, contributing to increasing consumer confusion, indifference, and even skepticism towards ecolabeling schemes.

The overuse of environmental claims by firms may also be viewed as greenwashing by consumers. Consequently, there seems to be a need for more efficient auditing of the various stages of the ecolabeling process (Lyon and Maxwell 2011). But such regulatory intervention is costly, often imperfect, and not necessarily legally binding. Today, no binding legal framework at the international level exists in order to ensure the reliability of green claims, i.e., in order to verify whether the relevant labeling criteria are respected as defined by ISO 14021. To fill this gap and to call for a legal framework, private initiatives have denounced firms that use “misleading” and/or inaccurate advertising regarding their actions in favor of the environment. In 2008, several environmental NGO’s decided to attribute an award, called the Pinocchio prize, to firms who promoted themselves to be greener than they actually are (<http://www.prix-pinocchio.org/>).

In order for ecolabels to be effective, consumers must receive, process, and believe the information transmitted by labels. Therefore, the type and the content of information transmitted by ecolabels are of crucial importance. The aim of an efficient label design is to provide consumers with information that increases understanding of environmental targets and improvements in a credible manner and may simultaneously boost

pro-environmentally friendly preferences themselves (Delmas et al. (2013)).

Truthful ecolabels do not necessarily achieve environmental policy goals. Because consumers tend to focus only on the emphasized attribute, they often underestimate the overall environmental harm resulting from their consumption and use of the product. For instance, by buying fluorescent light bulbs, consumers may focus on the energy saving properties yet underestimate the environmental damages that result from the emission of mercury vapor occurring when these lights are disposed. A great many other behavioral biases may also distort the efficiency of ecolabeling schemes (Grolleau et al. 2015). For example, people in general overestimate their own virtuous behavior (optimism bias) and underestimate the degree to which they are implicated in contributing to negative externalities (attribution error). If these cognitive biases apply to ecolabeled products, people may allow themselves to increase consumption of environmentally friendly products thinking that their consumption is less harmful, while in reality their increased consumption may in fact contribute to a net deterioration of the environment (Bougherara et al. 2005).

An important determinant in consumer behavior is other-regarding and social preferences. And one can imagine the progressive shift of markets towards more sustainable production and ecolabeling may have the potential to increase consumer preferences for environmental-friendly attributes in the long run. In other words, exposure to an increasing and diverse range of ecolabels can trigger new social norms pertaining to acceptable environmental behavior.

However, even if exposure to green products may activate norms of social responsibility and ethical conduct, it may also give rise to a rebound effect (Cohen and Viscusi 2012). Recent research has shown that green consumption or pro-environmental conduct can modify people’s overall sense of morality and affect subsequent behavior either within or outside of the concerned domain (Clot et al. 2014). This moral compensation, also called the “licensing effect,” shows that doing something virtuous seems to give individuals license to later engage in more immoral

actions. The consequences of this phenomenon can be diverse and have negative impacts on the global environment through the reduction of pro-environmental behavior in subsequent decisions or in other domains.

Conclusion

In theory, ecolabeling is an efficient way of regulating environmental externalities, under the condition that the additional costs don't exceed consumer willingness to pay for environmental quality. When this is the case, voluntary approaches for environmental regulation are good alternatives to command and control policies. However, in real life, ecolabeling does not always achieve its initial goal of environmental improvements due to a variety of side effects. Intervention or regulation could then provide a way to correct these inefficiencies (Horne 2010).

But what should be the role of governments? Is new legislation the only way to regulate inefficiencies, or can efficient voluntary mechanisms be implemented to effectively encourage optimal behavior? It is important to be able to measure the global effectiveness of ecolabeling programs through assessments on an international level. The importance of environmental improvements in relation to trade distortion raises ethical and equity issues and calls for better insights on the economic implications of the interaction between ecolabeling schemes and (already existing) command and control policies.

Regarding the extent to which ecolabeling should be regulated, a distinction should be made between interventions in the certification process and scheme, and interventions regarding the use of ecolabels as a communication tool. A first task is thus determining which products and services should be targeted to use ecolabels. One might wonder whether consumers should systematically be encouraged to purchase ecolabeled products and services and firms pushed to adopt ecolabels and more sustainable production technologies. Complementary policies on both the national and international level should also target ecolabels. Nevertheless, the simple

provision of accurate product information is often not enough to induce behavior change, and thus social and behavioral research could help policymakers to better design ecolabels in order to improve their overall effectiveness (Grolleau et al. 2015). One promising new strategy emphasizes the importance of educating the youngest generations in environmental issues and the relevance of ecolabels to these issues.

References

- Amacher G, Koskela E, Ollikainen M (2004) Environmental quality competition and eco-labeling. *J Environ Econ Manag* 47:284–306
- Bonroy O, Constantatos C (2015) On the economics of labels: how their introduction affects the functioning of markets and the welfare of all participants. *Am J Agric Econ*, forthcoming 97(1):239–259
- Bougherara D, Grolleau G, Thiébaud L (2005) Can labeling policies do more harm than good? An analysis applied to environmental labelling schemes. *Eur J Law Econ* 19:5–16
- Clot S, Grolleau G, Ibanez L (2014) Do good deeds make bad people? *Eur J Law Econ*, forthcoming <https://doi.org/10.1007/510657-014-9441-4>
- Cohen M, Viscusi WK (2012) The role of information disclosure in climate mitigation policy. *Climate Change Econ* 3(4):1–21
- D'Souza C, Taghian M, Lamb P, Peretiatko R (2007) Green decisions: demographics and consumer understanding of environmental labels. *Int J Consum Stud* 31:371–376
- Delmas M, Naim-Birch N, Balzarova M (2013) Choosing the right eco-label for your product. *MIT Sloan Manag Rev* 54(4):10–12
- Dosi C, Moretto M (2001) Is ecolabelling a reliable environmental policy measure? *Environ Resour Econ* 18(1):113–127
- Dyck A, Zingales L (2002) The corporate governance role of the media. Chap. 7. In: *The right to tell. The role of mass media in economic development*. World Bank Institute, Washington, DC, pp 107–140
- Fischer C, Lyon T (2014) Competing environmental labels. *J Econ Manag Strateg* (forthcoming) 23(3):692–716
- Grolleau G, Ibanez L, Mzoughi N (2007) Industrialists hand in hand with environmentalists: how ecolabeling schemes can help firms to raise rivals' costs. *Eur J Law Econ* 24(3):215–236
- Grolleau G, Ibanez L, Mzoughi N (2009) Too much of a good thing? Why altruism can harm the environment. *Ecol Econ* 68(7):2145–2149
- Grolleau G, Ibanez L, Mzoughi N, Teisl M (2015) Helping eco-labels to fulfill their promises, Climate policy <https://doi.org/10.1080/14693062.2015.1033675>

- Harbaugh R, Maxwell J, Roussillon B (2011) Label confusion: the Groucho effect of uncertain standards. *Manag Sci* 57(9):1512–1527
- Horne R (2010) Limits to labels: the role of eco-labels in the assessment of product sustainability and routes to sustainable consumption. *Int J Consum Stud* 33:175–182
- Ibanez L, Grolleau G (2008) Can ecolabeling schemes preserve the environment? *Environ Resour Econ* 40(2):233–249
- Lathrop K, Centner T (1998) Eco-labeling and ISO 14000: an analysis of US regulatory systems and issues concerning adoption of Type II standards. *Environ Manage* 22(2):163–172
- Lyon T, Maxwell J (2011) Greenwash: corporate environmental disclosure under threat of audit. *J Econ Manag Strateg* 20(1):3–41
- Mason C (2011) Eco-labeling and market equilibria with noisy certification tests. *Environ Resour Econ* 48:537–560
- McCluskey J, Loureiro M (2003) Consumer preferences and willingness to pay for food labeling: a discussion of empirical studies. *J Food Distribution Res* 34(3):95–102
- Nunes P, Riyanto Y (2005) Information as a regulatory instrument to price biodiversity benefits: certification and ecolabeling policy practices. *Biodivers Conserv* 14(8):2009–2027
- O'Brien K, Teisl M (2004) Eco-information and its effect on consumer values for environmentally certified forest products. *J For Econ* 10:75–96
- Porter M (1991) Towards a dynamic theory of strategy. *Strateg Manag J* 12:95–117
- Teisl M, Rubin J, Noblet C (2008) Non-dirty dancing? Interactions between ecolabels and consumers. *J Econ Psychol* 29:140–159
- Thøgersen J, Haugaard P, Olesen A (2010) Understanding consumer responses to ecolabels. *Eur J Mark* 44(11/12):1787–1810

Economic Analysis of Brazilian Labor Law

Sergio Pinheiro Firpo and Luciana Yeung
Insper Institute of Education and Research,
São Paulo, Brazil

Definition

Being a civil law country, Brazil heavily relies on codified rules to regulate labor relations. For this purpose, back in the 1940s,

a comprehensive set of laws was consolidated in the *Consolidação das Leis Trabalhista* (CLT), which remains the main source of labor regulation in the country, even after some minor reforms in the end of the year 2017 (to be discussed ahead). Besides the CLT, the Federal Constitution itself has several clauses governing labor and employment relations. Finally, more and more judicial precedents have been taken and followed (although not in a mandatory manner) as sources of law.

Introduction

Labor law in Brazil is based on a paternalistic, overly protective, and dictatorial view of employees. The first and perhaps most important principle is that of the vulnerability of employees: these are considered almost legally incapacitated, with no knowledge whatsoever of what their desires are. This is the first and sharp contrast with the economic concept of rational agents: people do know what their preferences are and what costs they must incur in order to maximize their benefits.

Another assumption by the Brazilian labor law is that of adversarial relations between employers and employees. On one hand, employees are considered to be irrational, not knowing what their preferences are, and totally lacking bargaining power; on the other hand, employers are considered to be pure profit maximizers, willing to adopt inhumane conditions for his/her employees with the sole objective to reduce production costs. Because of that, legislators, judges, and law enforcers will stay vigilant to regulate this relationship. Under such circumstances, the scenario predicted by Coase, of cooperative bargaining, leading to efficient allocation of resources, becomes even more a piece of fiction. Labor relations, by definition and assumption, are considered to be of very high transaction costs. Then, as Coase's theorem clearly predicts, the law will directly impact on the efficient allocation of resources. The question is as follows: Has this impact been mostly positive or mostly negative in Brazil? The answer is not difficult to find. We provide in the second half of this

entry a brief review of the empirical literature showing the negative impacts on efficiency of these regulations.

Main Characteristics of Labor Law and Labor Justice in Brazil

Conflicts are pervasive in judicial relations in Brazil. The judiciary power is comprised of five big branches: Federal justice (for all matters related to the Federal Constitution), common state justice (for most civil cases, such as commercial, tort, family, consumer rights, etc.), military justice (for trials involving the military), electoral justice (running the electoral system), and labor justice. Within it, a three-tier structure exists, comprising local first-instance courts, second-degree regional courts, and a supreme labor court. In cases where labor conflicts involve constitutional matters, cases and appeals may be directed to Federal courts, even the highest Federal Supreme Court (STF).

Given such a structure, it is not a surprise that this is a highly conflictive branch of law. Official statistics published by the National Council of Justice (CNJ) – the “Justice in Numbers” (*Justiça em Números*) reports – testify this: in 2016, there were 1721 new cases in labor courts for every 100,000 people, or a total of 4.3 million new cases. With regard to pending cases, labor courts comprised 5.4 million, placed top 3 (6.8% of total), behind only to state courts and Federal courts. As a specialized branch of the judiciary, labor courts have far more cases than those of the electoral justice, which in 2016 received only 972,000 new cases.

Furthermore, “Justice in Numbers” also shows that the most frequent theme in all Brazilian courts, the top 1 subject of judicial conflicts, is “labor dismissal/dismissal fees,” comprising 11.5% of all cases brought to courts in Brazil. In 2016, they represented more than 5.8 million cases. Top 2 conflictive subject was “breach of contracts” (not including labor contracts), with far less, 1.9 million cases.

One can affirm, with certainty, that labor laws are not bringing harmony in employers-

employee’s relations, or in Coasean terms, the law is not reducing transaction costs.

Historical Background: The Creation of Labor Laws and Labor Justice in Brazil

Labor law and labor justice were born under a dictatorial regime in Brazil. The first compilation of labor laws, the *Consolidação das Leis Trabalhistas* (CLT), was made in 1943, 2 years after the birth of labor courts in the country. Both are creations of populist dictator, Getúlio Vargas. In his turn, Vargas came to power with the 1930 revolution, which put an end in the “old republic” in Brazil, which had been dominated by the rural coffee aristocracy. Vargas’ regime represented the beginning of the political power exerted by urban industrialists and bourgeoisie. His government was one of the extreme instabilities, pointed by several attempts of coup and a few effective ones, ending in a dramatic way, with the suicide of Vargas in 1954. He took power twice, in a total of 18.5 years (1930–1945 and, then, 1951–1954). It was during his first government, under a dictatorship, that labor law was created. It is usually considered to be a populist dictatorship, commonly observed in several Latin American countries at that historic time.

Those were years of intense social transformations in Brazil: industrialism, urbanization, and influx of foreign immigrants. In order to maintain social order, Vargas managed to control the various classes that were emerging, especially urban groups. Being a founding member of the “Brazilian Labor Party” (*Partido Trabalhista Brasileiro*, PTB), he had strong commitment with the labor class. Another important historical event from those years was the emergence of communist ideas, mainly brought by the Italian immigrants. In response to all these movements, the creation of the labor justice and of labor laws was a means to pacify and monitor potential labor movements adversarial to the government. The paternalistic approach adopted by this legal structure was a manner to compensate the

intense monitoring by the government that the new system established.

According to Lopez (1991), the unification of all previous labor laws under the CLT represented a “combination of state paternalism and fascism, which was the essence [of Vargas’ dictatorship]” (free translation). Some characteristics of the Brazilian labor structure reveal the accuracy of this analysis. Vargas created a “fake” unionism, which included monopoly of representation, i.e., only one union per professional category in a particular geographical region. Employees must pay compulsory union fees (deducted from their paychecks once a year). In turn, unionism is funded by public resources. Unsurprisingly, unions were mostly subordinated to the Federal government. By its turn, labor laws, greatly represented by the CLT, impose high financial costs and responsibility burden on employers. Legal rules are generally applied under the rule of strict liability (Yeung 2006). Labor law is based on the principle of vulnerability of the employee, under all circumstances. This means labor conflicts will never be solved as contractual conflicts by Brazilian courts. There is a long list of labor benefits that employers are mandated to provide. This, in turn, increases labor expenses up to 200% of what employees effectively receive as paycheck (Pastore 2007).

Paternalism toward employees is also reflected in procedural rules. Until a very recent reform in labor laws, employees had no costs at all to access labor courts: all that was required was a self-declaration of financial incapacity. This certification was, then, later submitted for approval by the judge. Decisions were made based on the judge’s subjective opinion; there was no requirement for financial statements. This procedural rule was granted by a combination of principles present in Brazilian labor laws, a rule issued in 1983 (Law Number 7115) and several labor jurisprudence. Furthermore, although provided by procedural law, “*bath faith litigation*,” or litigation without reasonable motif, is rarely punished in labor courts. With all this in place, it is not difficult to understand why there is so much conflict in Brazilian Labor Justice.

Distortionary Effects of Labor Law in the Brazilian Labor Market

There are three important features of labor law in Brazil that, in spite of being the source of several distortions in the labor market, they were not affected by the 2017 reform, which we shall discuss in the next section. These three aspects are *firing costs*, *labor taxes*, and the existence of the Brazilian Severance Indemnity Fund, known as *FGTS* (Fundo de Garantia do Tempo de Serviço). These three components have had important effects on the labor market over the past decades affecting employment levels, unemployment and informality rates, and labor productivity.

Firing (or dismissal) costs are the set of rules in labor law that cause the firm-employee separation, whose initiative is by the employer, to involve some pecuniary penalty to the firm, both by means of indemnification to the dismissed worker and by means of fines paid to the government, in addition to prior notice to the worker. Firing costs in Brazil are relatively high, as they involve a fine of 40–50% of a monthly salary for each year in the firm, plus 1 month as prior notice to the worker, who may not need to fully work during that period. Higher dismissal costs induce lower job variability and a lower level of formal employment because firms tend to hire less, since any reversals in product demand could lead to costly layoffs in the future. The effects on informality are the opposite, and higher dismissal costs induce higher levels of informality. Finally, because (i) dismissal costs have a reallocation effect from formal to the informal sector and (ii) the latter has an average productivity below that of the formal sector, the higher the firing costs, the lower the labor productivity. In addition to that effect, there is the productivity-reducing effect in the formal sector, because firms in that sector will typically operate below the optimal labor demand curve.

Labor or payroll taxes are those that are paid when hiring an employee. There are many taxes and charges on payroll in Brazil, and they may vary between 20% and 50% of overall labor costs. These taxes are, in general, proportional to the time of service and to the salary paid. The effect of levying labor taxes is well known and amounts to the

reduction of employment. These taxes shift the labor supply curve inward and upward, and therefore their effects on unemployment and informality are inversely proportional to what they have on employment. Finally, labor taxes reduce the marginal productivity of labor for two reasons. First, after an increase in labor taxes, workers who will remain employed are those in the points along the supply curve that are associated with lower reservation wages. Because reservation wages reflect productivity, average labor productivity will decrease after an increase in labor taxes. The second reason is related to the increase in informal sector induced by these taxes, which tends to operate under lesser conditions of productivity.

The last important feature of the Brazilian labor law that has not been affected by the recent reforms is FGTS, the Brazilian Severance Indemnity Fund, which was created in 1966. From the 1940s until 1966, Brazilian labor law stipulated that in the case of unjustified dismissals, firms should pay as compensation to the worker one monthly salary for each year worked at the firm. In addition, the law guaranteed job stability for workers who had an employment spell longer than 10 years with the same firm. FGTS was introduced with the aim to introduce flexibility in the job relations. Since 1966, workers have not had job stability. But they have kept access to the one monthly salary for each year worked at the firm. That amount is deposited by the firm into an individual FGTS account. Workers also are eligible to a 40% fine over the balance in the case of unjustified dismissals that is paid by the firm at the time of dismissal. Thus, FGTS is not a simple labor payroll tax: it is a mandatory savings account paid by the employer plus a fine that is appropriated by the worker. That account can only be accessed if the worker is fired. Voluntary dismissals do not allow withdraws from the FGTS account. Therefore, FGTS induces the opportunistic behavior of workers who always benefit from being fired: they may have granted access to their FGTS account and to the fine. Because the interest rates accruing FGTS accounts are lower than inflation, it is rational for workers to try to have access to their individual accounts. The overall effects of FGTS are excessive labor turnover, which reduces labor productivity. Since it increases

labor costs for the firms, it also has all those other effects already described for labor taxes: lower employment, higher unemployment, and informality rates.

Labor Law and Labor Courts After the 2017 Reform: An Economic Analysis

In November 2017, a so-called labor law reform took place in Brazil, 74 years after the creation of the CLT, which remained basically untouched during all this time. In fact, it was approved by the National Congress in July, under Law N. 13,467, in a very heated process, severely opposed by trade unions and labor movements. The law in its integrity has close to 900 articles. For the purposes of this paper, we highlight three very important ones. We then provide a brief economic analysis of these topics (Yeung 2017).

1. (Articles 611-A and 611-B) Individual and/or collective agreements gained much importance as compared to legally issued laws over the discussion of a wide variety of themes. For a civil law country, highly dependent on legislative rules, this was a significant change in the process of creation of formal rules for labor relations.

Economic analysis: This is one of the most controversial points of the 2017 labor reform. However, it is very much in line with an economic analysis of labor, especially under the concept of rational agents and of the Coase theorem: employees know what is good for them and, when bargaining conditions are guaranteed and transaction costs kept in low levels, the result of the cooperative bargaining between employees and employers will be efficient. This is especially true if negotiation occurs with the support of labor committees or labor representatives at the workplace (as predicted by Article 510-A). Furthermore, although Article 611-A predicts 15 matters over which cooperative agreements and private bargaining can stipulate, Article 611-B lists 30 matters (i.e., double the previous number) which there must be *no* private negotiation, over

which only the law may regulate. These are mostly laws related to situations in which negative externalities are created by the employer to the employees and to the society.

2. (Articles 545, 578, and 579) Employees are no more legally required to make contributions to trade unions: the rule that mandated each employee to make an annual contribution equivalent to the wage of 1 day of labor, stipulated by Getúlio Vargas, is finally banned. Now, in order to get the contribution, unions must have written consent by the employee (but the rule of monopolist trade unions was unchanged).

Economic analysis: The former rule mandating union contribution by each formal employee was a clear violation of the freedom of association, especially in the case of Brazil, where there exist monopolist trade unions. That law could only have been created under a dictatorial political regime. Economic analysis would go beyond in the normative recommendations and demand for the extinction of union monopoly.

3. (Art. 791-A) Sums of surrender are owed by the employee, in cases of judicial defeat. Curiously, this rule has normally been applied to all other civil cases, even those with legally vulnerable parties, such as consumers against enterprises. The only exception happened in labor courts, and now, with the 2017 reform, this exceptionality has also gone away.

Economic analysis: If employees have “nothing to lose” when accessing labor courts, certainly they will access. This is one main reason of the statistics with evidence of over litigation in the labor justice, as seen before. Another reason might be a potential bias pro-employee and anti-employer by labor courts. The combination of these two factors leads to the very adversarial relationships between the two parties. With the end of the gratuity to access labor courts, it is expected that there will also be a reduction in the numbers of labor litigation.

Few months after its real implementation, some anecdotal evidence show the 2017 labor

reform has already produced concrete impacts on courts, by significantly diminishing the number of labor cases filed. We believe it is still early to significantly draw such conclusion. Hopefully, future academic research will address whether this small step was able to significantly improve the environment of labor relations in the country.

Cross-References

- [Economic Analysis of Labor Law](#)

References

- Lopez LR (1991) História do Brasil Contemporâneo, 6th edn. Mercado Aberto, Porto Alegre
- Pastore J (2007) Trabalhar custa caro. Editora LTr, São Paulo
- Yeung L (2006) The need of modernizing Brazilian labor institutions. In: 14th World Congress – Industrial Relations Research Association, 2006, Lima, Peru. Social Actors, Work Organization and New Technologies in the 21st Century. Lima, Peru: Universidad de Lima – Fondo Editorial, vol 1, pp 117–133
- Yeung L (2017) Análise Econômica do Direito do Trabalho e da Reforma Trabalhista (Lei N° 13.467/2017). REI-Revista de Estudos Institucionais 3(2):891–921

Economic Analysis of Judicial Decisions

- [Empirical Analysis of Judicial Decisions](#)

Economic Analysis of Labor Law

Luciana Yeung

Insper Institute of Education and Research, São Paulo, Brazil

Abstract

The literature on labor relations has been very much divided, mainly between the classical economic approach and the legal/sociological

one. An economic analysis of labor law would benefit by abridging both perspectives and complementing with more. This entry provides brief comments of why whether a purely economic approach on labor or a purely legal approach based on capital vs. labor dichotomy is not sufficient to properly address the subject. Several topics related to labor regulation are discussed, and empirical references on those issues are provided.

Economic Analysis of Labor Law and Labor Relations

Since the beginning of humankind, labor was exercised for people's needs. The study of human labor, its relation with the environment and its impacts in human relations within groups, is possibly one of the oldest social studies. Since the Greek philosophers and the Medieval religious scholars, up to the modern and contemporary economists, social scientists, psychologists, industrial engineers – just to name a few – several brilliant minds have engaged their time and research to understand the phenomenon of human, either in a positive or a normative manner. Many of the classical titles in economics and sociology were inquiries on topics related to the nature and consequences of human labor, such as Karl Marx's *The Capital*, Max Weber's *The Protestant Ethic and the Spirit of Capitalism*, and, in some sense, Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations*.

However, in recent times, the study of labor has been absolutely divided. On one side, economics – and specifically labor economics – has been mainly concerned about players' benefits. The economic approach is concerned with the analysis of workers' jobs, income and benefits (short term and long term), pension, etc. For the employers' side, economic models have studied impacts on productivity, profits, flexibility, etc. This perspective sees workers and employers as if the only things these actors pursue in the labor arena are economic and material benefits and that the main relationship here is of sellers and buyers. This is the basic framework of labor economics:

a market in which demand (workers) and supply (employers) interact.

On the other side, there is a wide range of scholars analyzing labor relations (until very recently also known as industrial relations) in a very different perspective. This group includes labor sociologists, labor lawyers, and labor historians – among others, who do not view employers and workers as merely in a seller vs. buyer relationship; instead, these researchers see a naturally antagonistic relationship at the labor *arena*. The model of capital vs. labor is the axis of this analysis, and all outcomes derive from it. Especially in the case of labor law, it aims at deriving rules that would smoothen this conflict (whenever possible) or that would balance the opposing forces, normally by protecting the weakest part, i.e., workers. These scholars consider the *locus* of labor as a battlefield, in which conflictive relations pervade, as if confronting with each other was the only thing that matters to these actors in their daily encounters at the workplace.

Needless to say, both approaches are limited and insufficient.

An economic analysis of labor law should take both sides of the view and complement with more. It incorporates features of the demand-supply model of labor and especially recognizes that workers and employers equally face several kinds of *incentives* and *constraints*, which affect their decision-making. However, it also gives significant importance to the formulation, application, and enforcement of rules, whether contractual or regulative ones, in the labor arena. It recognizes that the relationship between employers and workers is not a mere one of seller vs. buyer, or supplier vs. enterprise, and that labor is not a simple commodity. Human relations and power relations matter much here.

Although economic analysis of law a priori adopts the main economic framework, which considers labor *locus* a market, it recognizes that this is a special one, in which failures are the norm: asymmetric information (either from the employer to the worker or from the worker to the employer), externalities, uneven bargaining and political powers, monopoly, and monopsony,

among others. Externalities are sources of high transaction costs, and as the normative approach of the Coase Theorem tells us, under these circumstances, legal rules have an important role in the determination of the levels of efficiency (Coase 1960; Cooter and Ulen 2007). In other words, in this market, institutions matter and matter a lot.

Special Rules for a Special Market

Market failures exist in all markets. Yet, in the context of labor, they are essential features, not exceptions.

Asymmetric Information

Whenever labor is done for others – i.e., not for one's own subsistence – it entails a contractual relation, either a formal or an informal one. It is a contract because it is a *promise* made by someone to deliver something in exchange of another thing, and this is an enforceable promise. Workers promise to work under such and such conditions, to deliver certain tasks and/or to obtain certain results, in exchange of salary and a certain set of benefits. Employers, in their turn, promise to provide certain working conditions in exchange of workers' labor, which is used to produce goods and services, which, in turn, generate receipts and profits. The presence or absence of a written paper does not alter the fact that this relationship has a contractual nature, because both formal and informal contracts have their valid mechanisms of enforcement; it does not mean that written contracts are less effective in the guarantee of productive outcomes. The comparison between formal and informal types of contract is not the main issue of this brief entry, and, for our purposes, one should only bear in mind that labor entails, indeed, a contractual relation.

Besides that, most of the transactions between employers and employees are non-instantaneous, long-term interactions. The longer the relationship, the stronger are the impacts of uncertainties. For instances, neither the employer nor the employee knows, at the time of the recruitment, whether the economy will be booming or will be slowing down

in the following years (which may impact in the employer's ability to pay higher salaries); both parties do not even know how long will the company survive in the market.

Finally, human interactions are characterized by imperfect and unbalanced information: one is never able to access the whole set of information related to the other party, due to limited cognitive capability and due to high transaction costs. This is true even for noneconomic human relationships such as marriage, friendship, etc. In the case of employers, even if they try hard to access candidates' true ability (even by means of sophisticated processes of screening), they will not be able to fully acknowledge the candidate's adequacy for the position. Employers will also not be able to access the candidate's real interests in joining their firm ("Does this candidate plan to stay here long? Or will she/he leave the company as soon as another opportunity appears in the rival company?"). On the employee's side, she/he, at the time of the recruitment, is not able to fully access the challenges of the new job (even if one seeks information with current employees); she/he might not be able to understand the easiness or difficulty to interact with the new boss. Not even is the candidate able to access vital information about the company's finances and organizational challenges.

Externalities and Interdependence of Utility Functions

Employers depend on workers to reach some of their desired outcomes; workers, on their side, depend on employers to reach their economic benefits. This is what economist calls an *interdependence* of their utility functions. Whenever it happens, negative externalities might be created: one's decisions or actions might inflict undesired costs to the other party.

However, positive externalities might also be created. In the labor context, whenever an employer decides to train or educate his/her workers, he/she creates benefits not only for the company, for the trained workers, but potentially to the whole society, since a highly skilled labor force generates technology and knowledge that may benefit the whole society, directly or

indirectly. This is true both for general and for firm-specific skills. The same happens when employers invest in workers' health and safety conditions and so on. Thus, giving incentives for employers to invest in his/her employees' well-being also generates social efficiency. Yet, one knows from economic models that externalities – both negative and positive – drive the economy away from the point of efficiency: in the presence of *negative* externalities, private parties tend to provide *excessive* amounts of products and services; on the other hand, *positive* externalities lead to *insufficient* amounts. In both cases, regulation is necessary to guarantee efficiency: for negative externalities, taxes and quotas must be stipulated; for positive externalities, public policies must provide subsidies or other incentives to stimulate their production.

Uneven Bargaining and Political Powers

Economists are unable to deal with the concept of power relations. This is a phenomenon that cannot be explained by economic rules nor economic models. Yet, they significantly affect bargaining outcomes, even in the context of Coasean bargaining, with impacts on efficiency (e.g., Galiani et al. 2014; Barnes 2004, among others). It is also unrealistically naïve to consider this market as having “normal” supply and demand forces, with equal bargaining power. For most systems, employer and employees – especially if these ones are treated individually – do *not* have equal stands in the negotiation of working conditions. Thus, to assume that this is an ordinary contract relationship, in which clauses are outcomes of cooperative and voluntary agreement, does not truly mirror reality. Sociological, cultural, and political variables may explain why workers' bargaining powers are, on average, stronger in one system than they are in others, but the common-law tradition which views labor regulation as another ordinary type of contract regulation might not be well suited for other countries; probably, there are places in which power imbalance between workers and employers is larger. Then, other types of labor rules must be considered, so to take the “power effect” into consideration.

Monopoly-Monopsony

The historical answer to uneven bargaining and political powers in labor relations was labor organizations, mostly (but not exclusively) trade unions. Yet, this creates a problem of another sort: unions operate, most of the time, as monopolies in the labor market. As one knows, monopolies create deadweight loss: due to the existence of unions, wages tend to be higher, and as a response, employers are willing to hire less labor. In addition to that, because membership is never mandatory, the presence of unions creates an even worse outcome, dual labor markets, in which unionized workers have higher benefits but at the cost of nonunionized ones. In poor countries, this is materialized, in its extreme, into the case of formal vs. informal labor markets, with a large portion of the labor force belonging to the second one. Needless to say, working and living conditions here are strikingly adverse. We will discuss some empirical findings about trade unions in a specific section ahead.

It is also not unusual to find cases of monopsony in labor markets: one or a few firms are the sole employers of a certain type of workers in a region. Under these circumstances, employers have abnormal market power to unilaterally set wages, working conditions, etc., to which workers have little chances to react. A monopsony may be as damaging for social welfare as monopolies.

Litwinski (2001) discusses both problems in the labor market and, within the American context, defends the application of antitrust law to unions (as it has happened in the USA since the beginning of the twentieth century, in the *Loewe vs. Lawlor* case in 1908) and of labor and antitrust laws to firms, whenever they operate as labor monopsonies. Further, he claims that “anti-trust law and labor law are somewhat like Siamese twins unhappily separated at birth. The natural affinity between their subject matter and concern was recognized by the early cases and statutes” (p. 50). According to the author, in the same manner that labor unions should be prohibited to exercise their monopoly power, firms should also be disciplined in their monopsony power.

Labor Laws for the Labor Market

Summing up, labor relations are based on contractual, (usually) long-term relationships, characterized by high levels of uncertainties and asymmetric information. Different to what happens in some other markets, bargaining power here is inherently uneven between suppliers (i.e., workers) and demanders (i.e., employers), and monopolies and monopsonies occur frequently. Besides that, workers' and employers' utility functions are interdependent. All this shows that market failures are abundant in this case. As economic theory tells us, regulation is needed to solve these failures. A third party – usually the government – must step in; otherwise, efficiency will not be achieved, and maximum welfare will be missed.

However, bad regulation might be worse than no regulation, and many times, the problem lays here.

The Dilemma of Labor Law in an Economic Perspective

Summing the previous discussion, one may consider the *locus* of labor as a market, but in which (legal) rules are of particular importance to equilibrate powers and interests. In this manner, both economists' and sociologists'/lawyers' perspective on labor can be equally considered, in a balanced manner. The main goal of labor law in modern democracies should be to provide sound institutions that will equilibrate information, bargaining power, and contractual relationships while fostering economic growth by promoting firms' efficiency boosting. This task should involve the *creation* of rules by the executive and legislative powers and also their *enforcement* by the judiciary. Besides that, if rules to promote social indicators (such as education, health, pension, etc.) are also developed, a country should be able to truly achieve economic development.

Yet, it seems that many countries have failed in this task, either by pending too much to one side – overregulation or “bad” regulation of the labor market, hindering potential economic growth (e.g., Latin American countries, Spain, Italy,

etc.) – or by pending to the other side: lack of regulation leading to low welfare, working conditions, or inequality (e.g., Asia's developing countries and, to some degree, the USA, as compared to other industrialized countries). In some sense, this reflects a one-sided view of labor, either excessively economic or excessively based on the balance of powers.

Special Topics

In this section, we discuss the main types of regulation in some areas of labor relations. We refer to some empirical studies that try to access the impacts of these regulations in the economy.

Labor Regulation (In General)

Botero et al. (2004) analyze regulation of labor, specifically laws on employment, unionization, and collective bargaining, and social security in a sample of 85 countries. Their main result is that countries in which labor regulation is overall more invasive, have higher unemployment rates – especially of the young.

In a theoretical study, Blanchard and Giavazzi (2003) show that labor regulation creates sharp intertemporal trade-offs for workers. In the short run, workers are better off with more regulation, because their wages will be higher; yet, in the long run, regulation brings higher unemployment. In a dynamic approach – in which there are new entrant firms in the future – their model predicts that employed workers will be worse off with deregulation, both because wages will be lower and the probability of them being unemployed will be higher. On the other hand, workers who would have been unemployed without the entrance of new firms benefit either by the new possibilities of being hired or with an increase in their wages. Although this study does not include empirical observation, it seems to be a more careful analysis of the implications of regulation and deregulation in the labor market, compared to those done initially in the classical economic literature.

In the other side of the discussion – the legalistic, noneconomic literature – there have been

claims that labor regulation should be (also) approached in a transnational manner, specifically under regional integration pacts (Trubeck et al. 2000). Although the idea seems coherent, these studies would greatly benefit from a more analytical and empirical approach, which could evaluate the concrete outcomes on the overall economy, and specifically on the variables directly affecting workers and employers.

Unions

Basic economic models show that unions impact markets by raising wages of unionized workers. Consequently, employment levels decrease, since employers try to substitute those more expensive workers with nonunionized, cheaper ones. In a context where there is no perfect mobility of factors (in this case, labor is the main factor), this creates a long-lasting, perverse effect in the economy: the presence of a dual labor market (Piore 1969; Doring and Piore 1971), in which the first one is marked by the presence of unionized, high-skilled workers, earning high wages, but in which the level of employment is lower. On the other hand, low-skilled workers are trapped in the market with lower wages and, usually, worse labor conditions. This happens because due to the lack of skills, inferior workers cannot move to the first market; on the contrary, firms tend to have higher mobility and may choose to operate in either the superior or the inferior market, depending on the presence of unions and their bargaining power. If it is the case that firms are not able to choose in which market to operate, in the long run, those facing unions will lose their comparative advantage and lose business to those who do not face unions. The result will be a decline in unionization in the overall economy (Posner 2003). In countries where unionization does not happen at the firm level, the analogy holds true for different sectors of the economy: those who face unionization will lose competitiveness and may perish in the long run; the opposite is true for those sectors not facing unionization.

For some decades, empirical literature has systematically brought evidence in this direction (e.g., Borjas 2016). Yet, one study shows these effects in a careful and detailed manner and deserves a more careful attention. Aidt and

Tzannatos (2008), besides linking union activities to monetary policies in an original but important fashion, show that it is not the simple, cross-country variation in union density that affects economic outcomes (as the literature has traditionally implied). The authors bring the *coordination* and *bargaining coverages* to the spotlight. They empirically associate the existence of coordinated bargaining systems – via labor unions or other formal and informal labor organizations – to better macroeconomic outcomes (e.g., lower unemployment) and more flexible labor markets, what may sound odd at a first glimpse. On the contrary, high bargaining coverages are associated to poorer economic performance. According to these authors, bargaining coordination guarantees positive outcomes and also enables labor markets to respond more adequately and in a less adverse manner to external shocks. In an effort to employ the classical economic theory of labor, coupled with an institutional perspective, these authors conclude that “it is the total ‘package’ of (formal and informal) institutions that matters for economic performance” and that “labour market coordination cannot and should not be thought of in isolation from the broader institutional environment” (p. 290).

Minimum Wage

Economists usually regard the effects of mandated minimum wages as similar to those of unions, because unions’ most usual demand is on higher wages. As the models predict, minimum wages create unemployment, especially for marginal workers, i.e., female, young and old, and black (Posner 2003). The effect is explained by a change in relative prices between less- and high-skilled workers, the first becoming more expensive (because minimum wages are set lower than the level of high-skilled workers’ productivity, not affecting, thus, their wages). If minimum wage is applicable only to some sectors and not to others, the effect is again equal to that when there is a dual labor market, of unionized vs. nonunionized firms: it may increase unemployment in the sector covered by minimum wage and decrease unemployment in the sectors in which no mandatory wage prevails. Finally, as Posner (2003) explains, minimum

wage may hinder on-the-job training of marginal workers (those who need it the most), because employers may not compensate the costs of the training investment by paying lower wages.

This topic has brought much attention to the economic literature, and Alan Krueger calls for a “History of Economic Thought on the Minimum Wage” (2015). Traditionally, evidence brought by this literature has corroborated the classical microeconomic model, as explained above. Yet, recently some studies have questioned that explanation, by showing that the effects of minimum wages on unemployment have been overstated. Hoffman (2016), for instance, finds no impacts of increases in minimum wages on young and low-skilled workers’ unemployment. In fact, the author finds some *positive* effects of wage increases on employment for a few states in the USA. The study also presents some other recent literature corroborating its findings.

Undoubtedly, although a “history of economic thought” already exists for this theme, more empirical studies, which carefully analyze the broad implications of minimum wages on economic variables, are still needed. The story is not over.

Unemployment Compensation

Traditional economic models predict that longer and more generous unemployment compensation leads to lower rates of employment and higher wages. The explanation is that, besides having less incentives to find a job, workers are comfortable to calmly search for better jobs, i.e., their outside options increase. Yet, empirical evidence coming from the literature is not strikingly convincing in this direction. By analyzing changes in the rules of unemployment compensation in the USA during the 2008–2009 recession, Nicholson et al. (2014) find that “[unemployment coverage] extensions may indeed have had detectable effects on the U.S. labor market, but some of the studies are contradictory” (p. 212). Ludsteck and Seth (2014) for Germany in the late 1990s and beginning of 2000s, and Howell and Rehm (2009) for the OECD countries in a period of 30 years, also find contradictory empirical evidence.

This is certainly a subject which deserves more in-depth empirical and analytic studies.

Outsourcing and Subcontracting

The current phenomenon of labor outsourcing takes place under two variations: cross-border outsourcing (transnational subcontracting) and within-border outsourcing (local subcontracting). The first one is a relatively recent trend, which intensified in the last three to four decades, in which companies from industrialized countries, mainly the USA and Western Europe, move their fabric plants abroad, to countries in which labor is cheaper. They may also simply stop their own national production and purchase products from other firms and factories in those poorer locations. The second type of outsourcing happens nationally and is a much older phenomenon: a firm, with its own employees, specializes in some activity – either manufacturing, assembling, or servicing – and hires other firms and/or outside workers to do activities in which it is not specialized. For instance, an automobile firm does not produce all the automotive parts, a bank may hire another firm to provide specialized security services, a movie theater may subcontract people and firms to offer popcorns, and a school may hire a company to be responsible for the cleaning services. Subcontracting, in this sense, materializes Adam Smith’s idea of division of labor which occurs in a society.

As the technology advances, the economy gets more specialized, and outsourcing – either transnational or local – becomes more pervasive. Along with this process, trade unions and other advocates of labor’s interests (labor sociologists, labor attorneys, public prosecutors, etc.) have loudly manifested their opposition to the increasing trends in labor outsourcing. They argue that it has been responsible for the degradation of labor conditions, including the decrease in wages and the increase of unemployment: *inside* workers lose their positions, and lower-paid jobs are created for *outside* workers (located in the same country or abroad). Once again, employers and workers seem to be in opposing sides, headed to divergent directions. What does evidence show?

Actually, very controversial results. In the topic of international outsourcing, literature is almost split: those who argue that it is beneficial for workers and others showing that it is very

deleterious to workers – for those in rich and in poor countries. Bachmann and Braun (2008) show part of the controversy in the literature. However, after analyzing a big panel of data, their conclusions are positive toward outsourcing, at least for some workers in Germany. They show that German companies cross borders to hire labor in cheaper countries, workers face higher job stability, and the effect is more significant for those in the service sector and less in the manufacturing sector. Yet, the authors remark that effects are strongly heterogeneous across the skill levels and age. Specifically, international outsourcing is significantly hazardous for medium-skilled and lower workers in the German manufacturing sector.

Anner (2011) evaluates the impact of outsourcing on the other side of the story, i.e., in countries of *destination*, and finds evidence of very negative results. In Central America – specifically Nicaragua and El Salvador – outsourcing reduced workers' bargaining power (due to spatial dispersion) and led to lower wages.

With regard to domestic outsourcing, or subcontracting, Autor (2003) shows that unionized firms in the USA tend to outsource *more* than nonunionized firms. This is a counterintuitive result and offers hints that unions are not always able to guarantee rules that benefit labor. Yet, the author is cautious about welfare impacts. In Brazil, Stein et al. (2015) use official data on eight million workers and employ the methodology of “fixed effects panel data,” i.e., they compare wage effects for a particular person when he/she migrates from being an inside worker to an outsourced one and vice versa. The main goal is to control for each worker's personal (observable and nonobservable) characteristics. Their results show that, after controlling for fixed effects, wage differential between inside and outsourced workers is of 3% only. Besides that, significant wage differentials are created when outsourcing occurs for low-skilled workers under such situations, wages are 12% lower compared to inside workers with the same characteristics.

Such controversial evidence in the literature shows the need for more detailed and careful analysis on this topic.

Other Labor Issues

Posner (2003) provides brief discussions on a long list of other labor issues, such as child and female labor, health and safety at the workplace, maternity leave, employment discrimination (race, sex, age, disability), pension law, etc. However, more empirical analysis on each of these themes is still needed, and many other themes are also still left uncovered by the literature. It is also important to directly link the explanation coming from economic models to institutional evaluations: What are the effective impacts of specific labor rules? (For instance, “What are the effects of creating quotas – based on gender, race, age, etc. – at a workplace?”) What are the political, sociological, etc., concerns behind the stipulation of some labor laws? (For example, “Why should the employer be always strictly liable for damages in case of accident at the workplace? Is this the manner to minimize damage costs and probabilities of accidents? If not, what kind of rule would be more effective?”)

Our discussion above offered some initial hints. But much should be assessed empirically.

Summary

We started this entry by arguing that the literature on labor relations has been very much divided, mainly between the classical economic approach and the legal/sociological one. An economic analysis of labor law would benefit by abridging both perspectives and complementing with more. It seems that the model of labor market is a plausible one. However, it is crucial to remember that, more than other usual markets, there are strong failures in place here. There are also problems of unbalanced bargaining and political powers. Because of this, labor regulation cannot mirror simple, ordinary contract regulations. A special look is needed.

Empirical literature, unfortunately, mirror the dichotomy of the economic vs. legal approach and, for several (if not all) issues, provide contradictory, inconclusive evidence. Many times, problems might have happened due to imperfect data or inadequate methodologies. As information technology advances, as well as the quality of

databases, one might expect higher quality and, hopefully, more conclusive evidence.

No matter what, although this topic is one of the oldest themes in the social sciences, there is still room for much research.

Cross-References

► Incomplete Contracts

References

- Aidt TS, Tzannatos Z (2008) Trade unions, collective bargaining and macroeconomic performance: a review. *Ind Relat J* 39(4):258–295
- Anner M (2011) Unionization and wages: evidence from the apparel export sector in Central America. *Ind Labor Relat Rev* 64(2):305–322
- Autor DH (2003) Outsourcing at will: the contribution of unjust dismissal doctrine to the growth of employment outsourcing. *J Labor Econ* 21(1):1–42
- Bachmann R, Braun S (2008) The impact of international outsourcing on labour market dynamics in Germany: the impact of international outsourcing on labour market dynamics in Germany. *Ruhr Econ Pap* 58(1): 1–29
- Barnes P (2004) The auditor's going concern decision and types I and II errors: the Coase theorem, transaction costs, bargaining power and attempts to mislead. *J Account Public Policy* 23(6):415–440
- Blanchard O, Giavazzi F (2003) Macroeconomic effects of regulation and deregulation in goods and labor markets. *Q J Econ* 118(3):879–907
- Borjas GJ (2016) *Labor economics*, 7th edn. McGraw Hill Education, New York
- Botero JC, Djankov S, Porta RL, Lopez-De-Silanes F, Shleifer A (2004) The regulation of labor. *Q J Econ* 119(4):1339–1382
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cooter R, Ulen T (2007) *Law and economics*, 5th edn. Addison Wesley, Boston
- Doring PB, Piore MJ (1971) *Internal labor markets and manpower analysis*. Heath, Lexington
- Galiani S, Torrens G, Yanguas ML (2014) The political Coase theorem: experimental evidence. *J Econ Behav Organ* 103:17–38
- Hoffman SD (2016) Are the effects of minimum wage increases always small? A reanalysis of Sabia, Burkhauser, and Hansen. *Ind Labor Relat Rev* 69(2):295–311
- Howell DR, Rehm M (2009) Unemployment compensation and high European unemployment: a reassessment with new benefit indicators. *Oxf Rev Econ Policy* 25(1):60–93
- Krueger AB (2015) The history of economic thought on the minimum wage. *Ind Relat* 54(4):533–537
- Litwinski JA (2001) Regulation of labor market monopoly. *Berkeley J Employ Labor Law* 22:49–98
- Ludsteck J, Seth S (2014) Comment on 'unemployment compensation and wages: evidence from the German Hartz reforms' by Stefan Arent and Wolfgang Nagl. *Jahrb Natl Stat* 234(5):635–644
- Nicholson W, Needels K, Hock H (2014) Unemployment compensation during the great recession: theory and evidence. *Natl Tax J* 67(1):187–218
- Piore MJ (1969) On-the-job training in the dual labor market. In: Arnold R, Weber FC, Woodrow LG (eds) *Public-private manpower policies*. Industrial Relations Research Association, University of Wisconsin., 1969, Madison, pp 101–132
- Posner RA (2003) *Economic analysis of law*, 6th edn. Aspen Publishers, New York
- Stein G, Zylberstajn E, Zylberstajn H (2015) Diferencial de Salários da Mão de Obra Terceirizada no Brasil. Working Paper 400, São Paulo School of Economics, FGV, CMICRO – No. 32, Working Paper Series (Ago 7). Available at: <http://bibliotecadigital.fgv.br/dspace/handle/10438/13883>
- Trubeck DM, Mosher J, Rothstein JS (2000) Transnationalism in the regulation of labor relations: international regimes and transnational advocacy networks. *Law Soc Inq* 25:1187–1211

Further Reading

- Aubuchon C, Bandyopadhyay S, Bhaumik SK (2012) The extent and impact of outsourcing: evidence from Germany. *Fed Reserve Bank St Louis Rev* 94(4):287–304
- Dickens W, Franco F, Landier A, Nicoletti G, Spector D, Solow R, ... Alesina A (2003) Macroeconomic effects of regulation and. *Europe* 879–908
- Laporšek S (2013) Minimum wage effects on youth employment in the European Union. *Appl Econ Lett* 20(14):1288–1292

Economic Analysis of Law

Alain Marciano

MRE and University of Montpellier, Montpellier, France

Faculté d'Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France

Abstract

The purpose of this entry is not to identify the central claims upon which rests an “economic analysis of law.” That goes far beyond what

could be done. Our goal is to characterize methodologically an economic analysis of law and, as a consequence, to establish and explain the distinction that exists between an “economic analysis of law” and “law and economics.”

Definition

Economic analysis of law is the field defined by the use of economics to analyze legal phenomena and the functioning of the legal system.

Introduction

The purpose of this entry is not to identify the central claims upon which rests an “economic analysis of law.” That goes far beyond what could be done. Our goal is to characterize methodologically an economic analysis of law and, as a consequence, to establish and explain the distinction that exists between an “economic analysis of law” and “law and economics.”

Usually, those terms are used interchangeably to describe any economic work dealing with law or legal rules. For instance, “The Problem of Social Cost” (Coase 1960) that represents the “origin [of ...] the modern *law and economics* movement” (Hovenkamp 1990, p. 494; emphasis added) and marks the passage from an “old” to a “new” law and economics (Posner 1975) is also viewed as the article that “established the paradigm style for the *economic analysis of law*” (Manne 1993). Symmetrically, an *Economic Analysis of Law* (Posner 1973) was viewed as “coursebook in law-and-economics” (Krier 1974, p. 1697). One could also quote Cento Veljanovski (1980, p. 160) and even Richard Posner, one of the founders of an economic analysis of law, who characterized his work as one of the “recent developments in *law and economics*” (1975, emphasis added). Many additional references could be cited that would confirm that the two expressions are usually viewed as synonymous.

Yet, sometimes, they are distinguished. In this regard, Ronald Coase is probably one of the most significant authors to quote. He explained that “two parts” coexist in law and economics (1996, p. 103; or Coase in Epstein et al. 1997, p. 1138), which are “quite separate although there is a considerable overlap” (Coase 1996, p. 103). The first part corresponds to what is called law and economics and implicitly corresponds to the analyses to which Coase attached his name. The second part, to which “Judge Posner is the person who has made the greatest contribution” (Coase in Epstein et al. 1997, p. 1138), is “often called the economic analysis of law” (Coase 1996, p. 103).

This is the distinction we want to emphasize in this text. Our point is that a better understanding of an “economic analysis of law” requires careful understanding of the differences with “law and economics” and, therefore, a careful understanding of what is law and economics.

A Negative Characterization of an Economic Analysis of Law: Law and Economics

In a “law and economics” approach, the focus is put on the economy, the economic system, or economic activities, and, since economic activities take place in an institutional, legal environment, a correct understanding of the economy and of economic problems requires to take into account how and how far legal rules do affect the economy. This is precisely what law and economics is. This is exactly what Coase did in “The Problem of Social Cost (1960).” He “used the concept of transaction costs to demonstrate the way in which the legal system could affect the working of the economic system” (Coase 1988, p. 35). Later, he added: “[F]or me, ‘The Problem of Social Cost’ was an essay in economics. It was aimed at economists. What I wanted to do was to improve our analysis of the working of the economic system” (1993, p. 250). From this perspective, Coase was one of the founders of “law and economics” in its modern form but not of an economic analysis of law.

What is important to correctly understand the distinction between law and economics and economic analysis of law is that a law and economics approach rests on a definition of economics by scope, object, domain, or subject matter. In other words, what distinguishes economics from other social sciences is that each of these sciences has its own subject matter. Once again, this was the perspective explicitly adopted by Coase, for whom “economists do have a subject matter” (1998, p. 93). It corresponds to “certain kinds of activities” (1978, p. 206) or, more broadly, “the working of the economic system, a system in which we earn and spend our incomes” (1998, p. 93). Or, in a slightly different way, economists study “the working of the social institutions which bind together the economic system: firms, markets for goods and services, labour markets, capital markets, the banking system, international trade, and so on” (1978, pp. 206–207). In other words, economists study the activities that take place on explicit markets. That’s the only set of activities that they can analyze with their tools. Coase again: economists “should use these analytical tools to study the economic system” (1998, p. 73).

This view has two implications. The first one is that economists should not study what is outside of the scope of their discipline, in particular legal rules, legal phenomena, and legal cases. They do not fall into the subject matter of economics and are not studied by law and economics. They are important but *only* to give “details of actual business practices (information largely absent in the economics literature)” (Coase 1996, p. 104). Second, law and economics does not only exclude certain objects from its domain of investigation and also excludes noneconomists from the analysis of economic activities. Coase was very clear about that: the subject matter is “the dominant factor producing the cohesive force that makes a group of scholars a recognizable profession” (p. 204), “the normal binding force of a scholarly profession” (p. 206), and what “distinguishes the economics profession” (p. 207). Thus, the delimitation or delimitation of the scope of economics establishes a distinction with other social sciences and guarantees the unity and the autonomy of economics.

What Is an Economic Analysis of Law

By contrast with “law and economics” presented above, an economic analysis of law implies a radical change in the object of study. The focus is no longer put on economic activities – defined as the activities that take place on markets – and the objective is no longer to understand how legal rules influence the economy. The legal system is no longer seen as the environment in which economic activities take place and therefore external to the object of study – the working of the economic system. It becomes the object of study. In fact, and very straightforwardly, an economic analysis of law consists in using economics to analyze the legal system and how it works or, to quote Lewis Kornhauser, “Economic analysis of law applies the tools of micro-economic theory to the analysis of legal rules and institutions” (2011). This means, in particular, that legal rules are no longer taken as given and exogenous. An economic analysis of law endogenizes legal rules. To quote Posner, an economic analysis of law consists in “the application of the theories and empirical methods of economics to the central institutions of the legal system” (1975, p. 39).

From a methodological perspective, an economic analysis of law rests on and requires a specific definition of economics – completely different to the definition of economics used in law and economics; the difference in definition of economics is such that it makes law and economics and an economic analysis of law incompatible. Indeed, an economic analysis of law does not and cannot rest on a definition of economics by subject matter or by scope or by domain, as it is the case with law and economics. Analyzing the working of the legal system is possible and legitimate, only if the very idea that there exists a subject matter specific to economics, to which economists should restrict their attention, is abandoned. Otherwise, there would be no justification to analyze, among other things, the behavior of criminals, judges, prosecutors, or attorneys and any of the phenomena that are usually studied in economic analyses of law. These behaviors and phenomena are not,

strictly speaking, of economic nature because they do not take place on markets. They can be studied by economists only because it is assumed that any kind of activity and of behavior or any phenomenon, even those taking place outside of markets, can be studied by economic theory. This is exactly the view adopted and promoted by Gary Becker who stressed that economic theory “applies to both market and nonmarket decisions” (1971, p. viii) or “the economic approach is clearly not restricted to material goods and wants, nor even to the market sector” (1976, p. 6).

Let us note here that such a change in perspective – the expansion of the domain of economics beyond its “traditional” boundaries – is a consequence of the assumption that no difference exists between market and nonmarket behaviors. Individuals are supposed to *always* behave in the same way. As Becker wrote, “human behavior is not compartmentalized, sometimes based on maximizing, sometimes not, sometimes motivated by stable preferences, sometimes by volatile ones, sometimes resulting in an optimal accumulation of information, sometimes not” (1976, p. 14).

This then means that all social sciences have exactly the same subject matter. All social sciences can study the same behaviors and the same phenomena. The only difference that exists between them is the method or the approach they use (Becker 1971). Then, from such a perspective, economics is defined or rather described or characterized – Posner (1987, p. 1) argued that economics cannot be defined – by its method. Economics is a “way of looking at human behavior” (Becker 1993, italics added). To use Posner’s words, economics is “a powerful tool” (1973, p. 3) or “an open-ended set of concepts” (Posner 1987, p. 2), which can be used to analyze any kind of human or social phenomenon – including legal ones. Then, “when used in sufficient density these concepts make a work of scholarship ‘economic’ regardless of its subject matter or its author’s degree” (Posner 1975). It is only if this definition of economics is adopted that an economic analysis of law is possible.

A Few Historical Landmarks

Cesare Beccaria and Jeremy Bentham – sometimes Gladstone (see Posner 1976) – are viewed as the “predecessors” (Stigler 1984, p. 303) of an economic analysis of nonmarket behavior (among others: Posner 1975, 1993, p. 213) and, more specifically, of economic analyses of crime and punishment. Indeed, they were the firsts to analyze illegal behaviors and illegitimate activities as the result of an “economic calculus” (Becker 1968, p. 209) or of a “rational choice” (Posner 1993, p. 213).

In the twentieth century, the first who developed an economic analysis of law is Guido Calabresi in the early 1960s. In “Some Thoughts on Risk Distribution and the Law of Torts (1961; see also Calabresi 1965)” by contrast with Coase, Calabresi used economics to analyze a legal problem – namely, the compensation of victims of car accidents in different systems of liability. This was acknowledged by Walter Blum and Harry Kalven (1967, p. 240), Posner (1970, p. 638) and Frank Michelman (1971, p. 648). However, Calabresi also insisted that his work should not be viewed as a form of economic analysis of law. We suggested elsewhere that his analysis should be viewed as a form of heterodox economic analysis of law, mainly for two reasons. First, he rejected the behavioral assumption that economic analyses of law use – individual rationality. Second, he eventually criticized standard – read, Posnerian – economic analyses of law because it assumes that the “world” – the conditions in which individuals act and live – is given and by analyzing how individuals chose in a set of given conditions. To him, the law could be used to change the world and not to promote its economic efficiency (Kalman 2014). In other words, Calabresi claimed that economics, and economic analyses of law, should not be only about the allocation of resources. And, one could add, Calabresi did not make the methodological move of explicitly defining economics as a method.

It was Becker who did this move, because and when he was the first economist who consistently and repeatedly used economics to analyze

nonmarket behaviors and to explicitly define economics as an “approach.” Among his writings, “Crime and Punishment: An Economic Approach” (1968) must be singled out as the first (modern) economic analysis of a legal problem, namely, crime and illegal behaviors. Indeed, Becker was the first to explain crime as the result of a rational decision, of an economic calculus, that is, of the comparison of costs and benefits. As mentioned above, this comes from the idea that all human behaviors are of the same nature and can be explained as if they were rational, an assumption that remains crucial to an economic analysis of law.

However, Becker’s direct contributions to an economic analysis of law were scarce. The ones who really founded an economic analysis of law are William Landes, Isaac Ehrlich, and Richard Posner.

Landes and Ehrlich were Becker’s PhD students (Fleury 2015, 2016). They transformed Becker’s insights into a specific field of research. Ehrlich studied the participation in illegitimate activities and deterrence (1967, 1970), and Landes analyzed the effects of fair employment legislation on the well-being of discriminated nonwhites (see Fleury 2014). Also of particular importance, Landes was the first who developed an economic analysis of courts (1971). Landes’s work was important because it was the first to really propose a model of the working of the judicial system, taking into account the two sides of the legal “market.” Thus, while Becker had introduced the assumption that criminals are rational, Landes introduced the assumption that prosecutors also are rational.

Landes played also an important role for having involved Posner in a program in law and economics launched by the National Bureau of Economic Research (Landes 1998). Posner became one of the most important figures in economic analyses of the law. Not only he invented the expression, in the title of his 1973 book, and launched the first journal devoted to an economic analysis of law – namely, the *Journal of Legal Studies* – but he also contributed to explicitly define the field, providing its methodological bases and incessantly opening up new domains

of analysis. He is one of the most important – quantitatively and qualitatively – contributors to the field. Posner’s contribution cannot be described or summarized, and, accordingly, it could be said that it is particularly difficult to summarize an “economic analysis of law.” However, let us note that Posner generalized the use of the assumption that individuals are rational in economic analyses of law; in particular, he developed and expanded the analyses of judicial decision making – explaining that judges make their decisions by maximizing a utility function. Actually, Posner linked this claim to another claim about the efficiency of the law, at the same time a positive and normative claim. Rationality and efficiency are two central claims, maybe the two pillars upon which rest economic analyses of law.

Conclusion

We suggest to distinguish between “law and economics” and an “economic analysis of law.” An economic analysis of law is based on a definition of economics as a method, a (set of) tool(s) that can be used to analyze the functioning of the legal system. The distinction was invented in the early 1970s but remains central to understand most of the analyses made at the intersection of economics and the law.

Cross-References

- [Austrian Perspectives in Law and Economics](#)
- [Austrian School of Economics](#)
- [Becker, Gary S.](#)
- [Coase, Ronald](#)
- [Coase Theorem](#)
- [Consequentialism](#)
- [Law and Economics](#)
- [Law and Economics, History of](#)
- [Posner, Richard](#)
- [Rationality](#)

Acknowledgments Special thanks to Magdalena Malecka for her comments on a previous version.

References

- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Becker GS (1971) *Economic theory*. Alfred Knopf, New York
- Becker GS (1976) *The economic approach to human behaviour*. The Chicago University Press, Chicago
- Becker GS (1993) Nobel lecture: the economic way of looking at behavior. *J Polit Econ* 101(3):385–409
- Blum WJ, Kalven H Jr (1967) The Empty Cabinet of Dr. Calabresi: auto accidents and general deterrence. *Univ Chic Law Rev* 34(2):239–273
- Calabresi G (1961) Some thoughts on risk distribution and the law of torts. *Yale Law J* 70(4):499–553
- Calabresi G (1965) The decision for accidents: an approach to non-fault allocation of costs. *Harv Law Rev* 78(4):713–745
- Coase R (1960) The problem of social cost. *J Law Econ* 1:1–44
- Coase R (1978) Economics and contiguous discipline. *J Leg Stud* 7(2):201–211
- Coase R (1988) The nature of the firm: influence. *J Law Econ Org* 4(1):33–47
- Coase R (1993) Law and economics at Chicago. *J Law Econ* 36(1 Part 2):239–254
- Coase R (1996) Law and economics and A. W. Brian Simpson. *J Leg Stud* 25(1):103–119
- Coase R (1998) The new institutional economics. *Am Econ Rev* 88:72–74
- Ehrlich I (1967) *The supply of illegitimate activities*. Unpublished. University of Chicago
- Ehrlich I (1970) *Participation in illegitimate activities: an economic analysis*. PhD dissertation, Chicago
- Epstein RA, Becker GS, Coase RH, Miller MH, Posner RA (1997) The roundtable discussion. *Univ Chic Law Rev* 64(4):1132–1165
- Fleury J-B (2014) *From opportunity theory to capital punishment: economists fighting crime in the 1960s and 1970s*. Mimeo
- Fleury J-B (2015) *Massive influence with scarce contributions: the rationalizing economist Gary S. Becker, 1930–2014*. *Eur J Law Econ* 39(1):3–9
- Fleury J-B (2016) Gary Becker. In: Marciano A, Ramello G (eds) *Encyclopedia of law and economics*. Springer
- Hovenkamp H (1990) The first great law and economics movement. *Stanford Law Rev* 42(4):993–1058
- Kalman L (2014) Laura some thoughts on Yale and Guido. *Law Contemp Probl* 77(2):15–43. Available at: <http://scholarship.law.duke.edu/lcp/vol77/iss2/3>
- Kornhauser L (2011) *The economic analysis of law*, Stanford Encyclopedia of Philosophy Archive. <http://plato.stanford.edu/archives/spr2014/entries/legal-econanalysis/>
- Krier JE (1974) Review of economic analysis of law by Richard A. Posner. *Univ Pa Law Rev* 122(6):1664–1705
- Landes W (1971) An economic analysis of the courts. *J Law Econ* 14(1):61–107
- Landes W (1998) *The art of law and economics: an autobiographical essay*. *Am Econ* 1(Spring 1997), reprinted in *Passion and Craft, Economists at Work*
- Manne HG (1993) *An intellectual history of the George Mason University School of Law*. <http://www.law.gmu.edu/about/history#manne>
- Michelman F (1971) Pollution as a tort: a non-accidental perspective on Calabresi's costs (review of the costs of accidents: a legal and economic analysis by Guido Calabresi). *Yale Law J* 80(3):647–686
- Posner RA (1970) The costs of accidents: a legal and economic analysis. *Univ Chic Law Rev* 37(3):636–648
- Posner RA (1973) *Economic analysis of law*. Little, Brown and Co, New York
- Posner RA (1975) *The economic approach to law*. Texas Law Rev 53(XX):757–782
- Posner RA (1976) Blackstone and Bentham, The revolution in social thought. *J Law Econ* 19(3), 1776:569–606
- Posner RA (1987) The law and economics movement. *Am Econ Rev* 77(1):1–13
- Posner RA (1993) Gary Becker's contributions to law and economics. *J Leg Stud* 22(2):211–215
- Stigler GJ (1984) Economics: the imperial science? *Scand J Econ* 86(3):301–313
- Veljanovski C (1980) The economic approach to law: a critical introduction. *Br J Law Soc* 7(2):158–193

Economic Development

Silvia Ručinská¹, Ronny Müller¹ and Jannik A. Nauerth²

¹Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

²Faculty of Business and Economics, University of Technology Dresden, Dresden, Germany

Abstract

Economic development describes a process of improving economic and social well-being of people in a specific area. In addition to qualitative and quantitative growth, it includes social and environmental aspects. For this purpose various economic measures with different focus are used to quantify economic development.

Introduction

The term economic development is generally used, but there is no clear definition of it. In

general the economic conditions and the standard of living are associated with it. Especially economists tend to consider economic development in terms of economic growth. Economic growth is the percentage change in the amount of a value from one period to another. In general there is a close relation between economic development and the growth of output in an economy. Economic growth describes just a quantitative change in the level of an economic measure. Beyond this development, it describes the improvement of economic and social conditions as a whole including qualitative as well as quantitative measures. Economic growth and economic development are different categories, but for explaining the economic development is a definition of the economic growth useful.

From the Economic Growth to Economic Development

A commonly used instrument to measure economic growth is the “gross domestic product” (GDP). There are three different approaches to measure it. Considering the GDP in terms of expenditures, it is defined as “the sum of the final uses of goods and services (all uses except intermediate consumption) measured in purchasers’ prices, less the value of imports of goods and services” (OECD a). The other approaches regard it in terms of income and production. Thus, it summarizes the economic power within one country into a number. Furthermore, it contains economic acting of all people within the borders of the country without taking in consideration of nationality. From this, economic growth describes the percentage change in GDP from one period to another. In general, values are reported quarterly, as well as yearly. The “gross national income” (GNI) is quite similar to the GDP, but it considers all citizens of a nation, even if they work in another country.

To refine the GDP, a per capita adjustment is used. It is calculated as the quotient of GDP and the number of residents. However, it is independent of any distributional dimension. Therefore, the GDP per capita does not refer to real

distribution, it is just statistical. In relation to population, one can distinguish between extensive and intensive growth. Extensive growth describes an increase in GDP without an increase in its per capita. Contrary to this, intensive growth describes itself as an increase in the level and per capita. To compare the GDP above countries, different price levels are taken into account with purchasing power parities. This adjustment uses baskets of goods to compare the purchasing power between countries.

Despite the advantages of the GDP, it is criticized frequently for excluding social and environmental issues. It should be noted that natural disasters can have positive impacts on it. Additionally, labor without payments as voluntary work or work at home is not included in the computation. Especially illegal work takes place as an estimation. Detailed information about the critics on GDP and the difficulties of measuring development are provided by Stiglitz et al. (2009).

It is usual to describe the economic growth by the model of economic growth. One approach to model growth was made by Harrod and Domar. They were influenced by “Keynesian” economic theory and developed models based on a connection between capital stock and production volume (Domar 1946), as well as investments and aggregate demand (Harrod 1939). The “neoclassical” approach of modeling growth goes back to Solow (1956). He developed a model laying down a long run equilibrium between investments and depreciation. As soon as the optimal capital stock is reached, the per capita income will not increase any further. Additional growth can only be achieved by technological progress. With respect to the utility maximization of the households, the “Ramsey-Cass-Koopmans Model” was modeled by Ramsey (1928) and expanded by Cass (1965) and Koopmans (1965). Different to the “Solow Model,” it considers the saving rate as endogenous depending on the optimization of infinitely living households. However, the technical progress is taken into account as exogenous. This critic encouraged Romer (1990) to develop a new model with endogenous technological change, depending on preferences of the households. However, the “Ramsey-Cass-Koopmans

Model” is the approach most commonly used in growth theory.

Measures of Economic Development

To describe and measure the economic development of economies on an easy way, the World Bank offers a classification by income every year. The current values (2014) state that a GNI per capita of \$1,045 per day or less describes a low-income economy, and the range from \$1,045 to \$12,746 describes middle-income economies. Countries with a GNI per capita and day over \$12,746 are referred as high-income economies. Besides that, low- and middle-income economies are denoted as developing economies (The World Bank 2015). Another classification is used by the United Nations (UN). To define less developed countries, three conditions are used: low per capita income, human assets, and economic vulnerability (UNCTAD 2015).

Taking into consideration those critiques on the GDP and the insufficiency of quantitative indicators to measure development, it is comprehensible that many alternative measures have been developed. The UN supports the “Human Development Index” (HDI), which includes live expectancy at birth, education, and the “gross national income” per capita to measure the development more adequately. Live expectancy at birth and education are used as more qualitative indicators, even if they are numerical measures. Education is taken into account with years of schooling from educated people and the expected years of schooling of uneducated people. Live expectancy is determined by average life span at birth.

Developed countries are reported at a HDI value over 0.8. Less developed countries are reported with a HDI in a range from 0.5 to 0.8, while some sub-Sahara states reported with a value under 0.5. Besides that, the HDI is highly correlated with the GDP, which is why it is criticized as being an inadequate measure of economic development too.

Regarding the distribution of wealth, the “Gini coefficient” is commonly used. It states a value

between zero and one (other scaling is possible), depending on the division of income, whereby a value of zero declares the complete equality in income and a value of one states the contrary. From that it is used as a measure of equality in a society. The OECD countries are reported with a value of 0.32 in 2012 (OECD b).

One famous critic on economic growth came from Meadows et al. (1972). This survey simulated growth scenarios considering world population, raw materials, and environmental pollution. In this context the term “qualitative growth” took place in the discussion. In contrast to “quantitative growth,” it considers issues as health, education, security, and equal opportunities in addition to the GDP.

As an alternative to the GDP, Bhutan proposed the “gross national happiness” (GNH) to measure economic development. It considers social justice, environmental protection, good governance, as well as cultural values. However, it is difficult to measure those qualitative indicators and make them comparable. As the consequence, the GNH was not successful.

Moreover, critics on the GDP took place in the discussion, so the German “Federal Office of Environment” proposed the “national wealth index” (Diefenbacher et al. 2010). It includes many indicators related to environmental issues, mobility, security, health, income, and consumption.

Independent from these approaches, many measures of economic development are existing, not just considering the growth.

References

- Cass D (1965) Optimum growth in an aggregate model of capital accumulation. *Rev Econ Stud* 32(3): 233–240
- Clark G (2007) *A farewell to alms – a brief history of the world*. Princeton University Press, Princeton
- Diefenbacher H, Zieschank R, Rodenhäuser D (2010) Measuring welfare in Germany – a suggestion for a new welfare index. *Umweltbundesamt, Dessau-Roßlau*
- Domar ED (1946) Capital expansion, rate of growth, and employment. *Econometrica* 14(2):137–147
- Harrod RF (1939) An essay in dynamic theory. *Econ J* 49(139):14–33

- Koopmans TC (1965) On the concept of optimal economic growth. In: *The econometric approach to development planning*. North Holland, Amsterdam
- Malthus TR (2007) *An essay on the principle of population*. Dover, Mineola
- Meadows DH, Meadows DL, Randers J, Behrens WW (1972) *The limits to growth*. Universe Books, New York
- OECD (a) Glossary of statistical terms – gross domestic product. <http://stats.oecd.org/glossary/detail.asp?ID=1163>. 30 June 2015
- OECD (b) Compare your country – income distribution and poverty. <http://www.compareyourcountry.org/inequality?cr=oe&lg=en>. 30 June 2015
- Ramsey FP (1928) A mathematical theory of savings. *Econ J* 38(152):543–559
- Romer PM (1990) Endogenous technological change. *J Polit Econ* 98(5):71–102
- Solow RM (1956) A contribution to the theory of economic growth. *Q J Econ* 70(1):65–94
- Stiglitz J, Sen A, Fitoussi J-P (2009) The measurement of economic performance and social progress revisited: reflections and overview
- The World Bank (2015) Country and lending groups. <http://data.worldbank.org/about/country-and-lending-groups>. 30 June 2015
- UNCTAD (2015) The least developed countries report 2014. United Nations Conference on Trade and Development. http://unctad.org/en/PublicationsLibrary/ldc2014_en.pdf. 30 June 2015

Economic Efficiency

Ram Singh
Department of Economics, Delhi School of
Economics, University of Delhi, Delhi, India

Definition

The term Economic Efficiency refers to the relationship between aggregate benefits and costs to the individuals concerned. Among the widely used efficiency criteria are the Pareto Optimality, the Kaldor-Hicks, the Cost-Benefit, and the Wealth Maximization criterion (Hicks 1939; Jain 2015; Jain and Singh 2002; Kaldor 1939; Sen 1970 and Scitovsky 1941). In this essay, we discuss economic efficiency as a tool for the social choice among the alternative legal rules. Discussion is carried out by using illustrative examples. We show that in several contexts the efficiency

can serve as a useful tool for comparing the legal rules. However, it has serious limitations as well. In several situations, the efficiency criteria can fail to compare legal rules or can lead to contradictory rankings. Moreover, the assumptions underlying some of the efficiency criteria do not hold always. We discuss merits and demerits of various efficiency criteria. It is shown that in the real world, economic efficiency has limitations as a guide for making social choice from among the legal institutions.

Introduction

The aim of the economic efficiency is to maximize the aggregate net benefits for the individuals concerned. It serves a useful tool when the objective is to maximize the net economic gains for the individuals involved. A strand of law and economics literature uses economic efficiency as a yardstick to compare outcomes under different legal rules or institutions. In other words, economic efficiency is used as a basis to compare and rank the legal rules and institutions.

The law and economics researchers use several efficiency criteria. Notable among the frequently used ones are the Pareto optimality, the Kaldor-Hicks, and the Wealth maximization criterion. In several legal contexts, these criteria can serve as a useful tool for comparing relative efficiency of the legal rules. However, on several occasions, these criteria can fail to compare various alternative legal arrangements.

For the ease of illustration, let us start with an example. Suppose there are 50 fishery units located on a water stream. Each unit earns a profit of 15 for its owner. There is plan for a cloth factory to start at upstream. If factory starts, it will generate a net profit of 600 for its owner. However, the factory will discharge chemicals in the water stream. The polluted water is going to be injurious to the fish. In the absence of any corrective measure, assume that the fisheries will suffer a harm of 10 each, that is, a total harm of 500. To keep things simple, we can assume that the factory has no other effects, good or bad, on the fisheries and also for any other third party. Moreover, assume if

a chemical treatment device is installed at the factory, it can completely solve the problem of water pollution. The device will cost 501. Now consider the following legal arrangements.

Rule 1: The factory is allowed to operate but will have to fully compensate the fishery for the loss inflicted on them – “full” compensation means that the profit of fisheries will be restored 15 each, i.e., the profit levels in the absence of the water pollution caused by the factory (see Singh 2004, 2007a). *Rule 2:* The law does not allow the factory to operate at all. Clearly, under this rule profitability of fisheries will remain unaffected. *Rule 3:* The factory is allowed to operate without any obligations to compensate the loss suffered by the fisheries. In this case, the profit enjoyed by the fishery owners will come down from 15 to 5. *Rule 4:* The factory is allowed to operate only with the treatment device installed, and half of the cost of the device is to be borne by the factory owner and the remaining half is to be split equally among the fishery owners.

Suppose the factory operates without the treatment device. Then, even after fully compensating the fisheries by paying them 10 each, the factory owner will be left with a profit of 100. Therefore, Rule 1 is more desirable than Rule 2. A shift from Rule 2 to Rule 1 makes one party better off (owner of the factory) without reducing well being of anyone else. This argument in favor of Rule 1 over 2 is at the heart of the widely used efficiency criterion called “Pareto optimality.”

Pareto Optimality

According to the Pareto optimality criterion, a rule or state of affairs x is Pareto superior to another state y if and only if the following conditions are satisfied: (a) every concerned individual or party involved considers the state x to be at least as good as the state y ; and (b) at least one person strictly prefers x over y . Under such a scenario, every individual will either strictly prefer x over y or will be indifferent between the two options. In our example, clearly Rule 1 is Pareto superior to Rule 2. In fact, if the factory owners were to pay 11 to

each fishery owner, all parties can be made better off compared to their well being under Rule 2.

The argument can easily be extended to the social desirability of legal arrangements. A legal rule x is Pareto superior to rule y if at least one person strictly prefers x over y and everyone else is indifferent between the two. Extending the argument further, alternative x is called Pareto optimum if there is no other legal arrangement which is feasible and is Pareto superior to x . In other words if x is Pareto optimal, then there cannot be an alternative t which is available as a social choice option and is Pareto superior to x .

This point can be seen by relabeling Rules 1 in the above example as alternatives x and Rule 2 as y . Further, label Rule 3 as z and Rule 4 as w . So the set of legal alternatives which are available for social choice can be written as

$$L = \{x, y, z, w\}$$

Now, it is easy to see that legal rule x is Pareto superior to rule y . In a meaningful sense, rule x is socially more desirable than y . Therefore, the value judgment based Pareto criterion is rather appealing. If a feasible state of society say u is Pareto superior to some other state v , then plausibly u is socially more desirable than v .

From the definition, it is clear that Pareto optimality of a rule or an alternative is defined with respect to the given set of alternatives and the set of individual concerned. Any change in either of these sets can alter the optimality status of a rule or a social choice option. For instance, in our factory-fisheries example suppose for some reasons the alternative x is not available anymore. Then the social choice set will be

$$L' = \{y, z, w\}$$

where y , z , and w are the legal alternatives as defined above. Now, y becomes Pareto optimum since neither z nor w is Pareto superior to y .

Moreover, Pareto criterion can easily fail to compare the alternatives available and thereby fail to be a guide to the social choice process. To see this, consider a comparison among Rules 2 and 3, i.e., legal arrangements y and z . It can

be seen that neither y is Pareto superior to z nor the other way round. Similarly, it can be seen that legal rule x and z are not comparable on Pareto criterion. In fact, in any choice situation involving alternatives u and v if one party strictly prefers u over v and another strictly prefers v over u then the two alternatives cannot be compared, regardless of how the rest of individuals in the society feel about these two alternatives. Moreover, if the choice involves only two such alternatives, then both will be Pareto efficient!

This shows a serious shortcoming of the Pareto criterion. In real world, the social choice involves legal changes which lead to net gains to some citizens but impose net cost on the others. Specifically, consider a change in the legal status-quo such that the proposed change entails net gains to most citizens but imposes net cost on a few members of the society. Under such situations, the Pareto criterion cannot be the guide to decide between the proposed change versus the status-quo.

Kaldor Efficiency

What is known as the Kaldor efficiency criterion aims to overcome the above limitations of the Pareto criterion. According the Kaldor criterion, a shift from alternative y to x is better if the gainers can compensate the losers and still be better-off. It is noteworthy that according to this criterion, the gainers are not required to actually compensate the losers – only requirement is that in principle they should be able to do so. On this logic, if we revisit the comparison among the Rules 2 and 3, Rule 3 is better than Rule 2 – since the gainer (factory owner) benefits by 600 out of which he can fully compensate fisheries by paying 500 and still be left with a net gain of 100. It can easily be seen that in the above example, each one of the Rules 1, 3, and 4 is better than Rule 2. In other words, alternatives x , z , and w are Kaldor superior to y . Legal alternatives x and z are equally efficient, and both are better than the 4th rule, i.e., the alternative w . The point is that unlike the Pareto criterion, the Kaldor criterion can compare all of the above legal alternatives. In general, the social

ranking can be produced by a pair wise of comparison of the alternatives.

The Kaldor criterion is the basis of the widely used cost-benefit and wealth maximization efficiency criteria. Going back to our earlier example, a shift from rule y to rule x imposes a cost of 500 on fisheries but entails a benefit of 600 to the factory owner. Since the benefit is greater than the resultant costs, the cost benefit criterion will rank x better than y . By similar logic, rules x , z , and w are superior to y .

Wealth Maximization

The wealth maximization criterion ranks the alternatives according to the levels of total wealth under them. On this count too, compared to legal rule y , under each one of the rules x and z the total wealth for all the parties involved is higher by 100; under w the corresponding figure is 99. Therefore, these rules are superior to the rule y according to the wealth maximization criterion. The rules x and z are all equally good. Moreover, these two alternatives are socially best, since under these rules the level of wealth is maximized.

It can be seen from the above discussion that the Kaldor, the cost-benefit, and the wealth maximization criteria provide a similar ranking of the legal rules. Moreover, the social ranking produced is complete, in that using these criteria we can compare any two legal rules. However, it should be noted that the virtues of these three efficiency criteria hold under the assumption that the wealth level is the sole determinant of individual utility levels (well beings). In the above discussion, our underlying assumption has been that well being of each party involved increases with the monetary gains enjoyed by the party.

The Kaldor criterion also serves as a useful metric when production and consumption choices involve only one good and the individual utility functions are increasing in the amount of good consumed. In that case, the criterion requires allocation resources to maximize production of the good. Once that is done, any distribution of the good among consumers is Kaldor efficient. However, such assumptions are restrictive.

Limitations of Efficiency Criteria

In real world, the consumption basket has numerous goods and services in it. Besides, the utility levels may depend on factors other than the individual consumption levels of goods and services. In several such contexts, the Kaldor criteria and its derivatives can lead to contradictory or inconclusive ranking of legal choices. Moreover, several decision makers may be guided by the non-economic considerations or might not be able to work in line with the objective of economic efficiency (see Singh 2003 and Schäfer and Singh 2018). Therefore, the real world applicability of the efficiency criteria is limited.

On top of it, the efficiency criteria completely ignore the issue of fairness, equity, and justice which can be paramount considerations when it comes to choosing from different legal regimes. Nonetheless, these criteria are widely used in law and economics. On account of the ease of working with it, most of the economic analysis of legal rules and institutions is based on the wealth maximization criterion. For a discussion on the underlying assumptions, the use, and the debates surrounding the use of this criterion see, R. A. Posner (1985). For use of the criterion in specific contexts see Singh (2007a, b) and Schäfer and Singh (2017).

Cross-References

- [Cost–Benefit Analysis](#)
- [Economic Analysis of Law](#)
- [Posner, Richard](#)
- [Strict Liability Versus Negligence](#)
- [Wealth Maximization: Efficiency and Equality Considerations](#)

References

- Hicks JR (1939) The foundations of welfare economics. *Econ J* 49:696–712
- Jain SK (2015) *Economic analysis of liability rules*. Springer, New Delhi, pp 23–33
- Jain SK, Singh R (2002) Efficient liability rules: complete characterization. *J Econ* 75(2):105–124

- Kaldor N (1939) Welfare propositions of economics and interpersonal comparisons of utility. *Econ J* 49: 549–552
- Posner RA (1985) Wealth maximization revisited. *Notre Dame J Law Ethics Public Policy* 2:85–105
- Schäfer H-B, Singh R (2018) Takings of land by self-interested governments economic analysis of eminent domain. *J Law Econ* 61:1–40
- Sen AK (1970) *Collective choice and social welfare*. Holden-Day, Inc., San Francisco, Chapters 2
- Singh R (2003) Efficiency of ‘simple’ liability rules when courts make erroneous estimation of the damage. *Eur J Law Econ* 16:39–58
- Singh R (2004) ‘Full’ compensation criteria: an enquiry into relative merits. *Eur J Law Econ* 18:223–237
- Singh R (2007a) Causation-consistent’ liability, economic efficiency and the law of torts. *Int Rev Law Econ* 27:179–203
- Singh R (2007b) Comparative causation and economic efficiency: when activity levels are constant. *Rev Law Econ* 3:383–406
- Scitovsky T-d (1941) A note on welfare propositions in economics. *Rev Econ Stud* 9:77–88

Economic Growth

Maria Rosaria Alfano

Economics Department, Seconda Università degli Studi di Napoli, Italy

Abstract

Considering its quantitative property, economic growth has often been considered an index of wealth. Nonetheless, it does not represent the well-being of a given country. In fact, it lacks information on how this wealth is redistributed or on the indirect effects of the said production, such as environmental consequences. Economic literature highlights the difference between economic growth and development, attributing to the latter a holistic definition, which takes into account additional factors, such as collective well-being, social equity, life expectancy, quality of institutions, and environmental quality. Although, at a first glance, the boundaries between concepts of growth and development may be considered as clearly distinguishable, they often tend to disappear when analyzing the two variables

from a long-term perspective. In both economic theory and practice, there are several significant variables, such as the quality of human capital and institutions or environmental efficiency, which play a key role in the opportunities for development in most high-income countries. Within this framework, development becomes a fundamental feature of economic growth, when the latter is interpreted as a general improvement in quality of life, as opposed to a mere quantitative increment in production.

Definition

Economic growth is the increased capacity of an economy to produce goods and services, comparing one period of time to another. It is measured through the growth rate of the gross domestic product (or gross national product, or national income), which is calculated in real terms in order to obtain an indicator that is not influenced by inflation.

Economic Growth

Three main approaches are attributed to economic growth theory:

- Classical economists – with authors such as Smith (1776), Malthus (1798), Ricardo (1817), Ramsey (1928), Young (1928), Schumpeter (1934), and Knight (1944).
- Keynesian and neoclassical economists – who can be divided into two categories, characterized by the different role that technology plays in the production function. On the one hand, Solow (1956) and Swan (1956), Cass (1965), Koopmans (1965) regarded technology as a given variable and built models of what is referred to as exogenous growth. On the other hand, Keynesian authors, such as Harrod (1939) and Domar (1946), as well as some neoclassical authors, such as Romer (1986), Lucas (1988), Rebelo (1991), Grossman and Helpman (1991), and Aghion and Howitt

(1992, 1998), directly included technology variables into the production function, implementing models of what is referred to as endogenous growth.

- Modern economists – who are all the economists interested in the immaterial aspect of growth and in the role of institutions in the economy. The most important authors in this category are North (1990), Evans (1995), Coleman (1990), Putnam (1993), Fukuyama (1995), and Sen (1999).

A different classification of economic growth theories is based on the perspective of divergence or convergence of the economies of different nations (de la Fuente 2000). Conventionally, convergence is considered when poorer economies grow at higher rates than richer ones, thus reducing income differences between them. Subsequently, divergence is determined by the opposite mechanism. According to the former, a general equalization of income levels would be observed in the long run, while an increase in the existing gap is expected when considering the latter. Indubitably, the source of such discrepancies lies in the differences in the formulation of the production function and of technological progress dynamics. Specifically, one of the conditions for economic convergence is the presence of decreasing returns to scale, which allows poorer economies, with initially lower levels of capital, to grow faster than the richer ones. Similarly, increasing returns to scale bring about divergence between the economies of different nations. Another element that characterizes the long-run growth path is related to technological progress. Though widely accepted that an increase in technological investment ensures higher future growth rates, it is not clear whether this would ultimately lead to a convergence or divergence effect. In fact, since significant differences in growth rates are not sustainable, if the technological return is a decreasing function of its accumulation, equalization in the level of technical efficiency will be observed. Finally, another factor of convergence is associated to the allocation mechanism that allows focusing investment where productivity and return on capital are greatest. Since investment

in poorer countries is concentrated in the agricultural sector, which is characterized by low labor productivity, the reallocation of resources towards the manufacturing sector allows for a rapid increase in mean productivity. Such a boost is not feasible in rich countries, where investment is already focused in highly productive sectors.

Classical Economists

When individually analyzing the three theoretical approaches to economic growth, one can immediately recognize that, aside from the main basic concepts, the classical literature contains some crucial perspectives that would later be reconsidered in modern theory. In his book *An Inquiry into the Nature and Causes of the Wealth of Nations*, Smith, founder of modern economic science, considers a nation's growth as the result of an enhancement in productive processes, which delivers an increase in wealth and production with the same resources. The latter is made possible by the augmentation of labor specialization, given by an increase in production and knowledge. Thus, the issue of low resources is overcome through enhanced efficiency. Continuous growth in the economy is thereby guaranteed.

Ricardo, on the other hand, expresses a different view on the matter, focusing on the problem of income distribution and its impact on economic development. Motivated by the conviction that development depends on the accumulation of capital, in other words, the share of profits devoted to investment, Ricardo dedicated his work to the study of the determinants of capital accumulation. With reference to an economy composed of landowners, capitalists, and workers, though no technical progress, the author showed how an increase in production could only be achieved through a greater exploitation of the employed labor force and cultivated land. According to his analysis, this process leads to a rise in rents and a continuous reduction in profits (due to the diminishing returns on land), thus discouraging investment and hampering growth. The economist Thomas Malthus, who studied the problem of long-term growth from a demographic perspective, also reached a similar conclusion. In this case, one of the key hypotheses is the absence of

technological progress. The main assumption behind Malthus' work is the existence of a positive relationship between the wealth of a nation and its birth rate. Specifically, he argued that any wage level in excess of the subsistence one would incentivize individuals to expand their families, thus resulting in an increased consumption of resources for the livelihood of the population and a reduction in the proportion devoted to the accumulation of capital. Therefore, for Malthus as for Ricardo, the long run is inevitably characterized by stagnation of the economy, unless (as in China) policies directed towards controlling the level of births are implemented.

Technology is once again the fulcrum and engine of economic growth in Schumpeter's theory. The latter emphasized the entrepreneur as being the heart of the innovation process that allows the generation of short-term monopoly profits. The guarantee of compliance to rights (patents) protects the investment made by firms in the field of research. Like Smith, Schumpeter also argued that the accumulation of knowledge and the implementation of human capital produce increasing returns in the long run.

Nonetheless, this model is differentiated by the fact that entrepreneurs see innovation as the primary goal of their activity, because competition between firms is entirely based on the destruction of the monopolistic position of competitors and the acquisition of a market niche. This process, defined as "creative destruction," brings the economy into a market characterized by monopolistic competition, where increasing returns are the consequence of continuous increments in production technology and human capital. The State's role in facilitating the process of economic growth becomes crucial therefore it is based on the implementation of policies capable of creating the most fertile conditions to allow businesses to grow and create knowledge and innovation in a specific area. The said policies are economically justified when considering how avoiding less than optimal production becomes the goal of public intervention. The explicit introduction of technology in the production function also allows for the comparison of this model to those of endogenous growth.

Keynesian and Neoclassical

The Great Depression gave rise to a new generation of economists, who set out to reinterpret and reconsider the main features and determinants of economic growth, through the application of new formulations of the production function. Working within a Keynesian framework, Harrod (1939, 1942), and Domar (1946) were the first to highlight the instability of the capitalist system, given by the low substitutability between factors of production. Nonetheless, this hypothesis was overcome by Solow (1956, 1957) and Swan, who used a different production function to develop a model that had diametrically opposite consequences on economic growth. According to the latter, in the absence of technological improvements, per capita growth would tend towards a steady state (Uzawa 1965). The consequent assumption of the said model is that of convergence among different countries (Barro and Sala i Martin 1992, 1997). Subsequently, Romer (1994) and Lucas (1988) developed an analysis that took into account reproducible factors within the production function. The assumption of diminishing returns to capital was thereby overcome and endogenous growth models started to be developed.

In order to better grasp the theoretical journey described above, let us consider a generic production function with two factors of production, capital and labor: $Y_t = F[K_t, L_t, t]$, where Y_t is the production level at time t . In this expression, time takes on a role of proxy for technological development, given that, assuming equality of capital and labor, there will be a higher production level at time $t + 1$ compared to time t . The increase in capital stock at any given moment will thus be given by $\partial K / \partial t = I - \delta K = s F(K, L, t) - \delta K$, where δ is the depreciation of capital and I the level of investment, expressed in the latter part of the equation as the savings rate ($0 < s < 1$) multiplied by the abovementioned production function.

Harrod and Domar's model presents a peculiar formulation of the production function. The latter is that put forward by Leontief (1941); in other words it consists in fixed coefficients that could be

expressed as $Y = F(K, L) = \min(AK, BL)$, with A and B being positive constants. The equilibrium condition given by the full employment of factors of production is achieved at the point where $AK = BL$, whereas when the quantities of labor and capital invested are not at equilibrium, there will be cases where $AK > BL$ or $AK < BL$. In the first case scenario, only an amount of capital $(B/A)L$ would be exploited; just as in the second case scenario, it would only be an amount of labor $(A/B)K$. These two situations would imply, respectively, an underuse of capital and labor. The hypothesis of non-substitutability implied in this production function translates into a process of continued instability, as a result of the typically Keynesian animal spirit reasoning. Thus, when starting from an ideal condition of general equilibrium, where the economy's growth rate (g) is equal to the growth rate in production (g^*), there would be a surplus of supply together with a recessive spiral any time that $g < g^*$ and, vice versa, an excess in demand with inflationary expansion whenever $g > g^*$. This process would inevitably result in an economic growth determined by macroeconomic unbalance and lacking any sort of natural force that could allow for long-term equilibrium.

With the aim of overcoming the limitation posed by the non-substitutability of factors of production, the Solow-Swan model utilizes a typically neoclassical production function. This model is characterized, first of all, by the absence of a technological development variable and thus of the properties typically attributed to functions that do include the latter: namely, the factors of production have a decreasing marginal productivity, the function presents constant returns to scale, and the marginal product of capital (or labor) is infinite for K (or L) tending to zero and zero in the opposite case scenario. Through the application of this specific production function, it is possible to reformulate the equation that explains the variation of capital stock over time as $\partial k / \partial t = s f(k) - (n + \delta) k$, where all variables are expressed in per capita terms and (n) represents the population's growth rate. This equation highlights that the capital's possible evolutionary routes would be:

- $s \cdot f(k) > (n + \delta) \cdot k$: arginal increase in capital over time
- $s \cdot f(k) < (n + \delta) \cdot k$: marginal decrease in capital over time
- $s \cdot f(k) = (n + \delta) \cdot k$: constant growth in capital, in other words, characterized by the per capita values (k , y , c), which remain constant over time, and the levels of these same variables (K , Y , C), which grow at the same rate as the population (n)

In Solow and Swan's envisioning, the first two cases are unstable and converge towards the third, which represents the only long-term equilibrium possible. This is instantaneously apprehended if $s \cdot f(k)$ is interpreted as the savings quota destined to investment and $(n + \delta) \cdot k$ as the real depreciation of capital. Net investment will always be positive in the first case and negative in the second, whereas there will be no incentives to changing the level of capital in the third case scenario. Convergence towards the steady state is inevitable, determining in the long term an absolute condition of equality between countries de la Fuente (2000). According to the authors, the countries starting with similar conditions will immediately converge to the same steady state, while the ones starting off as more disadvantaged will grow at a faster rate, to then also converge. This framework, known as absolute convergence, has been criticized by Barro and Sala i Martin (1992, 1997, 2004), who sustained conditional convergence. The latter differs from the former in that it takes into account structural differences between countries.

The decreasing marginal productivity of factors of production plays a fundamental role in ensuring the process of convergence (Baumol 1986). During the 1980s, the growing international integration, together with the need to take into consideration variables that could capsize this hypothesis, led many economists to study and develop new models that could identify growth paths from within the production function. A simplified formulation of the said models is given by the equation $Y = AK$, where A represents a positive constant for the level of technology. In comparison to the production

function developed by Solow-Swan, this one is characterized by the absence of decreasing returns given by the presence of non-reproducible factors in K , particularly human capital. The latter's increasing returns counterbalanced the decreasing returns to capital, resulting in complex constant returns for each individual enterprise. When considering these new assumptions, convergence is no longer guaranteed, be it absolute or conditional. Other approaches aimed at eliminating decreasing returns to capital are based on the concept of learning by doing, introduced by Arrow (1962). The latter consists in the possibility of implementing productivity simply according to the experience acquired during previous production cycles (Sheshinski 1967). Starting from this intuition, first Romer and then Lucas implemented growth models where the accumulation of physical and human capital still guarantees constant returns for the enterprise; however, the positive externalities given by spillover effects result in increasing returns for the overall sector. In Romer's model, the (Cobb-Douglas 1928) production function becomes $Y = AK^\alpha(KL)^{1-\alpha}$, where α represents the quota of national per capita income destined to capital. The growth rate is constant, and this results in a permanent condition of steady state for the economy. Moreover, while an increase in knowledge will determine an immediate rise in the economy's growth rate, the boost of the savings quota destined to investment will permanently influence growth. Also in this case scenario, the investment in learning and new knowledge acquired by the entrepreneur will create positive externalities for the entire sector through the spillover effect, thus justifying policy interventions supporting innovation on the part of the State.

Human capital then assumes a central and relevant role in Lucas' model, where the emphasis is on the role of education as the only sector really capable of producing innovation. In his model, workers, for every unit of time, are asked to choose between work and education. Clearly, with the former option, they would benefit from a higher level of current income, whereas with the latter they would be able to improve their skills, thus obtaining higher productivity and income in

the future. Having defined H as total human capital, and h as the knowledge of a single worker, the following relation may be formulated: $H = hL$; if (u) represents the time a worker dedicates to the productive process, $(1-u)$ will subsequently be the quantity of time dedicated to education and training, so that the equation determining the accumulation of human capital is $\dot{H} = H\phi H(1-u)$, where ϕ is a scale parameter, whereas the production function including human capital is given by $Y = AK^\alpha(uH)^{1-\alpha}$. From these equations it can be concluded that human capital is, in particular, a function of the time dedicated to education and training and that its marginal productivity is constant. In terms of general macroeconomic equilibrium, it will be found that, once the steady state is achieved, the stock of physical capital, like production and other related variables in the model, will continue to rise at a rate equal to the endogenous growth in human capital.

The Role of the Public Sector in Economic Growth Models

In neoclassical growth models, the public sector holds a role of agent responsible for the exogenous stimulation of investment and the education and training of human capital. Barro (1990) built a growth model where the goods and services produced by the public sector are considered as production inputs for the public sector. In his model, goods are divided between three main categories:

- Private, rival, and excludable goods, produced by the public sector
- Public, non-rival, and non-excludable goods, produced by the public sector
- Goods produced by the public sector and subject to congestion, i.e., rival non-excludable goods, such as roads, water supply, and sewers

In the first model, where G refers to the public sector's aggregate expenditure, the very nature of the (private) goods that are produced results in the quantity of available inputs being equal to $g = G/n$, where n represents the total number of producers. The introduction of the public sector in the Cobb-Douglas (1928) production function

results in the following equation: $y = Ak^{1-\alpha}g^\alpha$, where returns of capital are assumed to be decreasing and the level of public expenditure to be fixed. Each agent adopts a production technology that combines capital and public goods provided by the State, where y , k , and g , respectively, represent product, physical capital, and productive public investment per worker. A specific assumption is that entrepreneurs, given a fixed level of public expenditure, will determine the quantity of private inputs (k) that will be employed.

If the government is looking to maintain a neutral balance, one potential financing methodology could be the introduction of a tax proportional to the quantity of outputs equal to $\tau = g/y$. The latter allows us to derive the condition of efficiency determined by a given quantity of goods produced by the public sector. In fact, since each goods unit requires an increased use of resources, the natural condition of efficiency will be given by $\partial y / \partial g = 1$, which would in turn imply $g/y = \alpha$. Under these assumptions, Barro and Sala i Martin (1992, 1997, 2004) demonstrates that the profitability of investment is independent of the economy's growth rate, which will instead be influenced by the quantity of services produced by the public sector. Moreover, with a constant τ (og/y), the model does not allow for periods of transition, but rather gives a growth rate equal, in each period, to the steady state's growth rate.

In conclusion, it is interesting to analyze the possible scenarios determined by the government managing to produce efficiently, thus respecting the condition $g/y = \alpha$. If marginal taxation (τ) is equal to zero, returns on investment, be it private or social, will be identical. However, in the case where $\tau > 0$, private returns will be lower, and a condition of Pareto efficiency may be achieved by employing a lump-sum form of taxation on consumption or through the introduction of a subsidy to the purchase of capital goods on the part of the government.

On the other hand, in the case where the public sector produces public goods as defined by Samuelson (1954), the production function is modified according to the introduction of the hypothesis of

non-rivalry in consumption. Consequently, the whole quantity G of each producer will be presented as opposed to the relative per capita quota, resulting in $y = Ak^{1-\alpha}G^\alpha$. Non-rivalry implies that the marginal product of public goods will be given by the effect of the variation in G on aggregate output $Y = y n$ and the corresponding condition of efficiency will be given at the point where $G/Y = \alpha$. Having taken into account these minor differences, the model maintains all the implications already presented in the case of the production of private, rival, and excludable goods.

However, different implications are given when the public sector produces goods that are subjected to congestion, in other words, being characterized by rivalry and non-excludability in consumption. The phenomenon of congestion derives from the difficulty of excluding enterprises from the consumption of a given rival good, which would thus be exhaustible. Thus, for a given level of production (G) of a good, each enterprise will see a decrease in the quantity available to it, as other enterprises increase their consumption of the said good. Consequently, the production function for each producer will include a relation between the total quantity of product G and the aggregate quantity of private inputs K ; therefore, $y = Ak (G/K)^\alpha$. Since the increase in capital (k) and in production (y) of a given enterprise congests the inputs (g) available to the remaining enterprises, in the absence of a proportional tax on outputs or on income, there will be an excessive consumption of “public” goods. The introduction of a tax equal to $\tau = G/Y$ will, on the other hand, be able to match the rate of returns on social and private investment, thus bringing about a Pareto optimal growth rate.

Modern Thought

While an increasing number of academics concentrated on formalizing and arguing in ever greater detail the role of human capital, education, technology, and innovation in economic growth, a group of economists, motivated by the certainty that the abovementioned elements do not constitute the engines of growth, but rather represent the consequences of the latter, set out to determine new variables that would result in being essential

for development. The said research brought them to the identification of a series of elements that may be defined as immaterial and consist in the role of institutions, regulations, faith, and cooperation within society.

Institutions and Economic Growth

According to Acemoglu (2008), institutions influence the growth trajectory. The first approach to this facet is attributable to North, who defined the concept of institutions as the rules of the game within a society or, more formally, the limits conceived by man to determine human interaction. First and foremost, it must be noted how North's definition of institutions moved away from that generally conceived in everyday language. Entities such as parliament, enterprises, universities, or associations are defined by the author as organizations and are thus considered agents for change. The key concept used as a determinant of growth is, in fact, the relationship and interaction between institutions and organizations. Since the latter are born and die as a response to the set of incentives and limits imposed by institutions, the understanding of the nature of the latter, as well as how they change, represents a key for the historical evolution of countries. When analyzing the role of institutions within the economic system, the level of cooperation is identified as being a fundamental element. By applying game theory to a neoclassical structure with limited rationality, North understood that the role of institutions consisted in the highest possible reduction of transaction costs in the exchanges between agents, with the aim of rendering them convenient and encouraging their diffusion. It is clear that the level of cooperation is inversely correlated to the number of existing institutions, since when there is knowledge of the different parts and of repeated exchange (high level of cooperation), each agent does not need to feel protected by formal regulations and thus does not consider the presence of institutions as necessary. On the contrary, when there are impersonal exchanges and a high risk of disloyal behaviors, institutions acquire a fundamental role in the reduction of information costs and in the control and implementation of agreements. In conclusion, it can be stated that

the quality of regulations (institutions) influences the level of production, through the diffusion of exchanges, and subsequently affects economic growth.

Another approach, which may be considered similar to North's, however more centered on the role of the State in the promotion of an efficient economic system, is that implied by Evans in his conception of the embedded autonomy. The latter describes an ideal condition that governs the relationship between the State and civil society. In "embedded" States, one would find public officials, and institutions in general, that are deeply intertwined with economic actors and dedicated to trying to understand and favor the interests of the latter. A rooted territoriality allows for the implementation of "fitted" policies aimed at fulfilling the competitive needs of enterprises, thus favoring their development and international competitiveness. Obviously, an excessive degree of embeddedness brings about the risk of a loss of authority and of impartiality. For this reason, the State must preserve its autonomy when the moment comes to take decisions, trying to avoid being a mere instrument for the satisfaction of individualistic interests and always preserving the collective well-being as its foremost objective. To sum up Evans reasoning, a simple comparative advantage in resources of a given country will not necessarily imply a success in the corresponding sector, but will rather depend on the functioning of the existing social and political institutions. If a wealth in resources is accompanied by a predatory State, which incentivizes renters and allows for widespread lobbying, thus losing the required autonomy in decision making, then development is unlikely to be achieved. On the other hand, a State that implements development policies and is characterized by the incentivizing of entrepreneurs, meritocratic recruitment, dedication, cohesion, and social loyalty will promote efficient and profitable investments on the part of enterprises, which will thus result in long-term growth and development.

Corruption and Economic Growth

The emphasis applied by North on the role of regulations in economic growth stimulated for

many economists an interest towards understanding how bureaucratic machines, with all their dishonesty and obstructive presence, are able to slow down the process through which technological progress transforms into new tools and productive processes.

Many theoretical and empirical articles in economic, social, and political literature have studied how corruption affects economic development. Their authors have mainly concentrated on the relationship between corruption and economic growth, not always finding coherent results. On the one hand, Leff (1964) and Huntington (1968) argued that corruption could be positively correlated to economic performance in the presence of a thick and cumbersome bureaucracy. According to their reasoning, bribery may allow firms to get things done, thus increasing their efficiency and enhancing economic growth. Subsequently, corruption could be considered as growth enhancing in that it acts as a lubricant within a rigid bureaucracy. On the other hand, corruption can be seen as a type of government inefficiency, since it discourages investments, due to the wide discretionary power of public officials, as well as reducing the quality of public infrastructure and services, decreasing tax revenue, and affecting the allocation of entrepreneurial skills, thus slowing down economic growth (Bardhan 1997; Mauro 1995, 1998). Specifically, corruption diverts the public budget destined to social services, securities, and education, health, and general services. This implies inefficient behavior on the part of the government, in its implementation of policies that are not always in the country's best interest (Powell 2004). The market allocation of resources is thus distorted, negatively influencing investments on the part of economic agents, reducing the quality of public infrastructure and services, and thus hindering economic growth (McMullan 1961; Tanzi and Dawoodi 1997; Mauro 1995). In addition, corruption affects both the total regional amount of public spending and its structure, directing expenditures towards sectors where bribes are easier to collect. In this respect, the econometric results found by Del Monte and Papagni (2007) on the Italian regional dataset "show two distinct negative effects of

corruption on economic growth. One effect seems to be that on private investment; the other is on the efficiency of expenditures on the part of public investment.” It is therefore derived that “policies to deter corruption and to increase the efficiency of local public institutions could give very positive impulses to economic growth” for the development of southern Italy.

Social Capital and Economic Growth

Within the field of study related to the immaterial aspects of economic growth, the main contributions are undoubtedly given by J. Coleman and R. Putnam, both of whom consider social capital as being an engine of growth. Despite the lack of an unequivocal and widely accepted definition of social capital, it can certainly be distinguished from the concept of human capital, given that it is related to the set of social relations available to a subject or a group of subjects. At the core of this definition is the relational element. From the latter, in fact, one can derive the definitions of three different conceptions of social capital. The first one links social capital to the idea of members of a collective adhering to shared regulations; the second identifies it as the ability of a single subject to activate and manage interpersonal relations; and the third one associates it to the individual’s capability to create networks that are useful for the achievement of one’s own objectives. From these definitions it can be inferred that social capital and institutions are two inversely correlated subjects. The higher the presence of social capital, the lower will be the need for institutions in terms of guarantor for the implementation of agreements, since social sanctions will play a more prominent role. Within this framework, Coleman’s approach is based on the neoclassical assumption of full rationality, however considering each economic agent, not individually, but as a part of a network of relationships. In the latter, there will be certain leading individuals that will impose their own rules; thus, the social relations between the subjects constituting the network will be the only leverage they will have to condition the approval and implementation of norms in their favor.

Putnam’s work, on the other hand, can be traced back to the function of social capital as

a determinant of the difference in growth between countries (Alfano and Baraldi 2012). In other words, he is interested in the analysis of the existence of loyalty and of ethical norms of reciprocity and cooperation, as a tool for analyzing the role that interpersonal relations can play in different countries’ experiences of growth.

The most important empirical contribution made by Putnam is summarized in his work “Making Democracy” (1993), which develops a comparative study of Italian regions, attributing the divergence in institutional and economic performance between the North and the South to the differences in their relative endowment of what he refers to as social capital. In fact, northern Italy developed faster than southern Italy because the former has been better endowed with social capital. Considering “a region’s chance of achieving socioeconomic development during this century has depended less on its initial socioeconomic endowments than on its civic endowments, [the] correlation between civics and economics reflects primarily on the impact of civics on economics, and not the reverse” (Putnam 1993, p. 157). In conclusion, Putnam takes into consideration a historical path dependency, according to which “social patterns, plainly traceable from early medieval Italy to today, turn out to be decisive in explaining why, on the verge of the twenty-first century, some communities are better able than others to manage collective life and sustain effective institutions” (Putnam 1993, p. 121) that promote economic growth.

Within this same field, Fukuyama’s study on trust has contributed to the increased attention to the relevance of social capital in economic growth. According to Fukuyama (1995), societies endowed with generalized trust enjoy a form of social capital that, together with traditional factor endowments such as labor and capital, contributes to their success in modern economic competition. From this conception of the role of loyalty, it can be observed that countries with an elevated endowment of social capital are differentiated by the presence of a great enterprise, while those with only a low level of social capital are characterized by a productive structure dominated by small enterprises, where both ownership and

management are family based. The author thus marks a red line between social capital, family, loyalty, and a country's ability to generate capital and utilize it in large-scale investments. In a society with little faith in those outside the family, one would encounter what is known as familism, a phenomenon that leads the family to close itself off, thus hindering relations with the wider community. The said community may be viewed as the perfect equilibrium between the excessive isolation of the family on the one hand and the disproportionate presence of the State on the other, which discourages action on the part of spontaneous groups. According to the author, it is between these two identities that social capital and trust based on shared values are built, thus creating the expectation of a correct behavior and allowing for the birth of great organizations with evolved systems of management control.

Last but not least, reference should be made to the economist Amartya Sen, who, in his work *Development as freedom*, highlights the inappropriateness of the indicator of economic growth as a proxy for the real well-being and economic progress in a society, thus providing the decisive impulse towards its abandonment, in favor of new measures of development. He highlights how economic growth is closely connected to a function of social well-being that is highly utilitarian and that does not take into account the distribution of wealth but only its total amount. By relegating the concept of economic growth to such an index, there is the risk of paradoxically considering as blossoming economies, those found in countries where there is a rapid increase in production on the one hand and a complete absence of other important elements on the other, such as freedom and life expectancy, the opportunity to escape avoidable illnesses, the possibility of finding one's desired employment, and to live in a peaceful society free from crime.

The importance of these elements in the economic growth of countries has become clear and widely accepted among economists. Even a portion of the current entrepreneurial class seems to have understood the importance of adopting policies that allow for corporate growth in a more socially friendly manner. The latter can

explain the ever-increasing implementation of corporate social responsibility (CSR) policies and social balances that highlight ethical behavior and attention to social issues on the part of the enterprise, with the aim of developing a more beneficial image both at local and at international level, thus subsequently generating higher profits. The application of this concept in balances and national policy strategies represents, without a doubt, the current frontier of the concept of economic growth. Indeed, in July 2010, the European Commission officially committed to a new policy towards sustainable and equitable development, centered on the very concept of CSR. The said policy has been named Europe 2020 Strategy and consists in a series of interventions for long-term economic growth, social cohesion, and environmental safeguard. The financial crises that started in 2008 highlighted the inappropriateness of the growth models that had thus far been followed by most states. With the aim of closing the gaps found in the said growth models and creating the preconditions for a different, smarter, more sustainable, and equitable type of development, a document has been developed describing key objectives to be achieved by 2020. The latter may be summarized as follows: with regard to labor, the goal is to increase by 75% the rate of employment for those aged between 20 and 64 years old; for research and innovation, the aim is to increase by 3% of the EU's GDP the investments in research and development; as for the safeguarding of the environment, the objective is to reduce greenhouse gas emissions by 20% compared to 1990, as well as increasing by 20% the supply of energy coming from renewable resources; with regard to education, the target is the reduction in school dropout rates by 10% and to increase by 40% the 30–34 year olds with higher education; finally, in the fight against poverty and marginalization, the European goal is to reduce by 20 million the number of inhabitants at risk of, or already living in, poverty. It is interesting to point out that current European policies on growth still contain a great portion of the economic thought analyzed above, specifically when it comes to theories on human capital, on the importance of research and development, and

on social responsibility. It should thus be noted that there is still a significant lack in policies aimed at developing social capital, and improving the quality of institutions, and that only in 2010 actions were taken towards the modification of growth models according to economic theories that had been formulated and proven years before. This institutional delay in embracing the suggestions transmitted by academics and researchers should, therefore, be taken into account as an important variable in future developments of models of economic growth.

Acknowledgement I thank Marco Bottone for helpful comment

References

- Acemoglu D (2008) Introduction to modern economic growth. Princeton University Press, Princeton
- Aghion P, Howitt P (1992) A Model of Growth Through Creative Destruction. *Econometrica* 60(2): 323–351
- Aghion P, Howitt P (1998) Endogenous growth theory. MIT Press, Cambridge, MA
- Alfano MR, Baraldi AL (2012) Is there an optimal level of political competition in terms of economic growth? Evidence from Italy. *Eur J Law Econ*. <https://doi.org/10.1007/s10657-012-9340-5>
- Arrow KJ (1962) The economic implications of learning by doing. *Rev Econ Stud* 29:155–173
- Bardhan P (1997) Corruption and development: a review of issues. *J Econ Lit* XXXV:1320–1346
- Barro RJ (1990) Government spending in a simple model of endogenous growth. *J Polit Econ* 98(5):S103–S125
- Barro RJ, Sala-i-Martin X (1992) Public finance in models of economic growth. *Rev Econ Stud* 59(4):645–661
- Barro RJ, Sala-i-Martin X (1997) Technological diffusion, convergence, and growth. *J Econ Growth* 2(1):1–26
- Barro RJ, Sala-i-Martin X (2004) *Economic growth*, 2nd edn. MIT Press, Cambridge, MA
- Baumol WJ (1986) Productivity growth, convergence, and welfare: what the long-run data show. *Am Econ Rev* 76(5):1072–1085
- Cass D (1965) Optimum growth in an aggregative model of capital accumulation. *Rev Econ Stud* 32(3):233–240
- Cobb CW, Douglas PH (1928) A theory of production. *Am Econ Assoc (Suppl)* 18(1):139–165
- Coleman JS (1990) Foundations of social theory. Harvard University Press, Cambridge, MA
- de la Fuente A (2000) Convergence across countries and regions: theory and empirics. *EIB Pap* 5(2):25–45
- Del Monte A, Papagni E (2007) The determinants of corruption in Italy: regional panel data analysis. *Eur J Polit Econ* 23(2):379–396
- Domar ED (1946) Capital expansion, rate of growth, and employment. *Econometrica* 14(2):137–147
- Evans P (1995) Embedded autonomy. Princeton University Press, Princeton
- Fukuyama F (1995) Trust: the social virtues and the creation of prosperity. Free Press, New York
- Grossman GM, Helpman E (1991) Innovation and growth in the global economy. MIT Press, Cambridge, MA
- Harrod RF (1939) An essay in dynamic theory. *Econ J* 49(193):14–33
- Harrod RF (1942) Toward a dynamic economics: some recent developments of economic theory and their application to policy. Macmillan, London
- Huntington SP (1968) Political order in changing societies. Yale University Press, New Haven
- Knight FH (1944) Diminishing returns from investment. *J Polit Econ* 52(3):26–47
- Koopmans TC (1965) On the concept of optimal economic growth. In: McNally R (ed) The econometric approach to development planning. North Holland, Amsterdam
- Leff N (1964) Economic development through bureaucratic corruption. *Am Behav Sci* 8(3):8–14
- Leontief WW (1941) The structure of the American economy, 1919–1929. Harvard University Press, Cambridge
- Lucas RE (1988) On the mechanics of economic development. *J Monetary Econ* 22:3–42
- Malthus TR (1798) An essay on the principle of population. W. Pickering, London, 1986
- Mauro P (1995) Corruption and growth. *Q J Econ* 110(3):681–712
- Mauro P (1998) Corruption and the composition of government expenditure. *J Public Econ* 69(2):263–279
- McMullan M (1961) Theory of corruption. *Sociol Rev* 9(2):181–201
- North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
- Powell BG (2004) The chain of responsiveness. *J Democracy* 15(4):91–105
- Putnam R (1993) Making democracy work: civic tradition in modern Italy. Princeton University Press, Princeton
- Ramsey F (1928) A mathematical theory of saving. *Econ J* 38(152):543–559
- Rebelo S (1991) Long-run policy analysis and long-run growth. *J Polit Econ* 99(3):500–521
- Ricardo D (1817) On the principles of political economy and taxation. Cambridge University Press, Cambridge, 1951
- Romer PM (1986) Increasing returns and long-run growth. *J Polit Econ* 94(5):1002–1037
- Romer PM (1994) The origins of endogenous growth. *J Econ Perspect* 8(1):3–22
- Samuelson PA (1954) The pure theory of public expenditure. *Rev Econ Stat* 36(4):387–389
- Schumpeter JA (1934) The theory of economic development. Harvard University Press, Cambridge, MA
- Sen A (1999) Development as freedom. Oxford University Press, Oxford

- Sheshinski E (1967) Optimal accumulation with learning by doing. In: Shell K (ed) *Essays on the theory of optimal economic growth*. Massachusetts Institute of Technology, MIT Press, Cambridge, pp 31–52
- Smith A (1776) *An inquiry into the nature and causes of the wealth of nations*. Random House, New York, 1937
- Solow RM (1956) A contribution to the theory of economic growth. *Q J Econ* 70(1):65–94
- Solow RM (1957) Technical change and the aggregate production function. *Rev Econ Stat* 39(3):312–320
- Swan TW (1956) Economic growth and capital accumulation. *Econ Rec* 32(2):334–361
- Tanzi V, Dawoodi H (1997) Corruption, public investment and growth. IMF, working paper 97/139
- Uzawa H (1965) Optimal technical change in an aggregative model of economic growth. *Int Econ Rev* 6(1):18–31
- Young A (1928) Increasing returns and economic progress. *Econ J* 38(152):527–542

Economic Impossibility

► Impracticability

Economic Integration

George Tridimas
Department of Accounting, Finance and
Economics, Ulster University Business School,
Belfast, Northern Ireland

Abstract

Economic integration is the establishment of a unified economic area where consumers and producers of different nations transact freely in a single market. Using the experience of the European Union, this essay offers a bird's-eye view of the trade-offs encountered when supranational structures pursuing collective objectives of integration may infringe on national sovereignty. The range of issues examined include: (A) Determination of policy with multiple veto players. (B) The advantage and disadvantages from centralising policy making. (C) The welfare effects of a customs union from changing the flows of trade and factors of production across different

countries. (D) The costs and benefits from adopting a single currency and its consequences for budgetary policy.

Definition

Economic integration is the creation of a unified economic area where firms and consumers from different nations buy and sell goods and services in a single market and owners of capital and labor can deploy their resources in any economic activity anywhere in the area. Integration encompasses economic, political, and legal dimensions that overlap with each other. Our understanding and assessment of economic integration is inextricably linked to the experience gained from the establishment, geographical expansion, and extension of functions of the European Union (EU). Its development has been extensively researched in economics, political science, international relations, organizational sociology, and law. The present essay does not aim to survey any part of this enormous literature.¹ Rather, it offers some pointers on the issues regarding the following issues: setting up of supranational structures that pursue collective objectives but in so doing may encroach on national sovereignty, determination of policy with multiple veto players, centralization of policy making, market competition, trade effects of a customs union, factor mobility, and adoption of a single currency, monetary and budgetary policy.

European Economic Integration

The construction of the EU started after the end of Second World War and is ongoing. Table 1 presents a brief timeline of landmark events in the development of the EU. To complete the

¹For textbook expositions, see amongst others Senior Nello (2011), Baldwin and Wplosz (2012) and Saurugger (2013). Detailed analysis of monetary integration can be found in Issing (2008) and De Grauwe (2012). The interested reader is also referred to the papers in the volume edited by Artis and Nixon (2007).

Economic Integration, Table 1 The evolution of the EU at a glance

1952	Belgium, France, Italy, Luxembourg, the Netherlands, and West Germany form the European Coal and Steel Community
1957	The six sign The Treaty of Rome, forming the European Economic Community (EEC) and the European Atomic Energy Community (Euratom)
1966	The Luxembourg Compromise is accepted, permitting member states to demand legislation to be adopted by unanimity when very important interests are at stake
1967	The three communities are united as “European Communities”
1973	Britain, Ireland, and Denmark join
1978	The European Monetary System and the European Currency Unit are founded
1981	Greece joins
1986	Spain and Portugal join
1986	The Single European Act is adopted, creating the Single Market
1990	Eastern Germany joins after the German reunification
1992	The Maastricht Treaty is signed, creating the European Union
1995	Austria, Sweden, and Finland join
1997	The Amsterdam Treaty is signed, amending the Maastricht Treaty
1998	The European Central Bank (ECB) is set up
1999	The <i>euro</i> currency is created by irrevocably locking the exchange rates of participating countries and monetary policy making is transferred to the ECB
2000	The Treaty of Nice Treaty is signed, amending earlier Treaties
2002	The <i>euro</i> ” becomes the sole currency of 12 out of 15 members (Britain, Sweden, and Denmark retain their national currencies)
2004	Poland, Hungary, the Czech Republic, Slovakia, Slovenia, Latvia, Lithuania, Estonia, Malta, and (the Greek part of) Cyprus join
2004	The Treaty of Rome establishing a constitution for Europe is signed
2005	Dutch and French voters reject the 2004 Constitutional Treaty; EU leaders suspend its ratification
2007	Bulgaria and Romania join
2007	The Treaty of Lisbon is signed, amending the Treaty on European Union and the Treaty establishing the European Community
2009	The Treaty of Lisbon enters into force after its ratification is completed
2013	Croatia joins

economic union, a number of intermediate stages must be accomplished:

1. Duty-free access to each other nation’s market, which requires the abolition of tariffs and non-tariff restrictions on trade between the member states.
2. Implementation of a common tariff on trade with non-member states to prevent cheaper imports entering the market through members states with lower external tariffs, known as a customs union.
3. Establishment of a free market in goods and services among the member states to ensure competition; this requires the harmonization of national laws and practices that regulate the market (including the relevant taxes), the abolition of anticompetitive structures, and the prohibition of state aid or other preferential treatment by member state governments to national firms.
4. Establishment of a free market in labor and capital, which is achieved by eliminating discriminatory treatment of workers on the basis of nationality, granting firms the right to establish in another member state and the removal of restrictions on capital flows.
5. Establishment of a monetary union where the member states adopt a single currency, so that prices and trading in the single market will not be affected by national currency fluctuations; in turn a single currency requires establishing a union-wide central bank to conduct monetary policy.

Each successive stage envelops its predecessor. Not all member states participate in the monetary union, with Denmark, Sweden, and the UK deciding to keep their national currencies, while countries that entered the union after the launch of the euro are expected to adopt the currency when their economies are ready to do so. A further step, one yet to be taken by the EU, is the fiscal union where taxes and transfers are decided centrally. It is however noted that contrary to the above checklist, the EU has adopted a protective agricultural policy. Each successive step implies a certain loss of national sovereignty, for example, a common tariff prevents a country to decide its own external trade policy or, with a common currency, a country can no longer implement an independent monetary policy.

Institutions of EU Governance

After the horrors of two world wars in a space of 20 years, the European economic integration was conceived as a form of international organization to bind together rival European powers, especially France and Germany, in order to prevent another war. That is, economic means were used to accomplish a political objective. What followed was a series of international treaties that over time established an intricate system of Europe-wide governance (control of decision making), changed the scope of national legislative powers, and created a body of legal acts and court decisions (known as the *acquis communautaire*) by which all member states abide, while on the economic side, it has shaped industry structures, labor market, trade, investment, and monetary flows.

Establishment of the EU by independent states meant the voluntary “pooling of sovereignty” to promote common interests and international public goods, like peace and prosperity, which are best pursued jointly rather than individually by each country. This also necessitated setting up bodies of collective decision making to pursue the common objectives and simultaneously governance mechanisms to check that the new bodies will act within the agreed limits without infringing on the rights of their creators. The main

institutions established to drive and administer the process of integration are the following.

The European Council which consists of the leaders of the EU countries sets the EU’s general political direction and priorities and deals with complex and sensitive issues that cannot be resolved at a lower level of intergovernmental cooperation. It meets four times a year and is chaired by the President of the European Council. The President of the European Commission and the EU’s High Representative of the Union for Foreign Affairs and Security Policy also take part in the meetings. The decisions of the European Council are taken by unanimity or by qualified majority, depending on what the EU Treaty provides for. Though influential in setting the EU political agenda, it has no powers to pass laws.

The Council of the European Union (not to be confused with the previous European Council), also informally known as the EU Council, or Council of Ministers, which brings together national ministers from each EU country to pass EU laws coordinate the broad economic policies of EU member countries, sign agreements between the EU and other countries, approve the annual EU budget, develop the EU’s foreign and defense policies, and coordinate cooperation between courts and police forces of member countries. As a general rule, the Council of the EU decides by qualified majority voting. From November 2014 a system known as “double majority voting” will be introduced. For a proposal to go through, it will need the support of 2 types of majority: a majority of countries (at least 15) and a majority of the total EU population (the countries in favor must represent at least 65% of the EU population). A blocking minority must include at least four Council members; if the latter fails, the qualified majority shall be deemed attained. When sensitive issues are decided, for example, security and external affairs and taxation, decisions have to be unanimous rendering veto powers to every single country. The Council of the EU represents the interests of the national governments of the member states.

The European Commission, which proposes policy measures, manages the day-to-day business of implementing EU policies and spending

EU funds and represents the EU internationally. It is an executive supranational body that represents the Community interests. It consists of 28 Commissioners, one from each country serving for a renewable term of 5 years. The President and members of the Commission are appointed by the European Council.

The European Parliament, which in an embryonic form, represents directly the interests of the peoples of Europe. Its role is to debate and pass laws in combination with the Council, debate and adopt the budget of the EU, and scrutinize other EU institutions. Its members are directly elected for terms of 5 years. Their number is 751 including its President; there is a minimum threshold of 6 members per country and a maximum of 96 (implying that smaller countries weigh higher in its composition). The members are grouped according to political affiliation and not by nationality. In comparison with national parliaments, it has significantly fewer legislative powers, but its powers have increased dramatically over time. It decides by simple majority.

The Court of Justice, whose task is to examine the legality of European Union measures and ensure the uniform interpretation and application of the EU law. It consists of one judge per EU country and eight "Advocates-General" whose job is to present opinions on the cases brought before the Court. Its members are appointed upon the common accord of the governments of the member states for renewable six-year terms. By far the largest part of the workload of the ECJ is to hear direct actions and give preliminary rulings. Direct actions concern violations of the EU law and include enforcement actions, where the Court declares whether or not a member state has infringed or complied with EU law; actions for judicial review, where the Court may annul an act of an EU institution for violating the EU law or for failing to make decisions required of them; and actions for damages, where the Court may determine the liability of an EU institution. In preliminary rulings, the Court on the request of a national court interprets a point of the EU law; this way it ensures the uniform interpretation of EU legislation.

The European Central Bank has been set up to manage the euro and maintain price stability in the

EU. More specifically, its tasks are the determination and implementation of monetary policy by setting key interest rates for the 17 countries that currently use the euro as their currency (Austria, Belgium, Cyprus, Germany, Estonia, Greece, Spain, Finland, France, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovenia, and Slovakia), the conduct of foreign exchange operations, the holding and management of the official foreign reserves of the euro area countries, and the promotion of the smooth operation of payment systems. The ECB is also responsible for framing and implementing the EU's economic and monetary policy. It is politically independent of the governments of the member states.

In comparing the role of the nation-states vis-à-vis that of the supranational bodies established to pursue and administer the objectives of European integration, two competing schools of thought appeared, namely, intergovernmentalism and supranationalism. Intergovernmentalism argues that the process of integration is controlled by the national governments of the member states which impose the policies that best suit their interests; hence, the institutions of international governance set up by the Treaties serve the purposes of their creators, usually the most powerful of the founding states, like France and Germany. On the contrary, supranationalism, or federalism, argues that economic integration, exchange, and cooperation across national borders generated a new transnational community whose interests are best served by setting institutional structures with some autonomous policy-making powers previously reserved for the nation-state.

Determination of Policy with Multiple Veto Players

The previous description of the governance organs makes clear that policy making in the EU is a complicated matter involving the strategic interaction of multiple players, where strategic means that the action of an actor takes into account the expected response of another actor whose interests are affected by such actions. The

interests of the national and supranational actors may not necessarily coincide on all issues at hand, whereas only the Commission has agenda-setting powers to initiate legislation, and decisions are subject to the veto power of the Council, the Parliament (a process known as co-decision), and, as practice has shown, the Court.

The qualified majority voting rule used by the EU (in approximately 80% of all its decisions) in combination with its supranational character raises a host of important issues. The first regards efficiency in decision making, which relates to how easy it is for the EU as a collective group to take a decision. It refers to how likely is to find a majority given the specific voting rule and the distance between the policy preferences of the national and supranational actors. In general the answer to that depends on the required majority, the number of countries, and the weight (number of votes) of each country. The second issue relates to the distribution of power among member states is approximated by the weight awarded to each member state in the Council of Ministers. In the EU, the countries with small populations, like Luxembourg, have been given greater weights than the more populous countries, like Germany. This goes a long way to explain the observed pattern of EU spending in favor of less populous countries. The third issue regards legitimacy of the EU. A decision is accepted as legitimate when it is accepted that the decision maker has

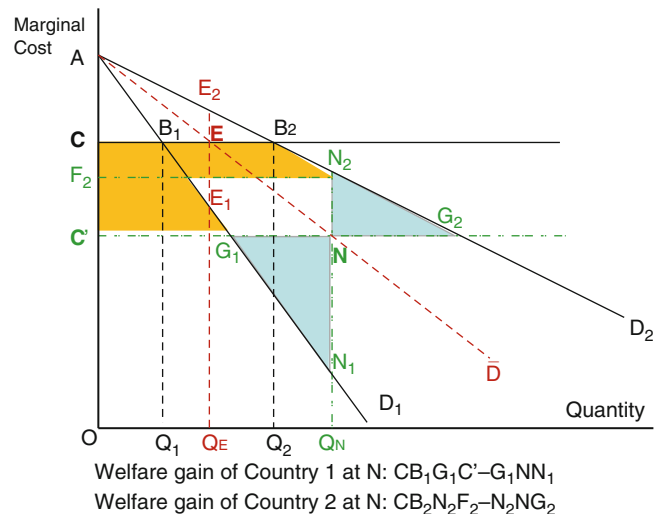
the right to take that decision. At one extreme, if the EU is considered as a union of states, as in a confederation, then legitimacy requires one vote per state. At the other extreme, if it is a union of peoples, then legitimacy requires equal power per citizen entitling more populous countries with increasing voting weight in the Council. Historically, the union-of-the-peoples approach has been the EU norm since more populous states have a larger voting weight.

Centralization Versus Decentralization of Policy Making

Economic integration requires that some policies are decided by the supranational institutions, “the central authority,” and a common policy applies to all member states. Examples include external tariff, market competition, environmental protection, arguably, monetary policy (if a single currency is adopted), financial regulation, and even some aspects of budgetary policy. This “harmonization” of policy opens up the debate of the merits of centralization versus decentralization. When the member states have different preferences for those policies but are forced to consume a greater (or smaller as the case may be) level of the service than it would have been optimal given their preferences when acting independently, a welfare loss results. This is depicted in Fig. 1 which shows the

Economic Integration,

Fig. 1 Diversity of preferences, economies of scale, and welfare of centralization



demand for a public service by two countries 1 and 2 and D_1 and D_2 , respectively, when the supply (marginal cost) of provision is C . Under decentralization the two countries act independently and consume Q_1 and Q_2 corresponding to the intersection of demand and supply. When they form a union, the central authority equates average demand to supply and provides the same quantity Q_E to both 1 and 2. Since $Q_1 < Q_E < Q_2$, that is, country 1 overconsumes and country 2 underconsumes the public service, centralization generates the welfare losses measured, respectively, by the decrease in consumer surplus EB_1E_1 and EB_2E_2 . Such losses depend on the diversity of preferences and the elasticity of demand (the distance between the demand curves and their slopes). Had the central authority the relevant information about the different preferences in the two countries, it would have been able to provide them with the individually optimal levels of the service; in the latter case, however, there would be no reason to form a union and centralize policy making.

A second argument in favor of decentralization relates to the political benefits it offers. Specifically, citizens of independent states can choose their own government, so that they have the incentive and opportunity to make informed decisions about policies that affect them and they exercise more effective control over the discretionary powers of politicians. This way the political principal-agent problem is mitigated and accountability of politicians to voters is improved. In response to this issue, the EU has adopted the principles of *subsidiarity* and *proportionality*. Subsidiarity means that a policy is assigned to the supranational (=central) authorities when it cannot be achieved by the national authorities. According to proportionality, the content and form of the EU action shall not exceed what is necessary to achieve the objectives of the Treaties.

By contrast, the arguments in favor of centralization underline the benefits from economies of scale, correction of externalities, and elimination of inefficient noncooperative behavior. Centralization often involves significant economies of scale where increasing all inputs by the same proportion increases output more than

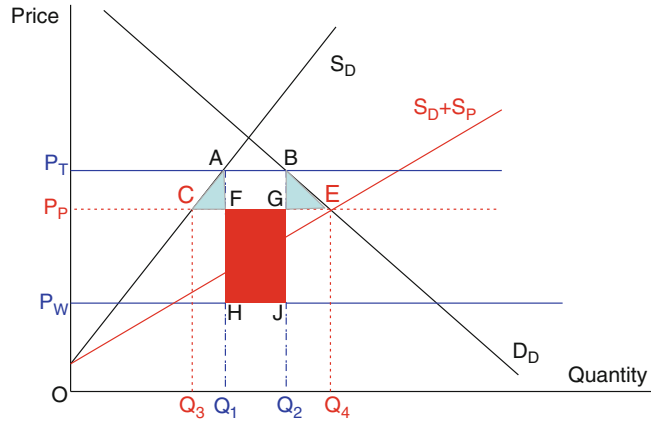
proportionally and costs rise more slowly than production, resulting in important gains for the consumer. Graphically, the presence of economies of scale that are exploited under centralization yields a supply curve at C' , lower than C . The new equilibrium obtained by the intersection of C' and is denoted by point N yields a larger equilibrium quantity for the two countries $Q_N > Q_E$. The gains from cost savings for countries 1 and 2 are shown, respectively, by the areas CB_1G_1C' and $CB_2N_2F_2$ and the corresponding losses from over- and underconsumption are represented by the triangles NN_1G_1 and NG_2N_2 . Thus, the net welfare gains from centralization for 1 and 2 are given by the differences $CB_1G_1C' - NN_1G_1$ and $CB_2N_2F_2 - N_2G_2N$.

Centralized decision making is also better equipped to address problems of positive or negative externalities (or spillovers), that is, situations where actions taken in one country (like burning fossil fuels or fishing) may increase or decrease the welfare of another country. However, the presence of externalities does not necessitate centralization, since such problems can be addressed by cooperation between the parties concerned. A third argument in favor of centralization is its ability to avoid noncooperative behavior by different countries. Countries competing against each other to attract business and mobile factors of production in their territories, which in turn increase the tax basis, may engage in games of competitive tax rate reductions (or other cost-cutting incentives, like relaxation of health and safety standards at the workplace) ultimately resulting in lower taxes and lower welfare in what is referred to as a "race to the bottom," that is, the lowest tax rate or protection for the workforce.

Trade and Growth Effects of Economic Integration

The economic rationale of integration among sovereign nations is that it promotes trade and growth and hence it increases welfare. In the short run, abolition of trade barriers allows economies to specialize according to their comparative

Economic Integration,
Fig. 2 Trade and welfare
 effects of a customs union



Net welfare effect of preferential trade: $ACF + BEG - HJGF$

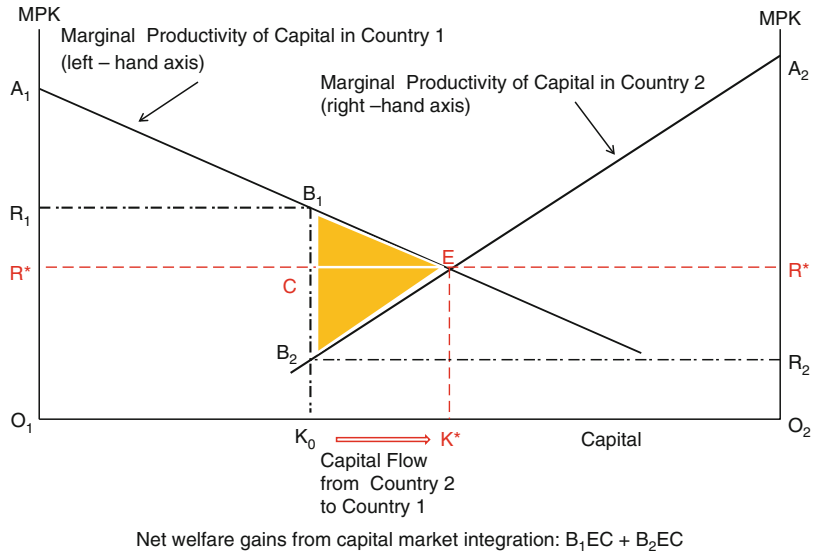
advantage which increases the volume of trade; however, against such benefits, one must set the costs from trade protection measures against non-members. These static effects of economic integration for an importing country are sketched in Fig. 2.

The lines D_D , S_D , and P_W denote, respectively, the domestic demand for the imported good, the domestic supply, and the world supply at the exogenously given P_W price. When the country applies a tariff to imports from all countries, the domestic price rises to P_T ; domestic supply and demand are shown by OQ_1 and OQ_2 , imports by Q_1Q_2 , and import expenditure by Q_1HJQ_2 . When a trade union is formed with a partner country that excludes other countries, domestic supply is represented by the line $S_D + S_P$ and equilibrium is obtained at E where D_D intersects $S_D + S_P$. Domestic demand rises to OQ_4 , and domestic supply falls to OQ_3 , implying the larger Q_3Q_4 volume of imports. Consumer welfare has increased by the sum $ACF + BEG$; this represents the gain from lowering the import price and is known as the trade creation effect. However, the Q_1Q_2 imports now enter the country at the P_P rather than P_W price paid to the partner country implying that the expenditure on Q_1Q_2 is $HJGF$. This is known as the trade diversion effect. Thence, the net welfare effect of forming the union is $ACF + BEG - HJGF$, which can be positive or negative. In other words, it is not a priori clear whether the preferential trade

arrangement will benefit or harm the country members.

The effect of factor mobility is shown in Fig. 3 using the example of capital mobility. Assume again a two-country setting, where the capital endowment of countries 1 and 2, respectively, are O_1K_0 and K_0O_2 (giving a total capital stock of O_1O_2 as shown on the horizontal axis). Lines MPK_1 and MPK_2 show the marginal productivity of capital in the two nations (or, equivalently, the national demand for capital functions), and R_1 and R_2 show the return on capital in the two countries before capital market integration, with $R_1 > R_2$. The areas defined by the MPK curves, the axis and the individual capital endowments, $O_1A_1B_1K_0$ and $O_2A_2B_2K_0$, show the outputs of countries 1 and 2, respectively. $A_1B_1R_1$ and $O_1R_1B_1K_0$ represent the sizes of labor and capital income in country 1, and $A_2B_2R_2$ and $O_2R_2B_2K_0$ show the corresponding magnitudes in country 2. When capital markets are integrated, capital flows freely from the low-return country 2 to the high-return country 1 until returns are equalized at R^* at the intersection of MPK_1 and MPK_2 ; country 1 ends up using domestically capital O_1K^* (although it owns OK_0), total output equal to $O_1A_1EK^*$, and labor income A_1ER^* (higher than before by $R_1B_1ER^*$). In country 2, output falls to $O_2A_2EK^*$, labor income falls to A_2ER^* , and capital income rises to $O_2R^*CK_0$ (upon adding the capital earnings from capital owned by country 2 but used in 1). Thus, the net effect of integration is again

Economic Integration,
Fig. 3 The effects of factor
mobility



represented by B_1EB_2 , which is the sum of the extra labor income in country 1, B_1EC , and the extra capital income in country 2, B_2EC . Clearly, although the overall integration has a positive effect on welfare, it creates winners and losers in each country, who may resist it with various degree of success depending on their political influence. Similar conclusions about increased output and its differential distribution are derived when we consider the labor market and allow immigration from a low-wage to a high-wage country.

Over the long run, integration implies that firms have access to a larger union-wide market and that they face stiffer competition than when they were confined to their national market only. As a result of more competition, costs and profit margins are squeezed and prices fall. Firms that survive become bigger and better able to exploit economies of scale driving prices further down to the benefit of consumer. The benefits of the restructuring process, however, are at risk from two sources. National governments bowing to political pressures may subsidize and otherwise assist failing firms, delaying the efficient restructuring of the economy. Secondly, the emergence of fewer but bigger firms may lead to price collusion, negating the benefits of lower prices for consumers. It is for these reasons that the EU has

introduced strict rules forbidding state aid and enforcing competition. Since integration encourages a more efficient allocation of resources, human and nonhuman, labor and capital are allocated more efficiently across different countries, boosting the rate of economic growth.

Monetary Union

Monetary union or monetary integration is an arrangement between participating countries where the exchange rates are permanently and irrevocably fixed, so that a single currency can be used by all members. Several benefits are associated with the adoption of a single currency:

- The elimination of exchange rate fluctuations and the ensuing uncertainty which discourages trade and investment.
- Reduction in transaction costs relating to conversion fees and commission charges incurred when exchanging different currencies.
- Seignorage gains from establishing the single currency as an international reserve currency; these arise from the willingness of the rest of the world to hold the single currency as an asset which then allows the monetary union to import more than it exports.

- (d) Reduction in the opportunity cost of keeping foreign reserves, since the union needs fewer reserves to manage its currency than the sum of reserves needed by each member state acting independently.
- (e) Greater effectiveness in pursuing aggregate stabilization policy. While many small open economies acting on their own cannot implement successful short-run aggregate demand management policies because of their dependence on international trade, a union can succeed by coordinating policy among member states.
- (f) For countries characterized by high inflation before the formation of the monetary union (like Italy and Greece), adopting a single currency managed by a central bank which is committed to price stability introduces a credible anti-inflationary assurance (or so the argument ran before the debt crisis of 2010).

However, such benefits may be accompanied by severe costs: giving up monetary policy independence and adopting the single currency imply that a country can no longer use the exchange rate to counteract demand and/or supply shocks. When domestic prices and wages are slow to adjust to such shocks, the exchange rate can be used to adjust domestic demand and supply to reestablish macroeconomic equilibrium quickly. With a floating exchange rate system, the exchange rate will adjust to maintain balance of payments equilibrium; monetary policy becomes effective to affect the domestic economy, but fiscal policy becomes ineffective (because capital will move in and out of the country responding to domestic and foreign interest rate differentials). On the other hand, under a fixed exchange rate regime, the authorities must respond to a surplus by reflation and to a deficit by deflation; thus, monetary policy becomes ineffective, while fiscal policy is now effective. These considerations point to the “impossible trinity” of having simultaneously a fixed exchange rate, independent monetary policy, and perfect capital mobility.

Whether or not a group of countries will benefit from forming a monetary union has been

examined by the “optimum currency area” (OCA) literature. This starts from the observation that the larger the area using the same currency, the larger its benefits; but as the area grows larger, it includes more diverse countries, which increases the costs of using the currency. Specifically, in the face of an adverse external shock that affects different countries differently (small economic losses for some but large for others), a single central bank cannot differentiate its policy responses according to the needs of the different countries. Asymmetric costs may deter the formation of a currency union. A group of countries can form a successful OCA when the following conditions are satisfied: (a) The labor force is mobile across the different countries. (b) The countries have diversified economies and produce and export similar goods. (c) The countries are open to international trade with each other. (d) They have adopted a mechanism of fiscal transfers that compensate each other for adverse economic shocks. (e) They share common preferences on how to respond to an external shock. (f) Perhaps more importantly, since none of the previous criteria may be fully satisfied, the countries have a sense of solidarity that their fates bound them together and so accept the costs of asymmetric shocks.

When countries lose the exchange rate as a policy instrument, they can use fiscal means to counteract any adverse shocks to aggregate demand, especially if fiscal transfers are not feasible (see (d) above). Despite the efficacy of discretionary fiscal policy in a system of fixed exchange rates, its use by member states of an economic union with a single currency is controversial. Acting independently, a member state running persistent large budget deficits adding to national debt may lead to severe difficulties. To bail out the highly indebted country, the central bank may increase the money supply, bringing inflation and depreciation of the single currency. Alternatively, debt accumulation may lead to higher interest rates for all country members crowding out private investment. If capital markets are efficient and assess the default risks of different governments, they will demand significantly higher interest on the debt of the profligate

government without a general interest rate increase. But if high indebtedness leads to fears about the financial stability of other countries, the commitment of the central bank to low inflation may no longer be believed, implying that the policy of no bailouts for fiscal profligacy lacks credibility. Accordingly, the risk of default is lower than otherwise. This generates a moral hazard problem where a member has an incentive for spending profligacy. Note that there may even be an adverse selection problem where only the worst offenders (those who run the biggest budgetary deficits, like Italy and Greece) are interested in joining the monetary union.

Cross-References

- [Court of Justice of the European Union](#)
- [European Nationality](#)
- [Efficiency](#)
- [Externalities](#)
- [Fiscal Federalism](#)
- [Power Indices](#)
- [Simple Majority](#)

References

- Artis M, Nixon F (2007) *Economics of the European Union*, 4th edn. Oxford University Press, Oxford
- Baldwin R, Wyplosz C (2012) *The economics of European integration*, 4th edn. McGraw Hill, Maidenhead
- De Grauwe P (2012) *The economics of Monetary union*, 7th edn. Oxford University Press, Oxford
- For annual scholarly updates on EU developments the reader is referred to the Supplement of the Journal of Common Market Studies, an academic publication dedicated to EU issues, The interested reader may also consult the EU website: http://europa.eu/index_en.htm (in English)
- Issing O (2008) *The birth of the Euro*. Cambridge University Press, Cambridge
- Saurugger S (2013) *Theoretical approaches to European integration*. Palgrave Macmillan, Basingstoke
- Senior Nello S (2011) *The European Union: economics, policies and history*, 3rd edn. McGraw Hill, Maidenhead
- The stock of scholarly work on the economic, political and legal aspects of European integration is enormous and, in view of the fast pace of the changes recent change, expanding rapidly. The following list is only a small sample of some of the most popular texts

Economic Performance

Marianna Succurro

Department of Economics, Statistics and Finance,
University of Calabria, Rende (CS), Italy

Abstract

Economic performance indicates the way in which a country or a firm functions, that is the efficiency with which they achieve their intended objectives. Alternative approaches to measuring both macroeconomic and microeconomic performance exist. Gross Domestic Product, among others, is one of the most important indicator to know how well an economy is performing and, given the existence of international standards for its calculation, it is also widely used across the world for international comparison. Efficiency and productivity, as well as long-term growth, are traditional indicators of firm performance. It is important to note, however, that all indicators are imperfect and, overall, it is important for economists and policy makers to look beyond the headline statistics to give a better overall picture of economic welfare.

Definition

Economic performance indicates the way in which an economy functions. In general, it signals the efficiency with which a country or a firm achieves its intended purposes.

What we intend to measure

...Every day, in every industrialized country of the world, journalists and politicians give out a conscious and unconscious message. It is that better economic performance means more happiness for a nation. This idea is rarely questioned... (Oswald 1997)

A prerequisite for examining alternative approaches to measuring economic performance

is an agreement on precisely what we intend to measure. This objective is closely related to the goals that societies and, hence, policy makers want to pursue. However, there is little doubt that economic policy makers need regular, timely, and accurate indicators of economic performance which, on the other hand, can relate to both countries (macroeconomic performance) and firms (microeconomic performance).

Measuring Macroeconomic Performance

The performance of an economy is usually assessed in terms of the achievement of economic objectives. These objectives can be long run, such as sustainable growth and development or short term, such as the stabilization of the economy in response to sudden and unpredictable economic shocks.

In general, to know how well an economy is performing, economists employ a wide range of economic indicators. Economic indicators measure macroeconomic variables that directly or indirectly enable economists to judge whether economic performance improves or deteriorates. Tracking these indicators is especially valuable to policy makers, both in terms of assessing whether to intervene and whether the intervention has worked or not.

Traditionally, the key measures of macroeconomic performance include:

1. Levels of real national income, spending, and output, three key variables that indicate whether an economy is growing or in recession. Like many other indicators, they are usually measured in per capita (per head) terms.
2. Economic growth, real gross domestic product (GDP) growth, GDP per capita.
3. GDP per hours worked as a measure of economic productivity, labor productivity.
4. Price levels and inflation.
5. Employment rate for the 15–64 age group and patterns of employment.
6. Unemployment levels and types.
7. Current account – satisfactory current account (e.g., low deficit).
8. Balance of payments.
9. Final consumption expenditure per capita, including government consumption.
10. Income inequality.
11. Investment levels and the relationship between capital investment and national output.
12. Levels of savings and savings ratios.
13. Competitiveness of exports.
14. Trade deficits and surpluses with specific countries or the rest of the world.
15. Debt levels with other countries.
16. The proportion of debt to national income.
17. The terms of trade of a country.
18. The purchasing power of a country's currency.
19. Wider measures of human development, including literacy rates and health-care provision. Such measures are included in the Human Development Index (HDI).
20. Measures of human poverty, including the Human Poverty Index (HPI).

Gross Domestic Product (GDP)

GDP is one of the most important indicators of economic performance. Since there are international standards for its calculation, it is also widely used across the world for international comparisons.

As GDP is a measure of a country's overall production for any given year, it is a reliable, albeit still imperfect, gauge of a country's economic performance. This is the justification for the great attention which both the general public and policy makers pay in all advanced economies to the regularly published GDP figures.

GDP, however, has several limitations.

Firstly, GDP doesn't take into account income distribution, since it could primarily benefit the top income strata of the population.

Secondly, it is characterized by well-known deficiencies related to the measurement of economic activities. Indeed, various nonmarket outputs, such as household activities and services provided free of charge, are systematically overlooked. The underground economy is difficult to capture, particularly certain criminal activities, although several attempts have been made to

harmonize the coverage of the underground economy at EU level in order to obtain comparable GDP measures. Some elements of GDP are fragile estimates, particularly those of the volume of publicly provided services and of the quality incorporated into products. Finally, some expenditures are unequivocally counted as positive contributors to economic performance, while the negative externalities associated with them – such as environmental damage – are neglected. As a result, GDP understates output.

Finally, even if GDP is a measure of market production, it has often been treated as if it were a measure of economic well-being. The measurement of GDP, however, does not address all aspects which are relevant for the material well-being of an economy; thus, it is not always a reliable guide to living standards. While the general public and many policymakers regard GDP as a measure of material well-being, this interpretation ignores the fact that production is not the ultimate goal of a society. In this context, production-based measures need to be complemented by a broader set of indicators if the aim is to assess well-being. These additional indicators would include social investment like infrastructure, education, access to health care, housing, as well as quality of life like material wealth, mental state, stress, perception of crime, environment, etc. Conflating GDP and well-being can lead to misleading indications about how well-off people are and entail the wrong policy decisions.

Making GDP a Better Measure of Economic Performance

Economic policy makers unquestionably need an economic performance indicator for short-term decision-making. Macroeconomic policy frequently operates with a time horizon of 1–2 years and, from this perspective GDP, as an indicator of current value added, is arguably the most informative gauge of economic performance. However, even in this area of economic policy, it is necessary to go “beyond GDP” by analyzing data on unemployment, inflation, short-term business activity, and consumer or business sentiment.

Although the usefulness of GDP is limited from a medium-term perspective, it still remains a viable indicator of medium-term performance. Thus, in conceptual terms, GDP remains the cornerstone of economic performance assessments. Nevertheless, it should be improved in various directions.

The most important starting points for improvements are (i) improving the measurement of service output in general and of government services in particular and (ii) making progress in measuring quality improvements.

Employment Rate

Unemployment rate is another important indicator of economic performance. However, it is heavily influenced by country-specific legislation and programs to combat joblessness. Moreover, whenever unemployment is too high and long-lasting, workers might quit the labor market, making intercountry comparisons particularly unreliable.

A more direct indicator could be the probability of being employed at working age. The employment rate in the population aged 15–64 years is often used. This basic indicator has already gained widespread acceptance in labor economics and statistics.

While such an indicator admittedly does not tell us anything about job quality or whether jobs match people’s expectations, it nevertheless means a lot when looking for a job or being exhausted by long periods of job search. It is also a sustainability indicator as it is an important parameter for the long-term future of retirement plans and public finances.

All Statistics Are Limited

It is important to note that all statistics have some limitation.

Real GDP will always be useful for showing the stage in the economic cycle. It is of some use in indicating living standards. But, it is far from the ultimate guide. Even employment rates can be partially misleading. For example, is the employment temporary or permanent? Employment figures have been better than expected, but there has been a rapid rise in labor market insecurity as well.

Therefore, there is always a need to look at several statistics at the same time, and, overall, it is important for economists to look beyond the headline statistics to give a better overall picture of economic welfare.

Measuring Microeconomic Performance

Microeconomic performance signals a firm's success in areas related to its assets, liabilities, and overall market strengths. To know how well a firm is performing, economists employ some economic indicators that directly or indirectly enable economists to judge whether economic performance has improved or deteriorated. These indicators usually include:

1. Efficiency and productivity
2. Long-term growth

Efficiency indicates that production proceeds at the lowest possible per-unit cost. Productivity is an average measure of the efficiency of production. It can be expressed as the ratio of output to inputs used in the production process, i.e., output per unit of input.

Note that productivity growth, in particular, is seen as the key economic indicator of innovation and it is a crucial factor in production performance of both firms and nations. Increasing national productivity can raise living standards because more real income improves people's ability to purchase goods and services, enjoy leisure, improve housing and education, and contribute to social and environmental programs. Productivity growth also helps businesses to be more profitable.

Firms' performance, particularly measured by efficiency and productivity, is a long-standing topic in economic studies (Coelli et al. 2004; OECD 2001).

Research developments in this direction are mostly related to methodological advancement. The empirical literature on firm efficiency has been dominated by a nonparametric approach – the data envelopment analysis (DEA) (Charnes et al. 1994; Seiford 1996). The main advantages of DEA compared to the standard econometric

technique are that: (1) it does not require any form of functional specification; (2) it is able to handle multiple inputs and outputs readily in any (in)efficiency theoretical paradigm. Based on the input and output data from DEA, a Malmquist Index can be constructed to measure productivity change. The criticism of the DEA method is related to its potential statistical shortcomings. A further development of this method is to use the bootstrap approach to obtain statistical properties.

Another well-developed method is the stochastic frontier approach (a parametric approach) (Aigner et al. 1977). Its principal advantage lies in the decomposition of deviations from the efficiency levels between noise (stochastic error) and pure efficiency; however, it faces the challenge of determining the appropriate functional forms. Recently, a semi-parametric method which combines nonparametric and parametric approaches has been applied (Bernini et al. 2004).

In addition to efficiency and productivity, firms' long-term growth – that is, sales, turnover, or profit growth – is also used to measure firm performance. Based on the production function (i.e., input–output), significant input factors are identified to explain the growth. This line of research departs from the SCP paradigm but does not seek explanations of firm growth from market structure or conduct.

References

- Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production functions. *J Econ* 6:21–37
- Bernini C, Freo M, Gardini A (2004) Quantile estimation of frontier production function. *Empir Econ* 29(2):373–381
- Charnes A, Cooper WW, Lewin AY (1994) Data envelopment analysis. Kluwer, Boston/Dordrecht/London
- Coelli TJ, Rao DSP, O'Donnell CJ, Battese GE (2004) An introduction to efficiency and productivity analysis. Springer, New York
- OECD (2001) Measuring productivity, OECD manual. OECD Publishing, Paris
- Oswald AJ (1997) Happiness and economic performance. *Econ J* 107(445):1815–1831
- Seiford LM (1996) Data envelopment analysis: the evolution of state of the art (1978–1995). *J Prod Anal* 7:99

Economics Imperialism in Law and Economics

Magdalena Małecka

TINT – Centre for Philosophy of Social Science,
University of Helsinki, Helsinki, Finland

Definition

Economics imperialism is a type of an interdisciplinary relation between the scientific discipline of economics and other disciplines of the social sciences. The disciplines that are often claimed to be “imperialized” by economics are sociology (see Granovetter and Swedberg 2001; Swedberg 1990), political science (see Hodgson 1994; Sigelman and Goldfarb 2012; Kuorikoski and Lehtinen 2010), anthropology (see Marchionatti and Cedrini 2016), geography (see Mäki and Marchionni 2010), and law (see Medema 2018; Davies 2010; Fink 2003). Hence, the law and economics movement is seen by some as one of the manifestations of the imperialism of economics in the social sciences (in this case – in legal scholarship). Below, I review the general debate on economics imperialism, and then I discuss the attempts to analyze law and economics as a case of economics imperialism.

Introduction

Recently, in the philosophy of science literature there is a discussion on scientific imperialism, of which economics imperialism is just an instance. This debate has revolved around the question of the permissibility of the application of scientific theories and methods outside the discipline in which they were initially introduced. Philosophers of science have attempted to clarify what it means for a theory or a discipline to be applied outside its own field or domain, and whether such an application can be understood as imperialistic (Dupré 1994, 2001; Mäki 2013; Clarke and Walsh 2009; Kidd 2013; Clarke and Walsh 2013; contributions to the volume edited by Mäki et al. 2018). It is debated whether scientific imperialism is a neutral or

normatively loaded term, how to establish criteria for assessment of imperialistic practices in science, how to categorize different instances of scientific imperialism, as well as whether the political metaphor of imperialism is useful at all in order to account for relationships between scientific disciplines. Mäki et al. (2018) argue that scientific imperialism challenges two central tenets of current scientific practice, as it calls into question two widely spread ideas: (1) that the broader the scope of a scientific theory, the better; (2) that interdisciplinary exchanges bring indisputable benefits (p. 1).

The term “economics imperialism” has been used in the pejorative sense by the critics of the idea of applying economics outside its domain (Fine and Milonakis 2009; Davis 2015; Hodgson 1994; Nik-Khah and Van Horn 2012) and in the approbatory sense by some economists (most notably by Becker 1971; Hirshleifer 1985; Stigler 1984; Lazear 2000; Radnitzky and Bernholz 1987; Tullock 1972) who have advocated the idea of expanding economic theories and methods to the topics studied by the less advanced, in their opinion, social sciences. “Economics imperialists” have argued that such an expansion is justified by economics’ “growing abstractness” and “generality” that economics achieves thanks to “the machine of maximizing behavior” (Stigler 1984); by the universal applicability of economics’ analytical categories (Hirshleifer 1985); and by economics’ intellectual rigor (Lazear 2000; Posner 1989). Moreover, Mäki (2009) emphasizes that the expansion of economics has been often justified in terms of the ideal of unification – economics is supposed to provide, in the opinion of “imperialists,” a unifying theory for the social sciences. Fine and Milonakis (2009) claim that the theoretical development in economics that made the expansion of economics possible was the marginalist revolution and its basic concepts of equilibrium (see: ► “[Equilibrium Theory](#)”), rationality (see: ► “[Rationality](#)”), scarcity – “universal in content and application” (p. 8).

Criticism of Economics Imperialism

However, critics of economics imperialism, apart from questioning the alleged advancement of

economics as a social science, often point out the importance of non-epistemic arguments and non-academic standing of economics for the successes of its imperialistic scientific practices. For instance, Nik-Khah and Van Horn (2012) claim that economics imperialists of the Chicago School wanted in the first place to influence policy approach by proposing their view on what constitutes the economic analysis of state policy. “Stigler’s program to study ‘governmental control’ constituted a new and distinct approach, rather than mere application of a core Chicago approach to a new domain” (p. 217). Amadae (2018) stresses the importance of politically motivated expansion of the rational choice theory – its attractiveness for nuclear deterrence made it prestigious in nonacademic contexts and in this way has strengthened its standing within academia as well. In this context the affinities of the Chicago School analysis with neoliberalism are often emphasized. Nik-Khah and Van Horn (2012) claim that economics imperialism cannot be really understood if one does not take into account the Chicago School economists’ involvement into the revival of the liberal project, known as neoliberalism – redefinition of the social and the political sphere through the concepts of neoclassical economics – and its impact on the real-world policy design. Therefore, partly in response to this “performative” aims of the Chicago School (Davis 2015), or “pedagogical” effects of the rational choice theory (Amadae 2018), some critics of economics imperialism have argued that the conceptualization of social relationships through the categories of neoclassical economics, or of rational choice, is inherently inadequate (Fine and Milonakis 2009) and can lead to ethically objectionable large-scale sociocultural consequences (Marino 2018) by influencing agents’ self-understanding (Clarke and Walsh 2009).

Is All Economics Imperialistic?

It should be noticed, however, that upon inspection we find out that economics imperialism, as each instance of scientific imperialism, is in fact a relationship between smaller epistemic units than scientific disciplines (Davis 2015; Małecka and

Lepenies 2018; Chassonnery-Zaïgouche 2018). The analysis of cases of economics imperialism makes it clear that it is never the case that the *whole* discipline of economics is “imperializing,” e.g., the *whole* sociology. Thus, as Davis (2012) argues, “economics is not all economics” (p. 210). It is usually a particular research program that develops within the discipline of economics and which is being transferred outside, as well as within, economics’ institutional borders.

In contemporary discussions, it is often claimed that only neoclassical economics, and more specifically the Chicago School and its price theory, are extended to other domains and scientific fields (Davis 2015; Nik-Khah and Van Horn 2012). However, Amadae (2018) argues that it is in fact not the neoclassical economics with its formalization of diminishing marginal utility, what constitutes economics imperialism, but rather the game theory with its view on agency as strategic competition between actors who satisfy their preferences. Guilhot and Marciano (2018) complicate the picture by showing that what is now called the economics imperialism in political science was the application of the rational choice theory to the questions of the political decision-making: it was possible rather due to the rising importance of the decision theory across the social and behavioral sciences after the Second World War. Furthermore, few researchers emphasize that some of the economic approaches, or research programs that we consider imperialistic today, were in fact quite marginal within economics, back in the days. Becker’s (see ► “Becker, Gary S.”) application of rational choice and price theory to the analysis of social phenomena is one of the examples (Chassonnery-Zaïgouche 2018; Vromen 2009). Chassonnery-Zaïgouche (2018) analyzes economic studies of discrimination and claims that Becker influenced, imperialized, with his work mainly the discipline of economics and has had, in fact, marginal impact on other social sciences studying discrimination.

Despite the seemingly intuitive appeal of the term “economics imperialism,” it is not clear what does it mean that one research program is imperializing another, or other. Economics imperialism has been defined, for instance, as

“colonisation of the subject matter of other social sciences by economics” (Fine and Milonakis 2009), as “a form of economics expansionism where the new types of explanandum phenomena are located in territories that are occupied by disciplines other than economics” (Mäki 2009), or as “the attempt to extend the core ideas of neoclassical economics to cover social science as a whole” (Hodgson 1994). Most of the definitions account for the metaphorical use of the notion of imperialism as we know it in political context. There is a discussion, however, to what extent the metaphor of imperialism should be taken seriously when talking about scientific practices. The notion of imperialism is normatively loaded and most scholars believe that one cannot ignore this normative dimension in the case of economics imperialism (Mäki 2013, who defines scientific, and economics, imperialism neutrally, is an exception here). Małecka and Lepenies (2018) provide the definition of scientific imperialism, that applies to economics imperialism, in which they account for the widely shared belief that there is something normatively problematic about economics imperialism. They stress the importance of epistemic and non-epistemic factors for being able to define and identify economics imperialism. Scientific imperialism is an activity that is related both to a certain view on the progressive character of a novel application of an existing research approach (the epistemic aspect) and to a power to favor this approach at the expense of other approaches in terms of academic and non-academic prestige, or/and resources (the institutional aspect).

Law and Economics as an Instance of Economics Imperialism

Law and economics is often seen as a result of the expansion of the Chicago School approach to law (Mercuro and Medema 2006; Medema 2015, 2018). Sometimes it is understood, more broadly, as an application of orthodox, or neoclassical economics (Jackson 1984; Davies 2010), or of game theory (Pearson 1997) to law. Mercuro and Medema (2006) even argue that law and economics

“has been the most successful of economists’ imperialistic forays into other disciplines” (p. 100).

The application of the price theory to the analysis of law is grounded on the premise that individuals are rational maximizers of their utility, that they respond to price incentives also in non-economic settings and that law can be treated as an incentive. Medema (2015) stresses the importance of Gary Becker’s (see ► “Becker, Gary S.”) works that advanced the analysis of all social phenomena with the tools of price theory (and econometrics) for the development of economic analysis of law. Medema (2015) differentiates this “new” law and economics, inspired by the work of Becker, as well as of Richard Posner (see ► “Posner, Richard”), from the “old” one, initiated by Aaron Director (see ► “Director, Aaron”), whose aim was to simply analyze the impact of legal rules on economic performance. According to Medema only the “new” law and economics has features of scientific (economics) imperialism (compare also Epstein 1997 on periodization of law and economics and Harnay and Marciano (2009) on the difference between law and economics and economic analysis on law).

Posner (1989) justifies the application of neo-classical economics to law by the rigor that this type of economic analysis allegedly offers, as well as by the possibility of funding the truly “scientific” analysis of law in this way. Cooter (1981) points out that the reason for the economics imperialism in law is the “discovery” by economists a niche in legal scholarship – a lack of quantitative reasoning. Cooter (1995) also emphasizes the importance of an attempt for unification for applying economics to law.

The historical case study made by Medema (2018) on the ad hoc Joint Committee of the American Economic Association and the Association of American Law Schools established in 1966 to explore the prospects for interactions between lawyers and economists challenges the commonly spread narrative about neoclassical economists conquering a new field. The study demonstrates that lawyers played a crucial role at the early stage of bringing economics to their studies and in replacing “the traditional methods of legal analysis” (Medema 2018, p. 110). However, later on

the application of neoclassical economics to law had been opposed by some lawyers as being too abstract an analysis, untested, irrelevant to the courtroom (Cooter 1981). Law and economics as a scientific project of explaining legal phenomena has been also scrutinized and criticized. For instance, Jackson (1984) criticized law and economics' scientism and its technocratic attitude that made it so dominant in the legal scholarship. For other critics law and economics is mainly a manifestation of "the neo-liberal project of applying neo-classical economics to state sovereignty" (Davies 2010, p. 64) that has been advanced by non-epistemic arguments and normative views on policy as complying with the goals of economic efficiency (Davies 2010; Fink 2003).

Cross-References

- [Becker, Gary S.](#)
- [Equilibrium Theory](#)
- [Posner, Richard](#)
- [Rationality](#)

References

- Amadae SM (2018) Economics imperialism reconsidered. In: *Scientific imperialism: exploring the boundaries of interdisciplinarity*. Routledge, p 4
- Becker GS (1971) *Economic theory*. Alfred A. Knopf, New York
- Chassonnery-Zaïgouche C (2018) Crossing boundaries, displacing previous knowledge and claiming superiority. In: *Scientific imperialism: exploring the boundaries of interdisciplinarity*. Routledge, p 31
- Clarke S, Walsh A (2009) Scientific imperialism and the proper relations between the sciences. *Int Stud Philos Sci* 23(2):195–207
- Clarke S, Walsh A (2013) Imperialism, progress, developmental teleology, and interdisciplinary unification. *Inter Stud Philos Sci* 27(3):341–351
- Cooter RD (1981) Law and the imperialism of economics: an introduction to the economic analysis of law and a review of the major books. *UCLA Law Rev* 29:1260
- Cooter R (1995) Law and unified social theory. *J Law Soc* 22(1):50–67
- Davies W (2010) Economics and the 'nonsense' of law: the case of the Chicago antitrust revolution. *Econ Soc* 39(1):64–83
- Davis JB (2012) Mäki on economics imperialism. In: *Economics for real: Uskali Mäki and the place of truth in economics*. Routledge, pp 203–219
- Davis JB (2015) Economics imperialism versus multidisciplinary
- Dupré J (1994) Against scientific imperialism. In *PSA: proceedings of the Biennial meeting of the philosophy of science association, Philosophy of Science Association*, vol 1994, no 2, pp 374–381
- Dupré J (2001) *Human nature and the limits of science*. Clarendon Press, Oxford
- Fine B, Milonakis D (2009) *From economics imperialism to freakonomics: the shifting boundaries between economics and other social sciences*. Routledge, London/New York
- Fink EM (2003) Post-realism, or the jurisprudential logic of late capitalism: a socio-legal analysis of the rise and diffusion of law and economics. *Hastings LJ*, 55, 931
- Granovetter M, Swedberg R (2001) *The sociology of economic life*, 2nd edn. Westview Press, Boulder
- Guilhot N, Marciano A (2018) Rational choice as neo-decisionism. In: *Scientific imperialism: exploring the boundaries of interdisciplinarity*. Routledge
- Harnay S, Marciano A (2009) Posner, economics and the law: from "law and economics" to an economic analysis of law. *J Hist Econ Thought* 31(2):215–232
- Hirshleifer J (1985) The expanding domain of economics. *Am Econ Rev* 75(6):53–68
- Hodgson GM (1994) Some remarks on 'economic imperialism' and international political economy. *Rev Int Polit Econ* 1(1):21–28
- Jackson N (1984) The economic explanation of legal phenomena. *Int Rev Law Econ* 4(2):163–183
- Kidd JI (2013) Historical contingency and the impact of scientific imperialism. *Int Stud Philos Sci* 27(3):315–324
- Kuorikoski J, Lehtinen A (2010) Economics imperialism and solution concepts in political science. *Philos Soc Sci* 40(3):347–374
- Lazear EP (2000) Economic imperialism. *Q J Econ* 115(1): 99–146
- Mäki U (2009) Economics imperialism: concept and constraints. *Philos Soc Sci* 39(3):351–380
- Mäki U (2013) Scientific imperialism: difficulties in definition, identification, and assessment. *Int Stud Philos Sci* 27(3):325–339
- Mäki U, Marchionni C (2010) Is geographical economics imperializing economic geography? *J Econ Geogr* 11(4):645–665
- Mäki U, Walsh A, Pinto MF (eds) (2018) *Scientific imperialism: exploring the boundaries of interdisciplinarity*. Routledge, Oxon
- Malecka M, Lepenies R (2018) Is the behavioral approach a form of scientific imperialism? An analysis of law and policy. In: Mäki U, Walsh A, Pinto MF (eds) *Scientific imperialism: exploring the boundaries of interdisciplinarity*. Routledge, Abingdon
- Marchionatti R, Cedrini M (2016) Economics as social science: economics imperialism and the challenge of interdisciplinarity. Taylor & Francis, Basingstoke
- Marino P (2018) Ethical implications of scientific imperialism. Two examples from economics. In: *Scientific imperialism: exploring the boundaries of interdisciplinarity*. Routledge

- Medema SG (2015) From dismal to dominance? Law and economics: philosophical issues and fundamental questions 22: 69
- Medema SG (2018) Scientific imperialism or merely boundary crossing? In: Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge
- Mercurio N, Medema SG (2006) Economics and the law: from posner to postmodernism and beyond, 2nd edn. Princeton University Press
- Nik-Khah E, Van Horn R (2012) Inland empire: economics imperialism as an imperative of Chicago neoliberalism. *J Econ Methodol* 19(3):259–282
- Pearson H (1997) Origins of law and economics: the economists' new science of law, 1830–1930. Cambridge University Press, New York
- Posner R (1989) Foreword. In: Faure M, van den Bergh R (eds) Essays in law and economics: corporations, accident prevention and compensation for losses. Maklu, Antwerpen
- Radnitzky G, Bernholz P (eds) (1987) Economic imperialism: the economic approach applied outside the field of economics. Paragon House Publishers, New York
- Sigelman L, Goldfarb R (2012) The influence of economics on political science: by what pathway? *J Econ Methodol* 19(1):1–19
- Stigler GJ (1984) Economics: the imperial science? *Scand J Econ* 86(3):301–313
- Swedberg R (1990) Economics and sociology: redefining their boundaries: conversations with economists and sociologists. Princeton University Press, Princeton
- Tullock G (1972) Economic imperialism. In: Theory of public choice. The University of Michigan Press, pp 317–29
- Vromen JJ (2009) The booming economics-made-fun genre: more than having fun, but less than economics imperialism. *Erasmus J Philos Econ* 2(1):70–99

Ecosystem Services

Sébastien Roussel

Department of Economics, CEE-M, Université Montpellier, CNRS, INRA, SupAgro, Université Paul Valéry Montpellier 3, Montpellier, France

Definition

Ecosystem services are services provided by the ecosystems and contributing to human well-being. These services play a crucial role in signaling the reliance of societies with regards to ecological systems and functions, as well as

biodiversity. Disservices stand for the opposite of ecosystem services.

Introduction

Ecosystem services correspond to a modern way to figure out the role of the environment for human society, in particular in economics through the valuation of ecosystem goods and services (Gómez-Baggethun et al. 2010; Braat 2016). These services have been promoted by increasing and complementary contributions both by policymakers and academic researchers towards their agenda to promote sustainable development, especially regarding the environmental issues at stake.

Over the past decades, landmark international initiatives to assess their social value range from the Millenium Ecosystem Assessment (MEA) (2005) and The Economics of Ecosystems and Biodiversity (TEEB) (2010) to the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (Pascual et al. 2017). In addition, international initiatives to set a policy agenda include among others initiatives by countries or group of countries (e.g., the US White House to include ES in Federal Policy (October 7, 2015) or the future EU Biodiversity strategy 2020), or initiatives by international bodies (e.g., the UN-REDD program to design and support Reducing Emissions from Deforestation and Forest Degradation + (REDD+) actions to cope with deforestation).

In terms of the history of economic thought, Mooney and Ehrlich (1997) date the modern history of the benefits got by human people from the ecosystems from the publication of Georges Perkins Marsh's book entitled *Man and Nature* in 1864. The concept of ecosystem service appeared effectively in the late 1970s (Ehrlich and Ehrlich 1981; Ehrlich and Mooney 1983), and the word service has been since used to reveal the benefits from ecological systems and functions in a utilitarianism framework, in order to highlight their social significance. According to Gómez-Baggethun et al. (2010), the role of Nature and then ecosystem services for human society depends on the considered economic perspective.

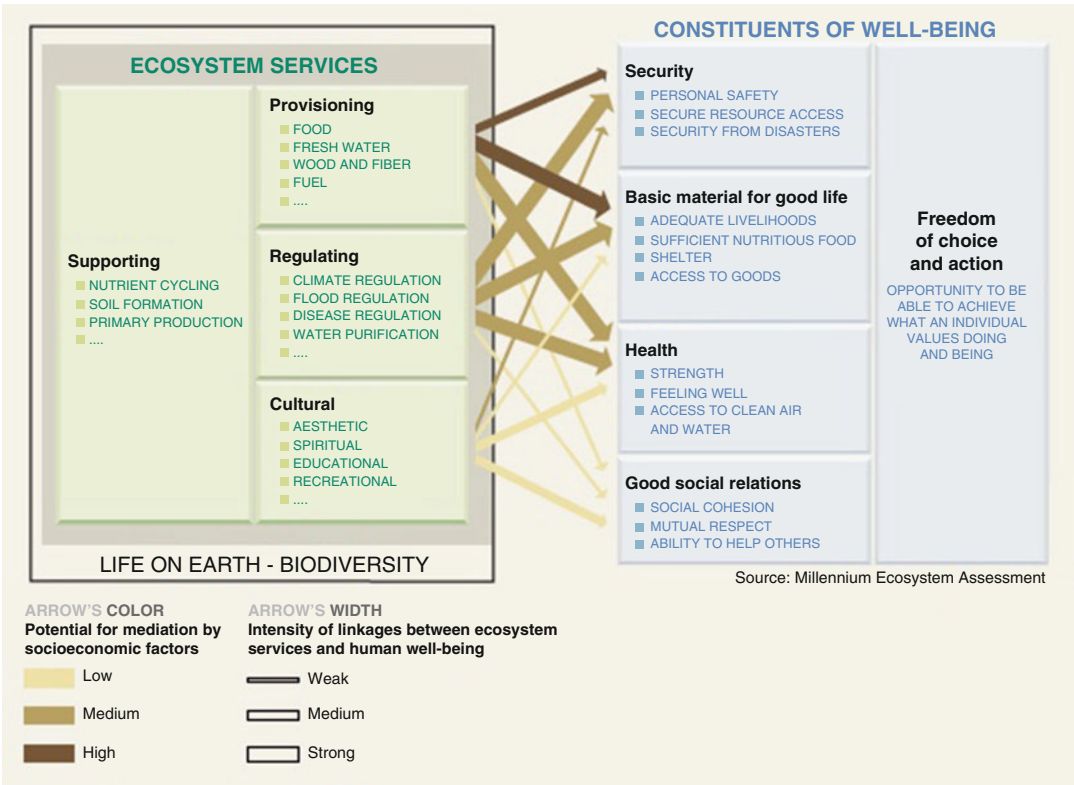
Firstly, Nature’s benefits switch from solely use values in classical economics to exchange values in neoclassical economics. Secondly and more recently, in the environmental and resource economics literature, Nature’s benefits are monetizable and exchangeable services, while in the ecological economics literature, there are controversies on monetization and commodification of Nature’s benefits. More precisely, the latter distinction depends on the status granted to the natural capital and the extent of the literature at stake, from the inclusion of market failures to social ecological economics.

Ecosystem Service Classification and Valuation

Many ecosystem service classifications have emerged over the past three decades with regards to the contribution of different types of

ecosystems to human well-being (MEA 2005; TEEB 2010). Following the revision of the System of Environmental-Economic Accounting (SEEA) under the umbrella of the United-Nations (UN), the Common International Classification of Ecosystem Services (CICES) was then developed to provide a hierarchically consistent and science-based classification (hosted by the European Environment Agency (EEA)).

Classifications of ecosystem services traditionally start from so-called supporting services that literally support other services mainly divided in provisioning services, regulating services and even maintenance services, and information services. Those services then contribute to well-being in providing security, basic material for good life, health, and social relations, and to a larger extent have public good characteristics. Figure 1 sums up these relationships in showing how ecosystem services and human well-being are linked from MEA (2005). We may underline that the status of



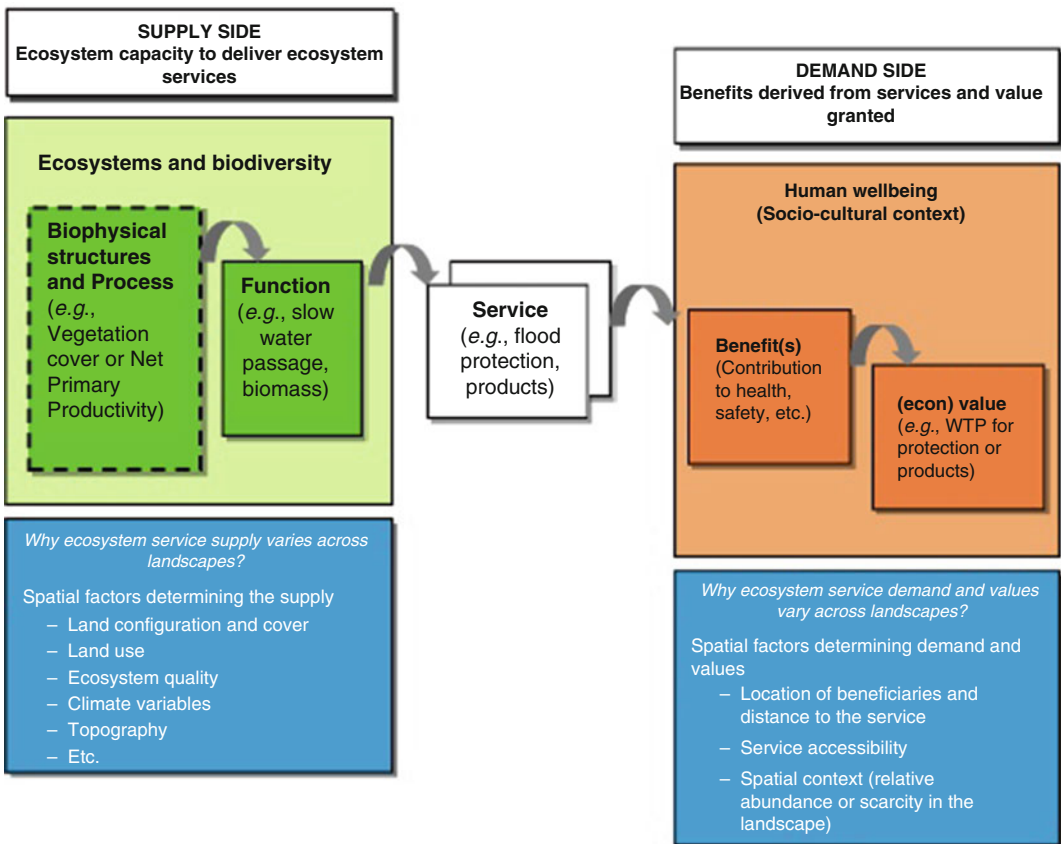
Ecosystem Services, Fig. 1 Relationships between ecosystem services and human well-being (from MEA (2005))

biodiversity remains debated among scholars as this can either be a support service through habitats for species or a single service as a whole. Moreover, relatively new contributions in the literature stress that multiple services are generated at the same time and consequently that this is suitable to use rather the concept of bundle of ecosystem services (Raudsepp-Hearne et al. 2010; de Groot et al. 2012).

Once classified and described with regards to an according typology, ecosystem services have to be valued to assess their contribution to society. Seminal papers in the literature range from methodological papers (de Groot et al. 2002) to effective value assessment even towards global values (Costanza et al. 1997, 2014; de Groot et al. 2012). Values are traditionally separated into ecological value, socio-cultural value, and economic value (de Groot et al.

2002). The economic value is then linked to use and non-use values and is measured in using economic valuation techniques either through direct market or indirect market/nonmarket valuation – avoided cost, replacement cost, travel cost, hedonic pricing, contingent valuation, and more recently choice experiment valuation – to derive willingness to pay (WTP) or willingness to accept (WTA) for the availability of these services.

Recent contributions highlight also the spatial dimension of ecosystem services and how this impacts their value (de Groot et al. 2012; Tardieu et al. 2015), either at the local scale or at the global scale through land use scenarios (Gascoigne et al. 2011). Consequently, the number, the location, and the characteristics of the beneficiaries may generate distance decays (Bateman et al. 2006; Tardieu et al. 2013). Figure 2 illustrates the spatial



Ecosystem Services, Fig. 2 Spatial dimension of ecosystem services: relationship between supply, demand and values (from Tardieu et al. (2013))

dimension of ecosystem services and therefore the overall relationship between supply, demand, and values (Tardieu et al. 2013) adapted from Haines-Young and Potschin (2010).

Environmental Law, Rules, and Incentives

Following the assessment of the significance of ecosystem services and their contribution to social welfare, the step forward consists in their inclusion in decision-making (Daily et al. 2009). Until the early 2000s, ecosystem services were indeed poorly included in cost-benefit analyses. The goal is then to integrate ecosystem services in landscape planning (de Groot et al. 2010) and more precisely strategic environmental planning and assessment (Kumar et al. 2013; Tardieu et al. 2015), to ensure the protection of the ecosystems. To this aim, environmental law can play a crucial role in allowing economic information to be used in decision-making.

According to Salzman (1998), environmental law can help for a better understanding of ecosystem services provision and then to assess their value by the creation of information markets. Indeed, the contribution of the economic analysis can take place only through a legal framework of rules and incentives. For example, the clean water acts regulation, the Code of Federal Regulations or the National Oceanic and Atmospheric Administration (NOAA) in the USA have created over the years information markets to gather data, to provide environmental impact statements, and to allow damages valuation on wetlands (Salzman et al. 2001). Moreover, the European Commission is currently preparing a Directive on the No Net Loss (NNL) of biodiversity and ecosystem services that can help to reach this aim (Wende et al. 2018).

Regarding the rules and incentives in environmental law, one could consider the following principles according to Salzman's "Five P's" (2005, 2013). These principles are respectively prescription, (financial) penalty, persuasion, property (rights), and payment. To sum up, prescription refers to the legal framework

ranging from the command-and-control regulation to the expressive function of law with regards to social norms. Financial penalties alter behaviors through financial signals (e.g., a tax mechanism). Persuasion refers to information instruments towards stakeholders regarding suitable management practices and behaviors. Property rights embody the allocation of private rights or (transferable) use rights. Finally, payment often refers to positive monetary incentives (e.g., a subsidy mechanism). These principles are complementary and are applied differently according to the ecosystems and the services at stake.

Considering the property rights and the payment principles, this contributes to the creation of market-based instruments to protect ecosystem services (Salzman 2005; Miteva et al. 2012) and to a larger extent to the implementation of payments for ecosystem services (PES) (Alix-Garcia and Wolff 2014). PES reward the landholders either to enhance their activities in favoring the provision of ecosystem services or to cover their opportunity costs in giving up their activities if these are harmful for the ecosystem services in question. PES are increasingly used in the forestry sector regarding water purification and supply or carbon storage for example (Delacote et al. 2016). Note that there is currently an increasing literature to assess the successes and failures of PES systems.

Last, brand new mechanisms to protect ecosystem services are biodiversity offsets that take place under legal constraints. Biodiversity offsets are activities that provide measurable and additional ecological gains that are equivalent to the ecological losses in an impacted area. Biodiversity offsets operate even as a market of "mitigation credits" under the control of regulators if we take the example of wetland mitigation banking in the USA (Vaissière and Levrel 2015). In this case, exchanges between gains and losses are implemented to achieve no net loss of wetland and services protection, under the supervision of the United States Army Corps of Engineers (USACE) and the United States Environmental Protection Agency (USEPA).

Cross-References

- [Non-Market Valuation](#)
- [Public Goods](#)

References

- Alix-Garcia J, Wolff H (2014) Payment for ecosystem services from forests. *Ann Rev Resour Econ* 6(1):361–380
- Bateman JJ, Day BH, Georgiou S, Lake I (2006) The aggregation of environmental benefit values: welfare measures, distance decay and total WTP. *Ecol Econ* 60:450–460
- Braat LC (2016) Ecosystem services. Oxford Research Encyclopedia of Environmental Science. <https://doi.org/10.1093/acrefore/9780199389414.013.4>
- Costanza R, d'Arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill RV, Paruelo J, Raskin RG, Sutton P, van den Belt M (1997) The value of the World's ecosystem services and natural capital. *Nature* 387:253–260
- Costanza R, de Groot R, Sutton P, van der Ploeg S, Anderson SJ, Kubiszewski I, Farber S, Turner RK (2014) Change in the global value of ecosystem services. *Glob Environ Chang* 26:152–158
- Daily GC, Polasky S, Goldstein J, Kareiva PM, Mooney HA, Pejchar L, Ricketts TH, Salzman J, Shallenberger S (2009) Ecosystem services in decision making: time to deliver. *Front Ecol Environ* 7(1):21–28
- de Groot RS, Wilson MA, Boumans RJM (2002) A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecol Econ* 41:393–408
- de Groot RS, Alkemade R, Braat L, Hein L, Willemen L (2010) Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. *Ecol Complex* 7:260–272
- de Groot RS, Brander L, van der Ploeg S, Costanza R, Bernard F, Braat L, Christie M, Crossman N, Ghermandi A, Hein L, Hussain S, Kumar P, McVittie A, Portela R, Rodriguez LC, ten Brink P, van Beukering P (2012) Global estimates of the value of ecosystems and their services in monetary units. *Ecosyst Serv* 1:50–61
- Delacote P, Robinson EJZ, Roussel S (2016) Deforestation, leakage and avoided deforestation policies: a spatial analysis. *Resour Energy Econ* 45:192–210
- Ehrlich PR, Ehrlich AH (1981) Extinction: the causes and consequences of the disappearance of species. Random House, New York
- Ehrlich PR, Mooney HA (1983) Extinction, substitution, and ecosystem services. *Bioscience* 33(4):248–254
- Gascoigne WR, Hoag D, Koontz L, Tangen BA, Shaffer TL, Gleason RA (2011) Valuing ecosystem and economic services across land-use scenarios in the Prairie Pothole region of the Dakotas, USA. *Ecol Econ* 70:1715–1725
- Gómez-Baggethun E, de Groot R, Lomas PL, Montes C (2010) The history of ecosystem services in economic theory and practice: from early notions to markets and payment schemes. *Ecol Econ* 69(6):1209–1218
- Haines-Young R, Potschin M (2010) The links between biodiversity, ecosystem services and human well-being. In: Raffaelli D, Frid C (eds) *Ecosystem ecology: a new synthesis*. BES ecological reviews series. Cambridge University Press, Cambridge
- Kumar P, Esen SE, Yashiro M (2013) Linking ecosystem services to strategic environmental assessment in development policies. *Environ Impact Assess Rev* 40:75–81
- Millennium Ecosystem Assessment (MEA) (2005) *Ecosystems and human well-being: synthesis*. Island Press, Washington, DC
- Miteva DA, Pattanayak SK, Ferraro PJ (2012) Evaluation of biodiversity policy instruments: what works and what doesn't? *Oxf Rev Econ Policy* 28(1):69–92
- Mooney H, Ehrlich P (1997) Ecosystem services: a fragmentary history. In: Daily GC (ed) *Nature's services*. Island Press, Washington, DC, pp 11–19
- Pascual U et al (2017) Valuing nature's contributions to people: the IPBES approach. *Curr Opin Environ Sustain* 26:7–16
- Raudsepp-Heare C, Peterson GD, Bennett EM (2010) Ecosystem service bundles for analyzing tradeoffs in diverse landscapes. *Proc Natl Acad Sci USA (PNAS)* 107(11):5242–5247
- Salzman J (1998) Ecosystem services and the law (Editorial). *Conserv Biol* 12(3):497–498
- Salzman J (2005) Creating markets for ecosystem services: notes from the field. *N Y Univ Law Rev* 80:870–961
- Salzman J (2013) Teaching policy instrument choice in environmental law: the Five P's. *Duke Environ Law Policy Forum* 23(2):363–373
- Salzman J, Thompson BH, Daily GC (2001) Protecting ecosystem services: science, economics and law. *Stanf Environ Law J* 20:309–332
- Tardieu L, Roussel S, Salles J-M (2013) Assessing and mapping global climate regulation service loss induced by terrestrial transport infrastructure construction. *Ecosyst Serv* 4:73–81
- Tardieu L, Roussel S, Thompson JD, Labarraque D, Salles J-M (2015) Combining direct and indirect impacts to assess ecosystem service loss due to infrastructure construction. *J Environ Manag* 152:145–157
- The Economics of Ecosystems and Biodiversity (TEEB) (2010) In: Kumar P (ed) *The economics of ecosystems and biodiversity: ecological and economic foundations*. Earthscan, London/Washington, DC
- Vaissière A-C, Levrel H (2015) Biodiversity offset markets: what is this really about? An empirical approach of wetlands mitigation banking. *Ecol Econ* 110:81–88
- Wende W, Tucker G, Quétier F, Rayment M, Darbi M (2018) Biodiversity offsets - European perspectives on no net loss of biodiversity and ecosystem services. Springer. <https://doi.org/10.1007/978-3-319-72581-9>

Efficiency

Fatih Deyneli

Department of Public Finance, Pamukkale
University, Denizli, Turkey

“An effective judiciary is predictable, resolves cases in a reasonable time frame, and is accessible to the public.”

Marina Dakolias 1999 Court Performance around the World: A Comparative Perspective

Definitions of Efficiency in Economics and in Law

The economic analysis of law relies upon micro-economic principles. Efficiency, a significant concept within microeconomic theory, is one of the key concepts in the economic analysis of law. There are different types of efficiency in economic literature. These types are defined as follows.

Productive Efficiency: The very first answer to the question “what is efficiency?” may be the ratio of the work done or output obtained by a certain amount of input. This is a state of maximum output using a certain source and technology during production. In other words, we can accomplish maximum output with the available source and technology we have. If the production with the sources and technology at hand is lower than maximum output, the state of inefficiency comes into picture (Arnold 2011).

Pareto Efficiency: Created by nineteenth century economist and sociologist Vilfredo Pareto (1848–1923), this term is a beneficial criterion in comparing the output from different economic institutions. If resources can be allocated evenly in which it is impossible to make any one individual better off without making at least one individual worse off, a Pareto improvement can be obtained. If allocation leads the way to Pareto improvement, this situation is called Pareto inefficient. In this situation, economy cannot have all that it could obtain from its resources. If we cannot find a way to make any one individual better

off without making at least one individual worse off, this situation is called Pareto efficient. In this situation, economy can have all it could obtain from its resources (Varian 2010; Besanko and Braeutigam 2011).

Kaldor-Hicks Efficiency: The application of Pareto efficiency concept to justice leads to a real problem. Firstly, there is one winner and one loser in justice, which originates from the compulsory characteristic of justice (Dimock 2005). Secondly, Pareto efficiency is also applicable to free market rules. Justice services, necessary for the implementation of law, are public commodities. Therefore, Pareto efficiency needs transformation before being used in the economic analysis of law.

Kaldor and Hicks transformed Pareto efficiency so that it can be added to the area of law which has a compulsory characteristic. Kaldor-Hicks efficiency criterion can be defined as: “a change is an improvement by Kaldor-Hicks criterion if the gainer value their gains more highly than the losers their losses” (Mathis 2009). This is basically a cost-benefit analysis. In cost-benefit analysis, when benefit exceeds cost, the project is commenced so it implies that winners may regulate losers (Cooter and Ulen 2011).

Justice System Efficiency

According to Posner, justice system has two expenditures, which are fault expenditure and direct expenditure. Fault expenditure is the failure of justice system in the distribution and other social functions. Direct expenditure is the cost of time, stationery, phone, and office work spent by judges, lawyers, and others. Posner states that fault expenditures are as real as direct ones and the economic aim should be to lower these two expenditures. The aim of justice system should be to lower both fault and direct expenditures (Posner 1973).

European Commission for the Efficiency of Justice defines justice system as a system composed of input, production, and output. These inputs, productions, and outputs within this system are summarized in Table 1 (Albers 2003). The demand side of justice system involves the

Efficiency, Table 1 Input, production, and output of justice system according to CEPEJ

Input	Production	Output
Number of courts	Average time of criminal cases	Number of resolved criminal cases
Court budget		
Information budget of courts	Average time of commercial cases	Number of resolved commercial cases
Numbers of professional judges		
Numbers of members of the jury	Average time of divorce cases	Number of resolved divorce cases
Numbers of deputy judges		
Numbers of justice staff	Average time of social security cases	Number of resolved social security cases
Numbers of other staff		
Average judge salary	Average time of immigrant cases	Number of resolved immigrant cases
Average justice staff salary		
Number of criminal cases	Average time of fiscal cases	Number of resolved fiscal cases
Number of commercial cases		
Number of divorce cases	Average time of labor cases	Number of resolved labor cases
Number of social security cases		
Number of fiscal cases		
Number of labor cases		

Source: Albers 2003

processing of incentives and constraints in lawyers' and parties' behavior (Rosales-Lopez 2008). The supply side of justice system involves factors such as the budget; number of judges, prosecutors, and assistant staff and their working time; and technology (Buscaglia and Ulen 1997).

The demand side of justice system and fault expenditure is mostly related to Kaldor-Hicks efficiency concept, while the supply side of justice system and direct expenditures is mostly related to the concept of productive efficiency.

Common Efficiency Indicators

Economic analyses have also begun to be used in the economic analysis of law. Quantitative measurements such as numbers of incoming cases, disrupted cases, and resolved cases and numbers of crime and accident have also gained importance in these analyses (Posner 2006). Due to different judicial systems in different countries, there are some differences among indicators. Despite these differences, common indicators about the efficiency, quality, and performance of justice services in international level have been used as reference point in many studies (Dakolias 1999). Developed for the measurement of justice system efficiency, these indicators benefit in objectifying the concept of justice. By this way, the concept of justice becomes appropriate for economic and econometric analyses.

The most commonly used one of these indicators is clearance rate. International institutions make use of other indicators to analyze the efficiency of justice services. Table 2 provides the websites of

Efficiency, Table 2 International institutions working on the efficiency of justice system

Name of institution	Web page
OECD	http://www.oecd.org
World Bank	http://www.worldbank.org/
European Commission	http://ec.europa.eu/justice/
The European Commission for the Efficiency of Justice (CEPEJ)	http://www.coe.int/t/dghl/cooperation/cepej/default_en.asp
National Center for State Courts	http://www.ncscinternational.org/
International Consortium for Court Excellence	http://www.courtexcellence.com/

international institutions for more indicators and analysis of these indicators on country basis.

Clearance rate is calculated by dividing the number of crimes that are “cleared” (a charge being laid) by the total number of crimes recorded. It may also be defined as the response of justice system to the clearance rate and demand in justice system – applications to the justice system. If clearance rate is more than 100%, courts begin to deal with the cases remained from the previous year. If clearance rate is less than 100%, there will be cases unresolved (Dakolias 1999). Clearance rate is measured as follows (CEPEJ 2012):

$$\text{Clearance rate} = \frac{\text{Resolved cases in a period}}{\text{Incoming cases in a period}} \times 100$$

References

- Albers P (2003) Evaluating judicial systems “a balance between variety and generalisation”. CEPEJ 12, Strasbourg
- Arnold R (2011) Microeconomics, 11th edn., South-Western Cengage Learning, USA
- Besanko D, Braeutigam R (2011) Microeconomics, 4th edn., NJ: Wiley, USA
- Buscaglia E, Ulen T (1997) A quantitative assessment of the efficiency of the judicial sector in Latin America. *Int Rev Law Econ* 17:275–291
- CEPEJ (2012) CEPEJ report evaluating European judicial systems – 2012 edition (2010). CEPEJ studies no 18, Belgium
- Cooter RB, Ulen T (2011) Law and economics, 6th edn., Pearson, USA.
- Dakolias M (1999) Court performance around the world: a comparative perspective. *World Bank*, 83 WTP no 430, Washington, D.C
- Dimock S (2005) Law and economics. Classic readings and Canadian cases in the philosophy of law. Prentice Hall, Toronto
- Mathis K, Shannon D (trans) (2009) Efficiency instead of justice? Searching for the philosophical foundations of the economic analysis of law. Springer Science +Business Media B.V, New York
- Posner RA (1973) An economic approach to legal procedure and judicial administration. *J Legal Stud* 2(2):399–458
- Posner RA (2006) A review of Steven Shavell’s foundations of economics analysis of law. *J Econ Lit* 44(2):405–414
- Rosales-Lopez V (2008) Economics of court performance: an empirical analysis. *Eur J Law Econ* 25:231–251
- Varian HR (2010) Intermediate microeconomics, 8th edn. W.W. Norton, New York

Efficiency, Types of

Henk J. ter Bogt

University of Groningen, Groningen, The Netherlands

Abstract

Efficiency is an important concept in the field of law and economics. The precise meaning of the term efficiency is not unequivocal and depends on the context in which it is used. In practice, efficiency often relates to economic efficiency. However, such forms as social and political efficiency can also be very important for the continuity of a society or an organization.

Efficiency types in organizations and society: Different kinds of measures which all indicate the optimal use of specific material or immaterial scarce resources in organizations and society.

Introduction

Efficiency is a well-known word in everyday language. It is also a core concept in the law and economics literature, a field which deals with the economic effects of legal rules and constraints and the way in which these rules and constraints can contribute to “efficiency.” To put it in more general terms, it could be claimed that the traditional emphasis of the law and economics literature has been on finding efficient legal rules and structures, i.e., instruments which promote economic efficiency. This objective implies finding an optimal allocation of the resources available and realizing a maximum quantity of outputs via these means. Laws and rules can play a role here (Coase 1960; Calabresi 1961). In principle people are supposed to act rational in the law and economics literature, which implies that based on their economic rationality, they strive for economic efficiency. In later additions to the law and economics literature, however, efficiency is used in broader terms,

referring to more aspects than only economic efficiency.

The specific focus on “efficiency” might suggest that the precise meaning of the concept is always clear. However, in economics different types of efficiency can be distinguished. And in later elaborations of the law and the economics literature, even more types have been added, such as social efficiency and political efficiency, each with their own specific definitions. All of these definitions may refer to “maximizing” something, i.e., an output, by using a minimum amount of “inputs.” What has to be maximized, and what the inputs and outputs are, varies and relates to the specific angle and subject area chosen. This optimization could relate, for example, to economic welfare, but also to public benefit (which can include immaterial and nonmonetary elements). Furthermore, in the later additions to the law and economics literature, human behavior has received increasing attention (see, e.g., Jolls et al. 1998; Korobkin and Ulen 2000). This has been done, for example, via concepts such as bounded rationality (people may want to choose the best alternative available to them, but they can only acquire full knowledge of a limited number of alternatives) and bounded self-interest (although people may be focussed on their own interests and utility, this does not necessarily imply that they maximize utility without caring about others). Despite these refinements, however, in practice the ultimate focus of much of the law and economics literature still seems to be on the realization of optimal economic efficiency.

Below, first, the basic concepts of inputs and outputs will be introduced, followed by a discussion of the more traditional economic efficiency concept and the elements in which it can be subdivided. Next, some other types of efficiency will be discussed briefly and a conclusion will be presented.

Inputs and Outputs

Some of the concepts mentioned above require some further explanation, because they are the basic components in the analysis of efficiency.

The various forms of efficiency all deal with inputs and outputs. In a production process, examples of the *inputs* required to produce a certain output, i.e., a good or a service, are direct and indirect labor hours, capital, materials, machine hours, ICT, housing, knowledge, and transport. These inputs can also be called resources. Together, the inputs are the costs made in the production of the outputs. The *outputs* are the “products” produced, for example, the goods and services, which are the result of a production process in an organization. So the production process transforms the inputs into outputs (which may lead to certain effects, i.e., outcomes considered desirable by an individual or society, resulting, e.g., in utility maximization). In order to be able to measure and compare the different inputs, they generally have to be expressed in monetary values.

Economic Efficiency and Its Dimensions

The term efficiency as commonly used generally refers to the ratio between the inputs employed and the outputs realized. More precisely, it refers to the maximization of output produced by a unit of input. If more output can be produced per unit of input, the efficiency increases. Linking outputs to inputs in this way could also be called measuring the productivity. Depending on how much output is produced by a unit of input, the common term “efficiency” can also indicate organizations as being less or more efficient. Therefore, the general criterion based on which different organizations are compared is their “efficiency.” However, in the academic economic and law and economics literature, many authors define “efficient” in more specific and absolute terms.

When the term efficiency is used in the field of law and economics, it generally refers to the so-called economic efficiency, which can be subdivided into two types: productive or technical efficiency and allocative efficiency. These categories together form the overall economic efficiency (Coelli et al. 1998, pp. 3–5).

In an economic context, the first type – productive or technical efficiency – is defined as

follows: “efficiency is the difference between the actual input/output ratio and the ideal ratio” (Jackson 2011, p. 16), all other things – such as the production technology and the quality of the output – remaining equal. So the ideal ratio is considered to be the result of the optimal possible use of a certain quantity of input, and in that situation the input yields the maximum quantity of output attainable on the basis of this input (i.e., the “production frontier,” as it is labeled by economists, has been reached). In economics, this type of efficiency is called *productive* or *technical efficiency* (terms which can be considered as synonyms). Productive (or technical) efficiency relates to an organization’s production costs. It is achieved when the organization’s outputs are produced at minimum average costs (i.e., only if this point of minimum average costs is achieved can one speak of an organization being “efficient”).

Although organizations and/or society can be productively efficient, this does not automatically imply that the products produced are valued by the citizens/consumers. The other dimension of economic efficiency, called the *allocative efficiency*, therefore addresses the question whether the “right products” are produced. If a company would produce steam engines in a productively efficient way whereas nobody needed these items, a major problem could arise concerning using the resources for this production process. Allocative efficiency indicates whether the resources in society are allocated in an optimal way, given a particular behavioral assumption, for example, utility maximization (but the behavioral assumption could also be, e.g., profit maximization or cost minimization). This means that allocative efficiency concentrates on whether a society’s resources (inputs) are allocated to those domains of activity and output which bring the citizens/consumers the maximum utility (where in this context utility is supposed to be related to the availability of economic goods). Expressed in more technical economic terminology, allocative efficiency refers to a situation where the price which consumers are willing to pay for a product (this price indicates the utility or value attached to the product) is equal to the cost of the resources used for realizing this product (i.e., the marginal

cost of the product). In an alternative definition allocative efficiency indicates the situation in society in which the available resources (inputs) are used in such a way that it would not be possible to increase one person’s utility without decreasing that of another. In economics, the latter example is also called a Pareto-efficient or Pareto-optimal allocation.

A concept related to the productive or technical efficiency category is the so-called *X-efficiency*. Being X-efficient means that an organization employs its internal resources in an optimal manner, i.e., in such a way that the average costs are reduced to a minimum. However, if the competition in the market is imperfect, for example, because of a monopoly, a situation of X-inefficiency may exist, also on a longer-term basis (Leibenstein 1966). Here the organization is not capable of achieving productive or technical efficiency and produces its outputs at higher costs per unit than technically considered necessary.

The above depicts the most frequently used and most important dimensions of economic efficiency. However, apart from the strictly economic forms of rationality, there are also other variants relevant in practice, which are not necessarily of a higher or lower order. Next, an outline is given of two of them.

Social Efficiency

Although social efficiency relates to an assessment of the welfare in a society in a mainly economic sense, it has more recently also been used in connection with all kinds of noneconomic topics (e.g., social, political, and cultural) which can play a role in and influence the public benefits. When used in this broader context, social efficiency refers to the maximization of certain “outputs” in society (e.g., the contribution to an objective such as sustainability or social justice) via the inputs available (see, e.g., Lefebvre and Vietorisz 2007).

The more traditional use of the term social efficiency, however, goes back to Coase (1960). In that context *social efficiency* indicates the maximization of the social welfare, which means

a maximization of the revenues obtained. This maximization of revenues suggests an economic focus. The revenues are considered to be maximized if the inputs and outputs are used in such a way that the social cost of production is equal to the social benefit (or, in more technical economic terminology, if the social marginal benefit of the consumption of the output is equal to the social marginal cost). If expressed in this way, social efficiency seems to be closely linked to allocative efficiency. However, the concept of social efficiency includes paying attention to the existence of the so-called externalities (and markets which are not perfectly competitive). When focussing on welfare at the level of society, the effect of production on different individuals or groups is considered with the aim to maximize the social welfare, i.e., attaining social efficiency.

As social efficiency and “externalities” are important in much of the law and the economics literature, they are dealt with in some more detail. External effects relate to the influence of the activity or action of an individual or organization on other individuals or organizations. Such an influence could be, for example, the pollution caused by the activities of a factory in the gardens of the people living in the vicinity. This effect, an externality, should be taken into consideration when social welfare has to be determined. In the case of the factory, elements should be taken into account like the negative effect of the production activities on the living environment of the people in the surrounding area and the value of their property. In principle, the factory could negotiate about a compensation for such negative effects with each individual neighbor. However, if the factory would have to negotiate with all its neighbors individually, its “transaction costs” may rise to an unacceptable level (in this case the transaction costs could include, e.g., the costs involved in searching for alternatives to avoid or compensate the pollution, negotiations, and the monitoring of the agreements made). Here certain legal rules which prescribe how to act in the case of an externality (and what kind of negative effects has to be accepted by one or both of the parties) could be helpful in reducing the cost of making contractual arrangements (and in reducing

transaction costs) and approaching an optimal economic result for society (Coase 1960; Calabresi 1961).

Although social efficiency as developed on the basis of the ideas of Coase and Calabresi mostly refers to economic welfare, Coase (1960, p. 43) suggested that the analysis “of different social arrangements for the solution of economic problems should be carried out in broader terms than” only the economic context (i.e., it should be broader than only the “comparison of production values as measured by the market”). If such a broader interpretation of the concept of efficiency is adopted, it seems to be very similar to the abovementioned maximization of social benefits.

Political Efficiency

In a political environment, the goal function, i.e., *ter what has to be maximized* (the “outputs”), may differ from that in a “strictly economic” context, which also applies to the inputs concerned. This could particularly be the case when special interest groups try to influence political decision-making (Buchanan and Tullock 1974, pp. 284–287, 302–305). Since the law and economics literature considers both the effects of legal rules and the way in which they develop and are ultimately formulated – which is basically a political process – several authors have also included the political context in their work.

A politician who has to decide about the content of a law could be interested in its effects in terms of both economic and social efficiency. Taking economic and social efficiency into account can be important in avoiding dissatisfaction from within society. However, the politician may also be interested in the effect of the legal rules on his/her voters and the number of votes he/she and his/her party will receive in the next elections, i.e., political efficiency (Buchanan 1993; Gavious and Mizrahi 2002). From the perspective of political rationality, i.e., to survive as a political party or a politician, a rational approach would be to try to maximize the electoral support by using minimal “efforts.” *Political efficiency*

could then be expressed “as a ratio of the amount of ‘effort’ (including money and other production factors) made by politicians to the amount of electoral support gained by means of the politicians’ policies” (ter Bogt 2003, p. 180). The elements on which the politician’s efforts should focus can vary from case to case. This is because elements such as the public interest and social welfare are generally not that clearly defined in practice and can include many different components, which can also differ among citizens (who are the voters in elections). Furthermore, the political support of voters may not only depend on the economic efficiency of the government policy but also on other factors, for example, equal treatment, a focus on sustainability, and social justice (see also Wildavsky 1966, pp. 308–309).

To Conclude

In principle, depending on the specific context, the type of rationality may vary, and herewith the relevant type and definition of “efficiency.” Economic and law and economics literature were traditionally mainly focussed on economic rationality and efficiency. However, social efficiency and the related concept of social rationality have now also become important elements in the law and economics literature. Social efficiency as it is mostly used in practice is closely linked to economic welfare in society. But it can also include “noneconomic” elements, such as sustainability or social justice, and pertain to the broader level of public benefits. The idea of political rationality and political efficiency applies more specifically to political environments (although it seems that in practice “politics” can play a role in any organizational context). The political efficiency idea is partly related to Granovetter’s suggestion that in order to better understand economic activity and the organizational arrangements chosen, it is necessary to take both economic and social relations into consideration (Granovetter 1985, pp. 490–491).

In this way the law and economics literature and the concept of efficiency have gradually been

extended. The literature no longer only focusses on economic aspects but also includes behavioral elements, such as bounded rationality, as well as goals and outputs other than strictly economic ones. It now tries to include more elements from the “real world.” This approach leads to different types of efficiency, playing a role in different contexts. As might have become obvious from the examples given above, the operationalization and proper measurement of the various concepts and types of efficiency in empirical research can be problematic, even in the case of economic efficiency. However, this difficulty does not imply that it would be a futile exercise to consider – and if otherwise not possible only in qualitative terms – the effects of changes in rules and structures on a specific type of efficiency in an organization or in society.

References

- Buchanan JM (1993) How can constitutions be designed so that politicians who seek to serve “public interest” can survive and prosper? *Constit Polit Econ* 4(1):1–6
- Buchanan JM, Tullock G (1974) *The calculus of consent: logical foundations of constitutional democracy*. University of Michigan Press, Ann Arbor
- Calabresi G (1961) Some thoughts on risk distribution and the law of torts. *Yale Law J* 70(4):499–553
- Coase RH (1960) The problem of social cost. *J Law Econ* 3(1):1–44
- Coelli T, Prasada Rao DS, Battese GE (1998) *An introduction to efficiency and productivity analysis*. Kluwer, Boston/Dordrecht/London
- Gavious A, Mizrahi S (2002) Maximizing political efficiency via electoral cycles: an optimal control model. *Eur J Oper Res* 141(1):186–199
- Granovetter M (1985) Economic action and social structure: the problem of embeddedness. *Am J Sociol* 91(3):481–510
- Jackson PM (2011) Governance by numbers: what have we learned over the past 30 years? *Public Money Manag* 31(1):13–26
- Jolls C, Sunstein CR, Thaler R (1998) A behavioral approach to law and economics. *Stanford Law Rev* 50:1471–1550
- Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. *Calif Law Rev* 88(4):1051–1144
- Lefebvre L, Victorisz T (2007) The meaning of social efficiency. *Rev Polit Econ* 19(2):139–164
- Leibenstein H (1966) Allocative efficiency vs. “X-efficiency”. *Am Econ Rev* 56(3):392–415

- ter Bogt HJ (2003) A transaction cost approach to the autonomization of government organizations. *Eur J Law Econ* 16(2):149–186
- Wildavsky A (1966) The political economy of efficiency: cost-benefit analysis, systems analysis, and program budgeting. *Public Adm Rev* 26(4):292–310

Efficient Market

Michael Mitsopoulos¹ and Theodore Pelagidis²

¹Brookings Institution, Athens, Greece

²Brookings Institution, University of Piraeus, Piraeus, Attica, Greece

Introduction

In this article, we deal with some pieces of evidence that are needed to explain the paradox of rapid GDP growth in the face of the dismal competitiveness and inefficiency of the Greek economy during 1995–2008. We show how Greece's economy structural inefficiencies – inefficiency of justice included – have hit the domestic economy, and we present their impact on the current turmoil of the economy. We offer a specific explanation of the current unfortunate state of the economy, and we briefly summarize the reforms already undertaken towards efficiency. We also suggest avenues of necessary reforms to increase efficiency and overcome the current crisis.

The Engines of Growth, 1995–2008

In Greece, certain positive developments led to the strong growth performance observed since the mid-1990s and up to 2008. Figure 1 shows how Greece clearly outperformed, since 1995–1996, the benchmark eurozone economy. However, it is absolutely crucial to look at the factors of “growth” to see why, at least in the great part, this was superficial, fragile, and not based on the improvement, the deepening, or the expansion of domestic production.

These developments include, primarily, the proper liberalization of the credit markets at the

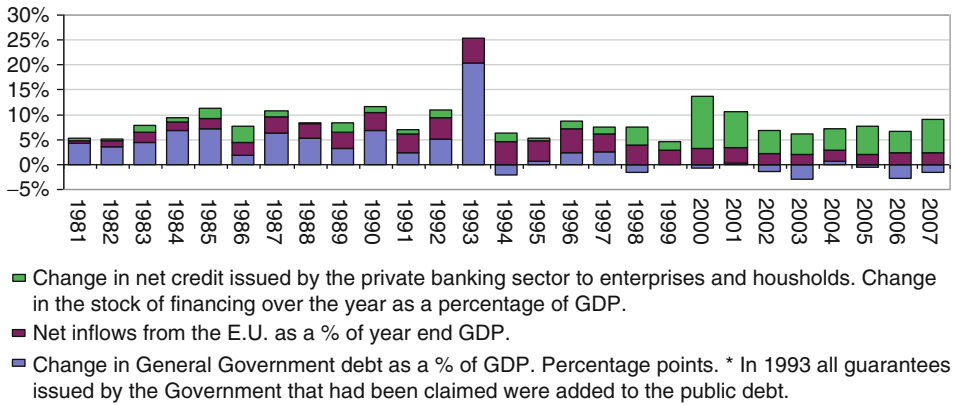
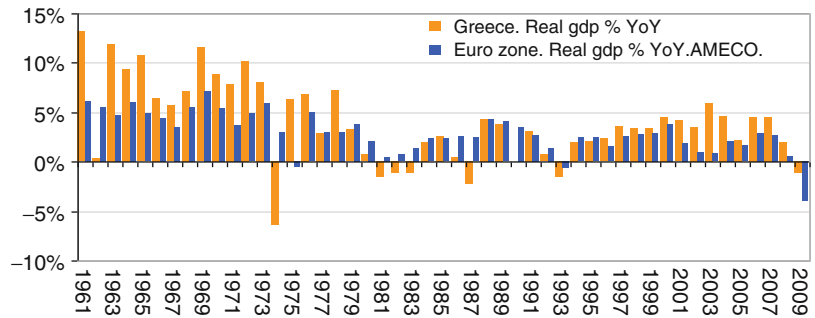
beginning of the 1990s, completed by the end of the 1990s. This was coupled with entry to the European Monetary Union. Combined these two developments lead simultaneously to macroeconomic stabilization and a steady increase of private credit after 2000. It has also to be stressed that the expansion of private credit replaced after the beginning of the 1990s the government deficit spending as the main way to finance the expansion of consumption in Greece, although data should be treated with caution. The most possible is that fiscal expansion reinforced private credit and private consumption expansion. As Fig. 2 shows by measuring demand injections to the GDP, the impact of these injections was important as a percentage of GDP for every year during a prolonged period that spans all the duration of Greece's strong performance.

The contribution of the stabilization of the macroeconomic outlook of Greece in the wake of EMU accession towards the expansion of private credit was significant, as is shown by the rapid fall of interbank rates after 1998 (Fig. 3), which reflect also the decline in the rates offered by commercial banks to households and businesses. (It also brought a significant fall of the inflation differential of Greece with respect to the eurozone average during the same period). It can be seen clearly how the expansion of credit to households fuelled the growth of private consumption during the past years (Fig. 4), and only just the period preceding the completion of the infrastructure projects that were prepared to be ready for the 2004 Olympic Games, private consumption kept accelerating in spite of a lull in the explosive growth of private sector credit.

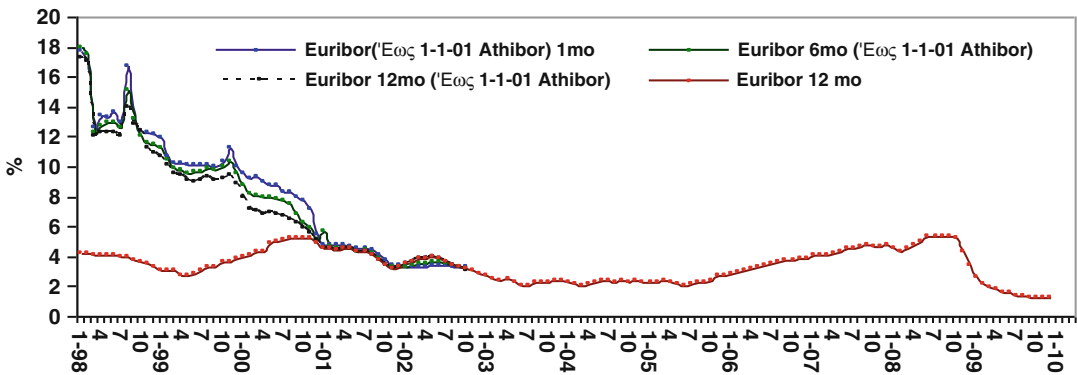
But this exception is easily explained by the peak in the investment growth rate during that time (Fig. 5).

Besides the credit expansion, two other factors contributed significantly to Greece's growth performance during the 2000s. First is the shipping and tourism industry. These secure significant annual revenue inflows of about 25% of GDP that are added to the domestic demand and help to mitigate the huge trade balance deficit. Secondly, the fiscal stimulus given by the 2004 Olympic Games nourished through public borrowing

Efficient Market, Fig. 1 Real GDP (Source: AMECO)



Efficient Market, Fig. 2 Demand injections (Source: Bank of Greece, Ministry of Finance, European Commission Budget, and EUROSTAT)

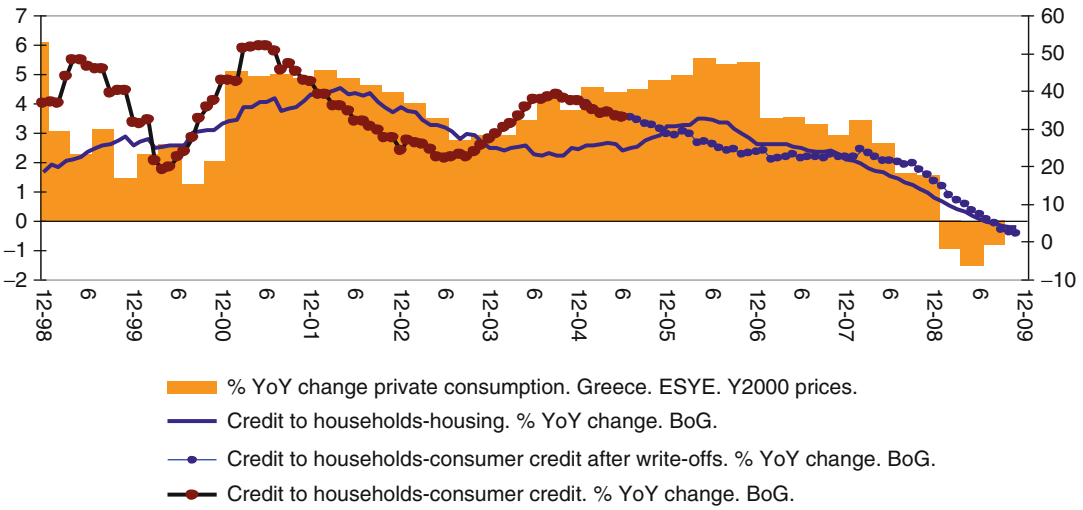


Efficient Market, Fig. 3 Interbank rates

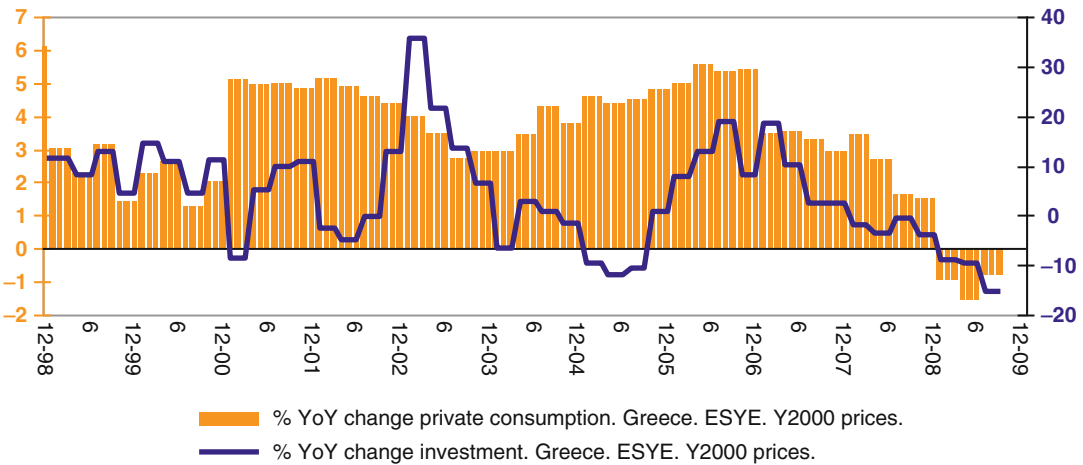
and that led to the improvement of certain key infrastructure facilities.

The rapid increase of new investment, both public and private (Fig. 6), also demonstrates the impact of the infrastructure investment that was largely financed by the EU structural funds. Still,

the rush into EU-financed infrastructure investment did not only contribute to investments and consequently to the creation of new jobs, as in the end many of these projects, when finished, actively boosted to some extent the productivity in the area surrounding Athens. The inflow of



Efficient Market, Fig. 4 % YoY change private consumption and Credit to households



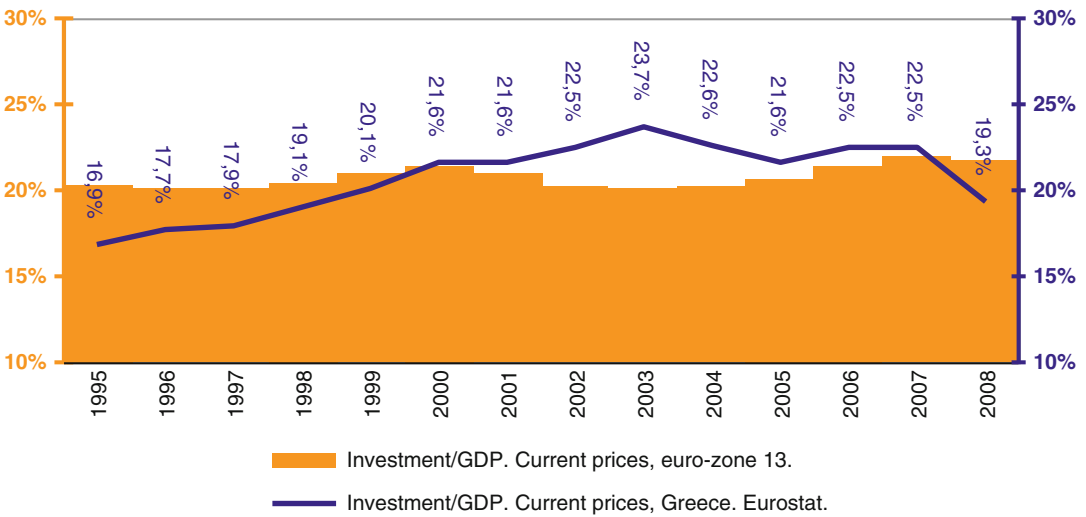
Efficient Market, Fig. 5 % YoY change private consumption & investment. Greece. ESYE. Y2000 prices

funds from the European Union, within the context of the European Union structural funds and the Common Agricultural Policy, also contributed largely to the improvement of key productivity enhancing infrastructure facilities.

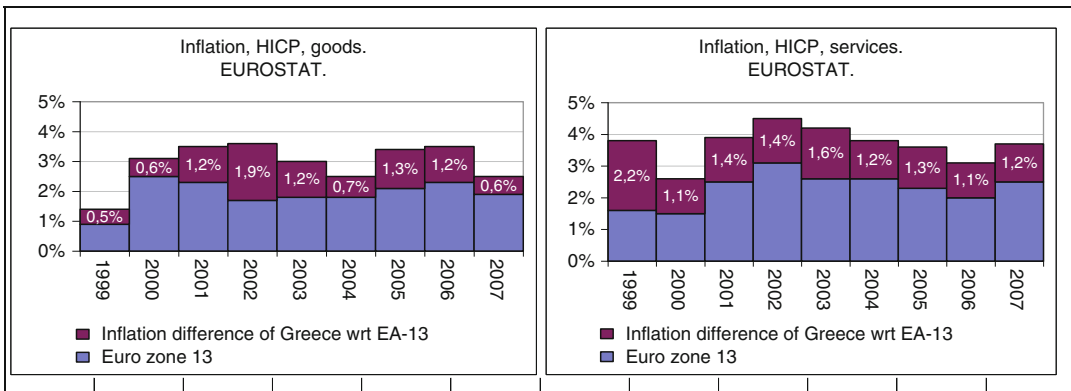
The Facets of Low Competitiveness and Inefficiency

However, at the same time, a wide range of factors persisted in contributing towards the poor

performance in certain other aspects of the Greek economy. The poor performance regarding competitiveness, to name just the most important one, is not only documented by numerous databases and surveys by international organizations and researchers but also by the persistent deficit of the current account in double-digit numbers (as a% of GDP), the persisting positive differential with the eurozone average inflation, and the unattractiveness of Greece to foreign direct investments that are practically zero (inflows minus outflows). Relatively recent research by



Efficient Market, Fig. 6 Investment as % of GDP



Efficient Market, Fig. 7 Inflation, HICP, goods & services. EUROSTAT

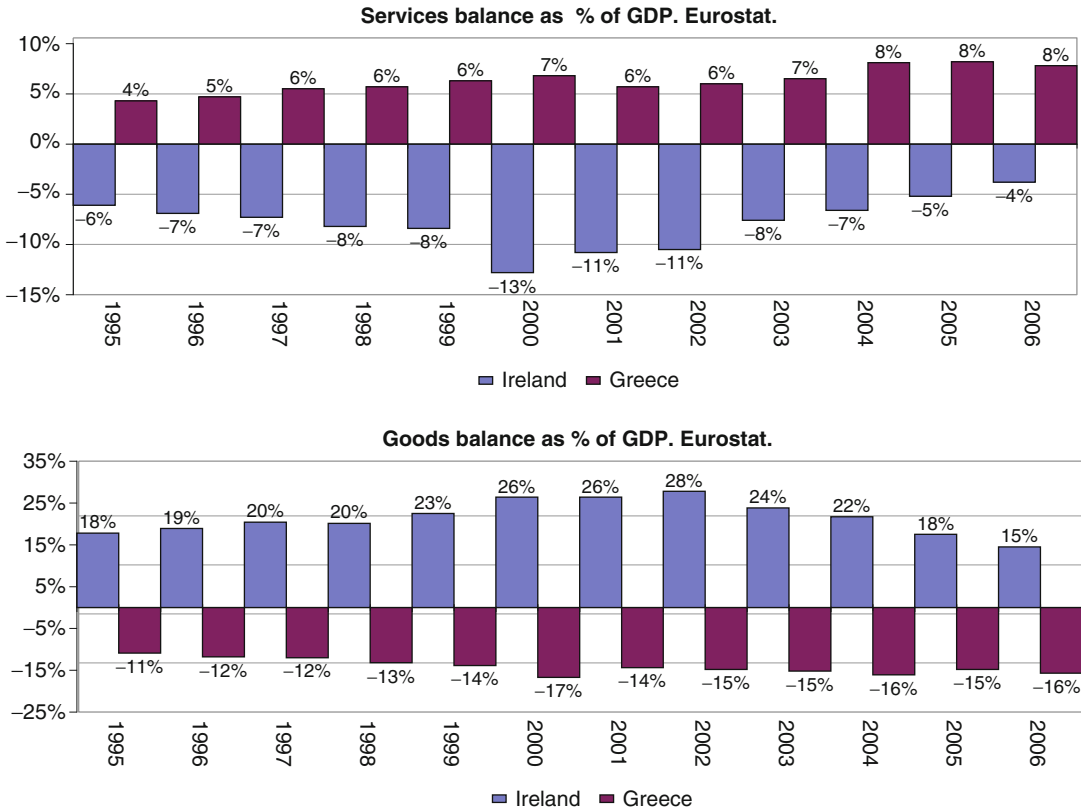
institutions like the OECD and the World Bank and a detailed presentation of numerous pieces of evidence indicate that the wide range of institutional weaknesses that prevail in Greece account, as a whole, for this dismal competitiveness performance.

The interesting part about the inflation differential of Greece with the eurozone (Fig. 7) is not that it is there, something that many would explain with the Balassa-Samuelson effect because of the rapid growth rate of the country. It is that it seems to emerge both in the goods (tradable sector) and services (non-tradable) subindexes, something that initially seems to refute the

Balassa-Samuelson argument (although to a certain extent, tourism that constitutes a significant part of services should be considered also as a “tradable service”).

An expository comparison with Ireland, where the inflation of goods is much lower than the inflation of services and that thus emerges as a textbook Balassa-Samuelson case, is most revealing.

The high inflation of Greece therefore seems to emerge as a result more of the demand increase, which is largely driven by the expansion of credit and the inflows from the EU structural funds as well as from tourism and shipping industry or



Efficient Market, Fig. 8 Services balance as % of GDP. Eurostat. Goods balance as % of GDP. Eurostat

public borrowing, which is not matched by a similar increase in the domestic supply of goods and services. And this is unlike the case of Ireland in which the surplus of the goods balance seems to finance a deficit in the services balance following again a pattern that fits well the standard predictions of the Balassa-Samuelson model.

The second piece of evidence that supports this argument is the excessive – and increasing – deficit of the goods trade balance, as a percentage of GDP (Fig. 8).

As a matter of fact, the deficit is of a magnitude that has never been seen in any country without the subsequent emergence of serious consequences. In the case of Greece, participation to the eurozone seems to have averted developments like the entrance into a spiral of high inflation and currency devaluations. As a result, the trade deficit in Greece can clearly demonstrate the existence

of a serious discrepancy between the growth of domestic demand and the increase of the domestic supply of both goods and services. It should be stressed that in the case of non-tradable services, the inflation differential is sufficient to document the discrepancy between supply and demand, but the emergence of such a differential for goods as well suggests the peculiarity of the case of Greece. Therefore, the evidence at hand would make it more appropriate to label Greece as a unique case of “quasi Balassa-Samuelson,” where exports are replaced by EU transfers and domestic credit expansion and the price level is pushed upwards both in the goods and in the services sector, which would actually be in line with the conclusions of recent research on the topic (Gibson 2007; Pelagidis and Toay 2007). The increase of the goods deficit follows as a natural consequence in this case, as increases in demand are satisfied by competitive and

available imported goods as there is no sufficient domestic supply of goods that can compete with the imports.

The third piece of evidence is the following: This persistent deterioration of the goods balance has been financed, besides from the surplus of the services account through foreign inflows in both Greek government bonds and into the stocks of Greek companies, at least until the present financial turmoil. However, it should be noted, rarely these inflows were FDIs. FDIs during the last 3 years were close to zero (\$0.9 bil. for 2006, \$2.5 bil. for 2007, and \$1.3 bil. for 2008, according to the Bank of Greece).

In fact, FDI inward flows for Greece as a percentage of GDP are very low for almost all years, something that is in line with the link between the attractiveness of the business environment and FDI (as described by authors such as Hajkova et al. 2007).

The performance of the goods balance together with the inflation differential with the eurozone for tradable goods suggests also that the cost of importing and distributing these competitive imported goods is higher compared to the eurozone. Furthermore, it suggests that the imports remain competitive in the domestic market in spite of this high cost of importing and distributing, which seems to be really damning for the competitiveness of the domestic supply of goods.

In spite of the mitigating effect of the surplus of the services balance, which is mainly driven by the performance of the shipping industry and tourism, the current account balance has remained for many years at a level (15% to GDP) that in any other country would have been associated with serious consequences. For the two sectors that contribute to the services account surplus, it should be noted that they are less affected by the regulatory environment of the Greek economy, either because they operate almost completely outside the Greek jurisdiction and administrative reality, for the case of shipping, or because they draw their competitive strength largely from the geographical attractiveness and the cultural heritage of Greece, as is the case for tourism.

These pieces of evidence manifest themselves in the compelling case for the low competitiveness of the Greek economy that is documented by a number of surveys (Fig. 9). The impressive part to note here is that a wide selection of different surveys, including those that measure governance and corruption, rank Greece in a roughly similar way even though they often use different methods that are either based on the evaluation of hard evidence, the responses to questionnaires, or a combination of both (Fig. 9).

Facets and Evidence of Institutional Inefficiency and Poor Governance

The OECD Regulation Database, the World Economic Forum competitiveness survey, the World Bank “Doing Business” and Governance Indicators, and the European Commission estimates (EU 2002, 2006), to name a few, all find that in Greece the administrative burden is also exceptionally high (Fig. 10), that regulation of markets is excessive, that government intervention limits competition as well as resource allocation and pricing decisions in crucial network industries, and that the regulation of professional and legal services (Figs. 11 and 12) is high as far as entry and price setting is concerned. At the same time, qualitative standards are excessively lax (Paterson et al. 2003) and that the business environment, as an aggregate, is unattractive.

These findings are complemented by more general statements that indicate weak institutions, poor governance (Kaufmann et al. 2005), and high levels of corruption that seem to follow as a consequence of the high administrative burden and the poor governance (Ackerman 2006).

As a matter of fact, the magnitude of the weaknesses documented by these pieces of evidence matches the size of the competitiveness deficit documented for Greece by the inflation differential with the eurozone, the current account deficit, and the low level of FDIs. It has to be added that, not surprisingly, Greece is found to be the OECD country which has the most to gain from rectifying these documented deficiencies, like product market regulation (Conway et al. 2006), in terms of

Doing Business in 2009, World Bank.	Ease of Doing Business Rank.	World Economic Forum 2008	GCI 2008-2009 rank	Transparency International	2008 Corruption Perceptions Index Country Rank	UN	Rank per capita income in US \$
No of countries	181		134		180		214
Greece /total	53%		50%		32%		19%
Singapore	1	United States	1	Denmark	1	Luxembourg	2
New Zealand	2	Switzerland	2	New Zealand	2	Norway	4
United States	3	Denmark	3	Sweden	3	Iceland	6
Hong Kong	4	Sweden	4	Singapore	4	Ireland	7
Denmark	5	Singapore	5	Finland	5	Denmark	8
UK	6	Finland	6	Switzerland	6	Switzerland	10
Ireland	7	Germany	7	Iceland	7	Sweden	13
Canada	8	Netherlands	8	Netherlands	8	Netherlands	14
Australia	9	Japan	9	Australia	9	Finland	15
Norway	10	Canada	10	Canada	10	Australia	16
Iceland	11	UK	12	Luxembourg	11	UK	17
Japan	12	Austria	14	Austria	12	United States	18
Sweden	17	Norway	15	Hong Kong	13	Austria	19
Belgium	19	France	16	Germany	14	Belgium	22
Switzerland	21	Taiwan	17	Norway	15	Canada	23
Estonia	22	Australia	18	Ireland	16	Australia & NZ	24
Korea	23	Belgium	19	UK	17	Germany	25
Mauritius	24	Iceland	20	Belgium	18	France	26
Germany	25	Ireland	22	Japan	19	Italy	32
Netherlands	26	New Zealand	24	USA	20	Japan	33
Austria	27	Luxembourg	25	Chile	23	Spain	35
France	31	Chile	28	France	24	New Zealand	37
South Africa	32	Spain	29	Uruguay	25	Hong Kong	39
Slovakia	36	China	30	Slovenia	26	Greece	40
Chile	40	Estonia	32	Spain	30	Cyprus	42
Hungary	41	Czech Rp	33	Cyprus	31	Bahrain	43
Tonga	43	Thailand	34	Portugal	32	Puerto Rico	45
Armenia	44	Kuwait	35	Dominica	33	Israel	46
Bulgaria	45	Tunisia	36	Taiwan	39	Slovenia	47
United Arab Emirates	46	Cyprus	40	South Korea	40	Portugal	48
Romania	47	Puerto Rico	41	Latvia	52	Czech Republic	55
Portugal	48	Slovenia	42	Slovakia	53	Estonia	56
Spain	49	Portugal	43	South Africa	54	Saudi Arabia	60
Luxembourg	50	Lithuania	44	Italy	55	Hungary	61
Turkey	59	Slovak Rpb	46	Seychelles	56	Slovakia	62
Italy	65	Italy	49	Greece	57	Antigua	63
Dominica	74	Turkey	63	Lithuania	58	Latvia	66
Albania	86	Brazil	64	Poland	59	Lithuania	67
Marshall Islands	93	Montenegro	65	Turkey	60	Croatia	68
Serbia	94	Kazakhstan	66	Namibia	61	Poland	69
Papua New Guinea	95	Greece	67			Russian Federation	73
Greece	96	Romania	68			Venezuela	74
Dominican Republic	97			Sudan	175		
.....				Afghanistan	176		
.....		Mauritania	131	Haiti	177	Liberia	211
Guinea-Bissau	179	Burundi	132	Iraq	178	Zimbabwe	212
Central African Republic	180	Zimbabwe	133	Myanmar	179	Congo	213
Congo, Dem. Rep.	181	Chad	134	Somalia	180	Burundi	214

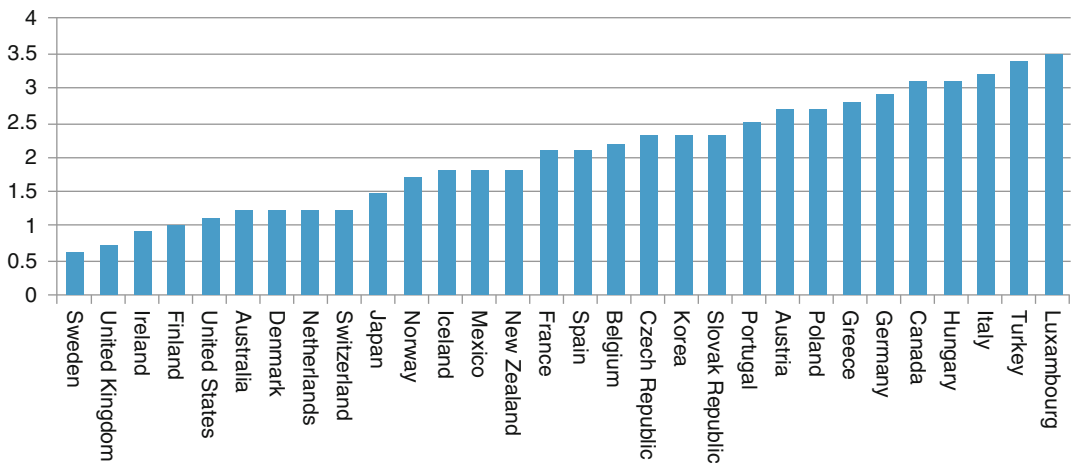
Efficient Market, Fig. 9 Competitiveness indexes

increased productivity. This performance can be labeled “dismal” not because of its absolute level, but because of the large discrepancy between the performance of the country on all these aspects

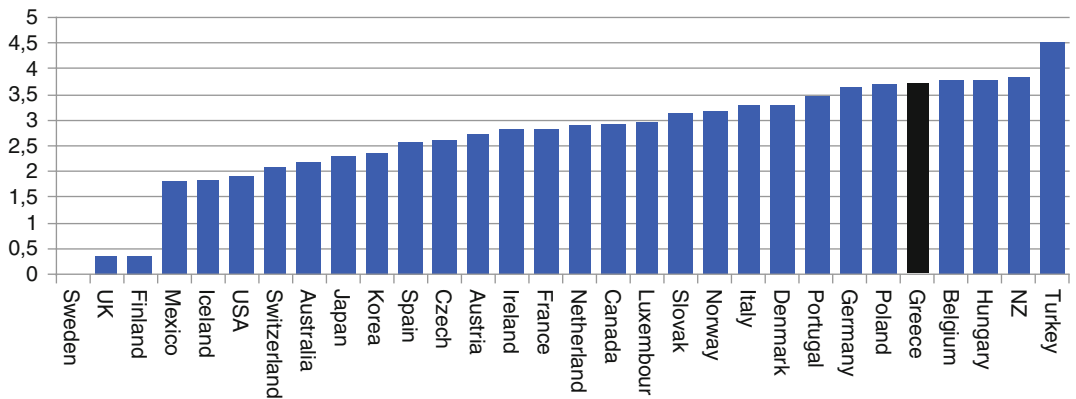
and the per capita GDP that it has achieved in the past years. In particular, following the strong performance till the 1970s and the strong performance of the past years, per capita GDP is

	AT	BL ²	CZ	DE	DK	ES	FI	FR	UK	GR	HU	IE	IT	NL	PL	PT	RE ²	SK	SI	SE	EU-25
Administrative cost share in GDP (in%) ¹	4.6	2.8	3.3	3.7	1.9	4.6	1.5	3.7	1.5	6.8	6.8	2.4	4.6	3.7	5.0	4.6	6.8	4.6	4.1	1.5	3.5
¹ Based on Kox (2005): Intra-EU differences in regulation-caused administrative burden of companies. CPB Memorandum 136. CPB, The Hague.																					
² BL combines Belgium and Luxembourg; RE combines the Baltic Member States, Malta and cyprus; EU-25 figures are GDP-weighted averages																					

Efficient Market, Fig. 10 Administrative costs by Member State



Efficient Market, Fig. 11 Regulation in professional services (Source: OECD indicator for regulation in professional services, 2007)



Efficient Market, Fig. 12 OECD indicator. Regulation of legal services, 2007 (Source: OECD. 6 = most stringent)

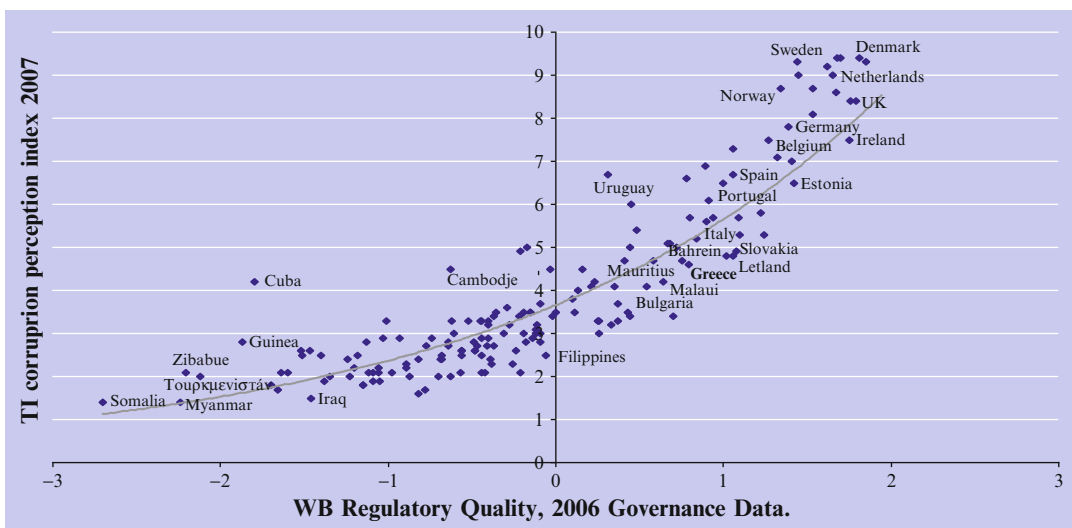
relatively close to the per capita GDP of the other OECD and EU member countries. And while Greece remains among the poorer members of these groups, it still can distance itself clearly from most other countries that do not participate in these two groups of privileged countries. On the other hand all the other performance indicators are clearly much weaker than the performance of all other OECD and EU member countries. Here Greece clearly is placed, repeatedly, in the middle of the sample of all the countries in the world, and not in the top 20% of the countries, as is the case with per capita GDP. Greece, ultimately, emerges as a country with almost first-class per capita GDP – until at least 2009 – but clearly second-class governance, institutions, business environment, and corruption (Fig. 13).

The factors that were analyzed previously and that document why Greece grew so fast in spite of these shortcomings can also reconcile the recent performance of Greece with the now extended literature, mainly of OECD Economic Department Working Papers (an indicative selection of related OECD and non-OECD related publications is OECD (2007a), Conway et al. (2006), Bassanini and Duval (2006), Nicoletti and Scarpetta (2005, 2006), Conway et al. (2005), Bassanini and Ernst (2002), Scarpetta and Tressel (2002), Scarpetta et al. (2002), Nicoletti and

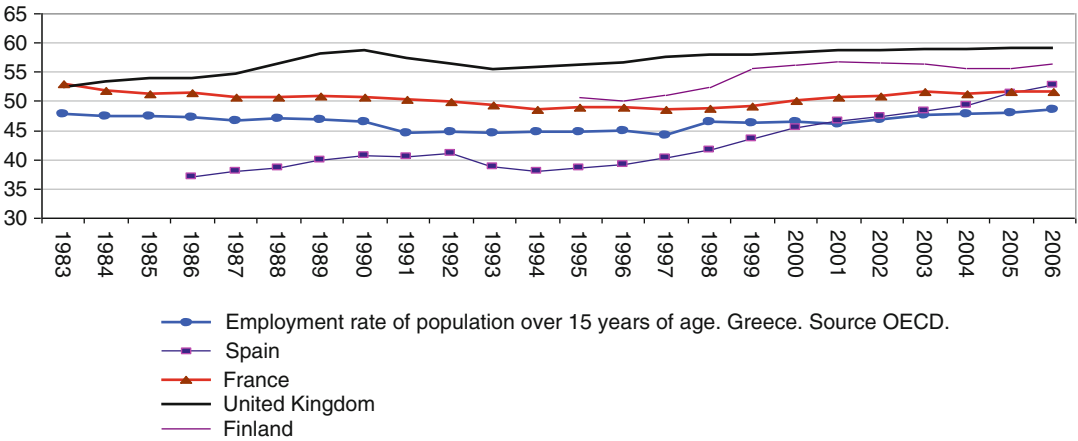
Scarpetta (2003), OECD (2003), Nicoletti et al. (2001), Conway et al. (2006)), which directly link the performance of an economy with the quality of the regulatory framework and the prevalence of competitive markets. In a similar way one can reconcile also almost all of the other weak performances of the country that range from research and innovation (Bassanini et al. 2000) to the protection of the environment, the quality of public health services and schools, and the performance of the higher education system (Bassanini and Scarpetta 2001; Mitsopoulos and Pelagidis 2007a; OECD 2007b). Even the weak performance of the judiciary which we will present below in section “[Last but Not Least: Inefficiency of Justice](#)” can be ultimately linked to this pattern (Mitsopoulos and Pelagidis 2007b; Djankov et al. 2002).

The Paradox of the Underlying “High Labor Productivity” in a Low Competitiveness and Inefficiency Context

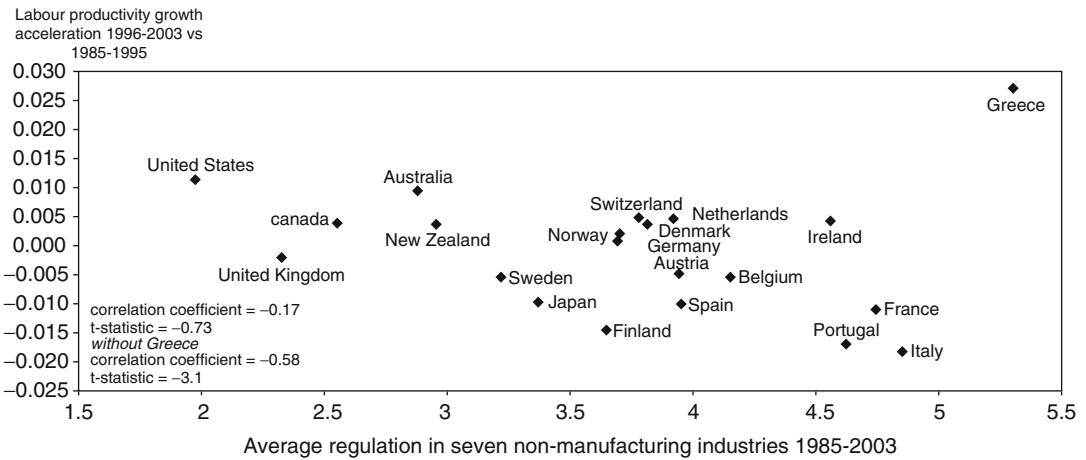
The result of the strong demand growth that is not driven by an increase in domestic supply that follows from an increase in employment (Fig. 14) directly affects the reliability of



Efficient Market, Fig. 13 Corruption and Regulation



Efficient Market, Fig. 14 Employment ratio for the population over 15 years of age

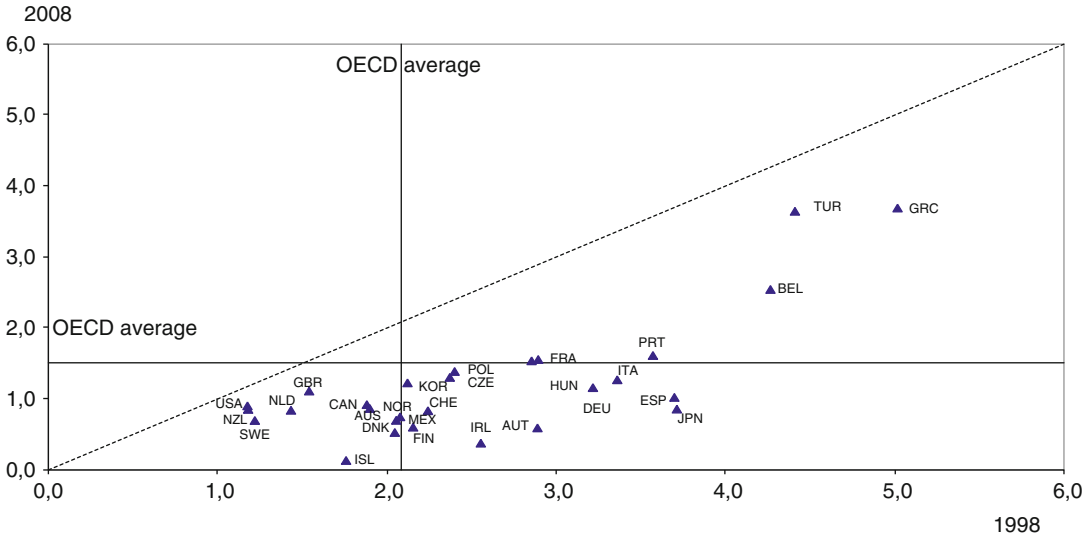


Source: OECD Productivity Database and OECD International regulation database

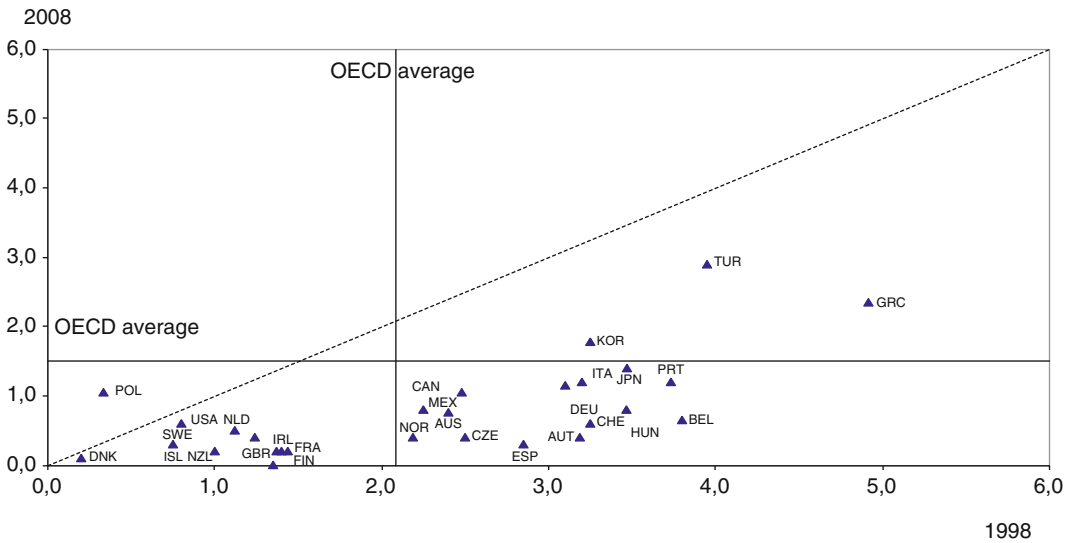
Efficient Market, Fig. 15 The scale of the indicators is 0-6 from least to most restrictive

productivity indexes that measure GDP to labor input – and that give a % of around 2.5–3% for Greece during these years – in various forms. This follows as the increase in the numerator (GDP) matches a restrained increase in the denominator (as can be seen in Fig. 14), thus measuring a large increase in the productivity per worker or per hour worked, in spite of the dismal performance of the Greek economy as measured by the rigidity index of relevant product markets (Fig. 15). It follows from the previous exposition that the use of such indicators is not correctly capturing the variety of the parameters that shape the performance of the

Greek economy during the past decade, often depicting Greece in a position that does not favor the drawing of reliable conclusions. This gives also an explanation to the puzzle of having on the one side high GDP and productivity rates and on the other side low competitiveness with twin deficits. At least to the extent, we take into account only domestic forces and not taking into account factors such as euro's overvaluation (at least to the extent that Greece's trade takes place with outside EU partners (around 50% of total)) and the asymmetric demand shocks.



Efficient Market, Fig. 16 State control. Involvement in business operation (Source: OECD. Index scale of 0–6 from least to most restrictive)

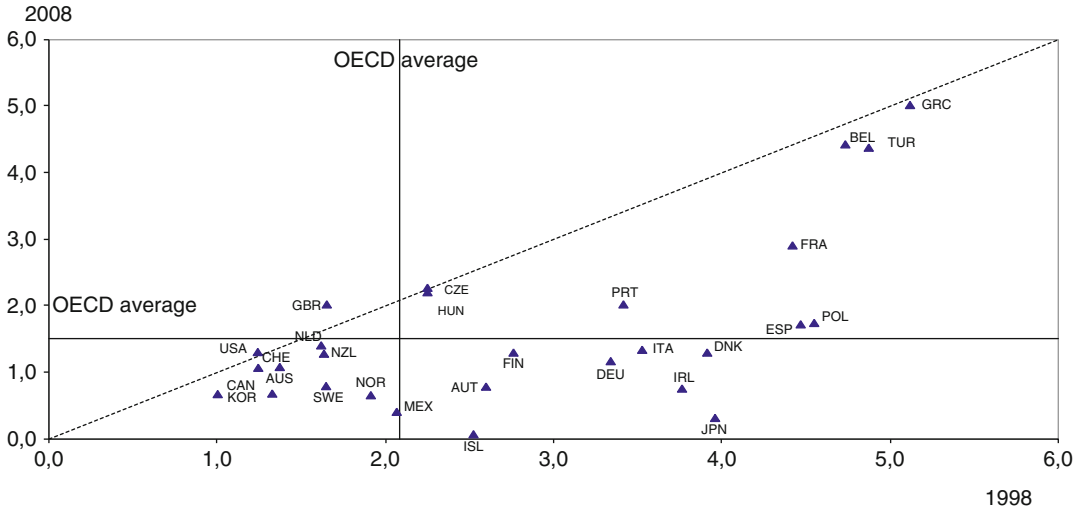


Efficient Market, Fig. 17 State control – price controls (Source: OECD. Index scale of 0–6 from least to most restrictive)

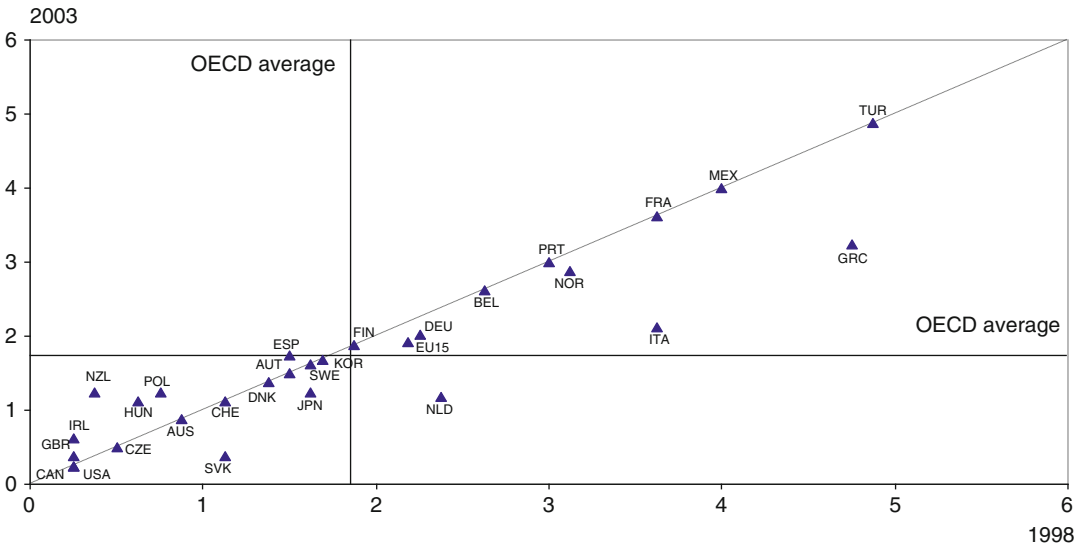
In this context, it is worth looking more on some other aspects of institutional rigidities which complement very well low competitiveness. Figure 16 summarizes product market regulation regarding state control through the involvement in business operation. Once again Greece appears to both have the most stringent

state involvement in the operation of businesses and to maintain these practices throughout the past decade.

Figures 17 and 18 concern the state involvement in business operations via price controls or the use of command and control regulation. “Command and control” includes a lot of



Efficient Market, Fig. 18 State control. Use of command and control regulation (Source: OECD. Index scale of 0–6 from least to most restrictive)



Efficient Market, Fig. 19 Employment protection legislation (However, the above data does not include the so-called shadow economy, which in the case of Greece it is estimated around 20–30 % of the official GDP. The “black

labour market,” as part of the shadow economy, includes a significantly flexible labour market that excludes employees from trade unions’ membership, social protection and insurance, working rights, etc.)

administrative mechanisms of hindrance of entrepreneurial activity/organization, in sectors such as “road and railway transports” and retail trade.

Product market rigidities are of critical importance for rigidities in the labor markets as well. Figure 19 shows Greece among OECD countries

with the highest employment protection legislation (EPL). It should be noted here that the market for nonpermanent, temporary employment in Greece is the main reason for the exceptional rigidity of the Greek labor market overall but that the market for permanent contracts is also

relatively rigid when compared with other OECD countries.

These kinds of structural institutional rigidities/inefficiencies constituted a true cost to society in the environment of a noncompetitive economy like the Greek economy. It meant and led to the exclusion of many others from the labor market, and especially the young that seek salaried labor. Under 26 years old unemployment was more than 35% and 20% for women and men correspondingly even before the crisis. This should be read as underutilization of a dynamic labor force and should not be considered solely as a major social or ethical issue.

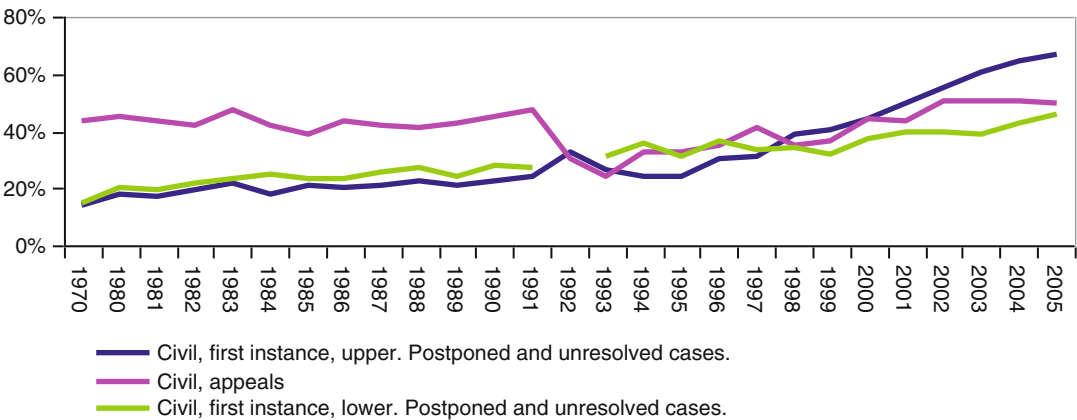
We mentioned extensive regulation of markets, high administrative costs, a business environment that is not favorable to entrepreneurship and, in the end, weak convergence and widespread corruption as drivers and cause of this low competitiveness and low efficiency, in spite of the reforms in the credit and telecommunications markets and the benefits accruing from EMU accession. Greece emerges, therefore, to benefit from certain reforms in terms of potential output because of the nature and importance of these positive developments, while it retains other, also significant in importance and magnitude, weaknesses that undermine the long-term growth potential of the country. These weaknesses are ultimately described as “rigidities and inefficiencies, weak non-independent institutions and governance,” and their proliferation was deeply built in the

equilibrium that was formed between the interest groups that accrued the rents that they secured through the regulation of markets and the inflation of the administrative costs (Pelagidis and Mitsopoulos 2009). One could also argue that the strong growth of the past years has also made the need for further reforms less pressing.

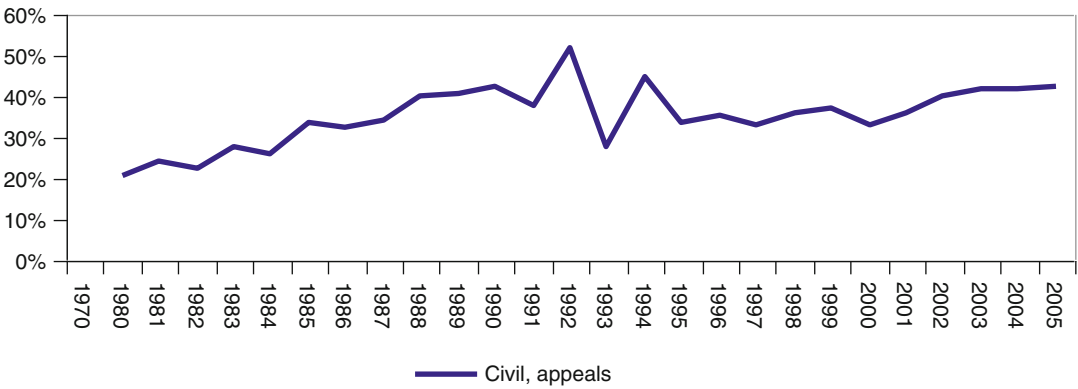
Last but Not Least: Inefficiency of Justice

The weak performance of the judiciary, which we present very briefly below, can be ultimately linked also to this pattern (Mitsopoulos and Pelagidis 2007b; Djankov et al. 2002).

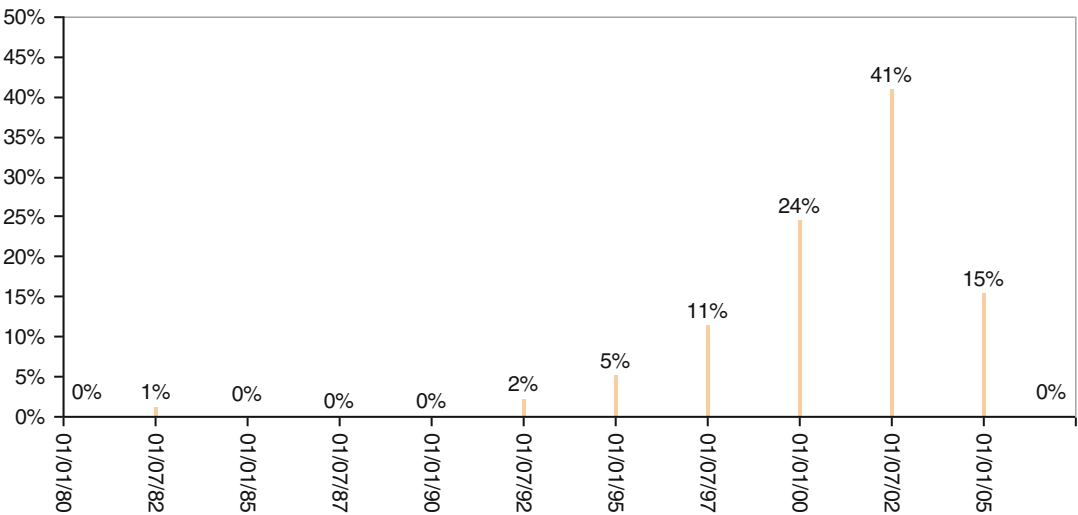
This unfortunate picture refers to cases that are quite old; therefore, the deterioration in the past years of the indirect measures of the time needed to dispose justice used in Mitsopoulos and Pelagidis (2007b) suggests that the cases that now enter the system may face even worse prospects. In particular, Fig. 20 shows that the ratio of cases that have not been heard to the sum of cases that are either pending or were introduced for the first time in the year has increased in upper civil courts from about 25% in 1997 to over 60% in 2005. The deterioration is visible, but less dramatic, in appeals courts and lower civil courts. Also, as shown in Fig. 21, for civil courts the appeal rate has increased slightly from 1997 to 2005.



Efficient Market, Fig. 20 Unresolved and postponed cases to total cases introduced



Efficient Market, Fig. 21 New introduced to appeals courts to concluded by first instance courts



Efficient Market, Fig. 22 % of cases first filed from the latter data of the range and till the latest date of the range. Cases to be retried

We also note in our sample in Mitsopoulos and Pelagidis (2010) for the Greek Supreme Court (*Areios Pagos*) some extreme delays in the process of serving justice (Fig. 22). Twenty percent of our cases have not been settled after almost 10 years since they were first filed, and even for these cases the process will be delayed for at least another year till the appeals court issues a new verdict. Sixty-five percent of the cases await a new verdict from the appeals court in spite of the fact that they were first filed between 5 and 10 years ago. On average, at the end of the year 2006, the cases tried by the *Areios Pagos*, and that were

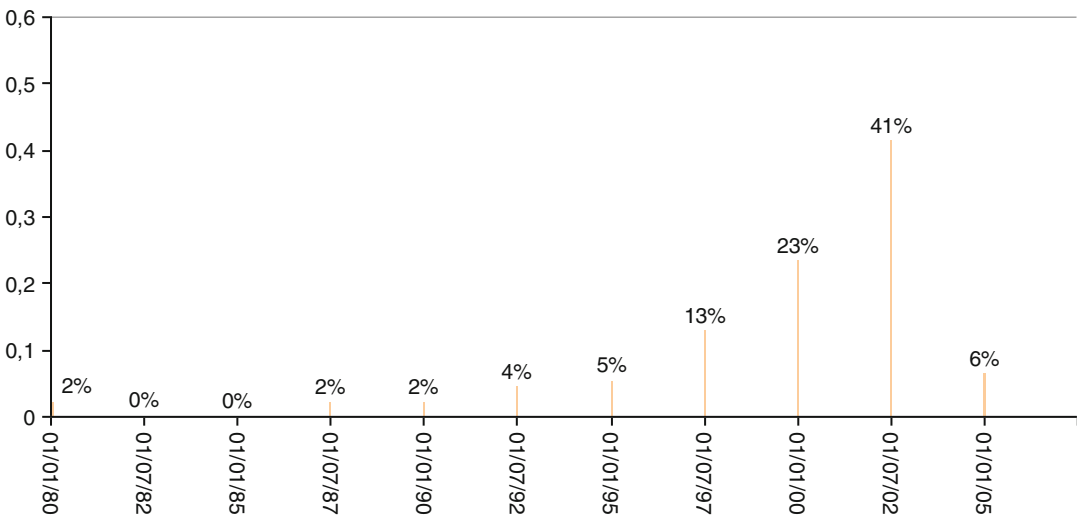
marked for a retrial, were 6.5 years old. We also find a small but not insignificant number of cases, 8% of our sample, in which we document repeated appeals to the civil *Supreme Court*. These originate among the 40% of all the cases that have been retried by the order of the *Areios Pagos*. It seems that for a number of these, the new decision of the appeals court, or the court that tries in its place, is challenged anew by one of the involved parties in front of the *Areios Pagos*. While the total number of these cases is small, about 8% of the total, they are about half of the cases that were initiated more than 12 years before 2006 and about one fourth of

the cases that are over 10 years old. This indicated that there is a small, but non-negligible, chance that a case will get entangled in a process of repeated retrials, and this process ensures that no final verdict is obtained for at least a decade. Our sample contained two cases that were judged for the third time by the *Areios Pagos* and thus are due to get tried for the fourth time by the appeals court, and one case where the second verdict of the *Areios Pagos* got challenged and was directly brought again to the court for retrial.

Such delays also apply to cases in which the Supreme Court reaffirms the decision of the previous court. Taking a random sample of 100 of these cases from the year 2006, we see that these cases, which form about 60% of the cases in the year 2006, were on average over 7.5 years old at the time the decision of the *Areios Pagos* was published. And while 70% of these cases was less than 7 years old, a significant 30% was older than 7 years with one case being first introduced almost 26 years ago and another case 33 years ago (Fig. 23). About 25% of these very old cases, which were first introduced more than 7 years ago, were cases that had been retried by the appeals court before, after a previous order to do so, but 75% of these old cases simply proceeded very slowly.

These long delays in the process of official arbitration are not caused because one of the involved parties is not satisfied with the outcome, say because the height of the compensation awarded did not meet its expectations. They are caused because the lower courts either do not follow procedures properly or because they ignore, or misinterpret, laws which is an objective measure for the low quality of the decisions that are taken by the lower courts. In itself though, the use of the appeals process for error correction, as opposed to first instance and appeals courts decisions that do not make mistakes, does not provide by itself information on the quality of the judicial system, as explained by Shavell (1993).

Measuring the quality is curtailed by the fact that we have a complete sample only from *Supreme Court* decisions that do judge only the lawfulness of the preceding court decisions, as is appropriate for a civil law country. Also, since we are not always informed about the full content of all the preceding court decisions, that is, both the appeals and first-level court, or in the case of previous retrials the content of the initial court decisions, we cannot construct measures that compare the initial with the subsequent decisions. Furthermore, the structure of civil law does not permit judges to commit errors that may be



Efficient Market, Fig. 23 % of cases first filed from the latter data of the range and till the latest date of the range. Cases reaffirmed

brought before the highest court that we may use to pinpoint a low quality of court decisions that refer to issues like the height of compensation. Therefore, we are left once again only with a measure of time that allows us to measure the time needed to serve finally justice. As in our previous work (Mitsopoulos and Pelagidis 2007b), where we employed an indirect measure for time, we find again, using this time a direct measure for the time needed to dispose justice, that in Greece justice is served often with delays that are excessive and which undermine the usefulness of the official arbitration mechanism.

The 2010–2013 “Troika Period”: Reforms for Achieving Efficiency

Below we offer a list of reforms already implemented by the Greek government in the context of the conditionality programs agreed with the troika and the creditors of the country during the period 2010–2013. The list is not exhaustive albeit representative. Reforms are leading the country towards a much more efficient economic structure.

Labor Market Reforms

- Significant reforms for the private sector from 2010 (reduction of severance payment, relaxing of layoff limits, activation of tools to reduce individual salaries, possibility to bypass general wage agreements at the firm level) have increased downward wage cost flexibility.
- Even more significant reforms tied down in MoU2 (rationalization of mediation process, reduction of minimum wage, and impact of sectorial and other agreements).
- More fine-tuning reforms rumored for 2013–2014.
- Much lower minimum wage.
- Large reduction in firing costs for high earners.
- Redefining threshold for group layoffs.
- Removal of all impediments to conclude wage agreements at the company level.
- Significant changes in the mediation process.
- A very important impact of these in the reality behind average labor cost indexes.

In the end, these reforms do establish a truly flexible and efficient labor market in Greece.

Fiscal Reforms

- Wide ranging social security reform
- Large cuts in public sector salaries and salaries of public utility companies
- Healthcare reform (online prescriptions, pricing of medicine, etc.)
- Reforms in budgeting and fiscal reporting
- Very large, across the board, tax increases (albeit with a significant inflationary impact)
- Accelerating drive against tax evasion

Structural Reforms (selectively)

- Licensing process and spatial planning reform (about 40% complete)
- Road haulage deregulation (about 50% complete)
- Cruise ship home porting regardless of flag
- Already some deregulation in professional fees and obligatory purchase of services (notary public, lawyers for small real estate transactions, company start-ups, fees of engineers, trademark submission without lawyer attendance, sale of baby milk outside pharmacies, etc.)
- Establishment of one-stop shops for company creation and gradual improvement in their operation (yet complicated process design does not work well)
- Export facilitation (abolishing obligation to register in “exporters registry,” a strategy to boost exports, link Piraeus port terminal operated by COSCO with rail network, etc.)

An Indicative List of New Taxation Since the First Memorandum 2010

- Multiple abolition of tax exemptions
- Increased of annual estate tax, up to 2% of administratively set value, and then additional estate tax through electricity bills
- Multiple VAT increases
- New personal income taxation law with higher rates, lower tax-free income
- 3 tax increases on tobacco and alcoholic beverages and VAT increase on nonalcoholic beverages

- Further increase of tax on mobile communications – highest in the EU
- 3 gas/petrol tax increases
- Multiple increases of taxes on electricity, especially for businesses – now probably highest energy prices in the EU for industry
- Introduction and increase of luxury tax
- Recurrent extraordinary tax on profitable companies and increased tax on dividends
- Introduction and increase of recurrent extraordinary tax on high personal incomes
- Tax on banks and increase of tax advancement
- Tax on TV advertisements
- Tax on violations of building permits
- Green tax
- Income tax on leased cars and private sector company cars (but not government official cars!!!)
- Introduction and then increase of tax on assumed income
- Increase in train and bus tickets
- Tax to avoid antismoking ban
- Fee to access hospitals
- Highest recurrent real estate taxes in the OECD

Since the initiation of the conditionality programs in 2010, and after two revisions and a debt haircut to private creditors in 2012, the improvement of the business environment, the remove of red tape, and the creation of a level playing field through the removal of restrictive to competition laws have been rather a second priority both for the Greek governments and for the official lenders of the former, when compared with tax increases and labor market reforms and at least for the two first years of the program implementation. While this has been changing slowly during 2013–2014, with the official lenders now demanding from the Greek government progress on these issues and while they now offer also extensive technical assistance, it must be noted that the delay in the acknowledgment of the importance of such measures may yet prove to be crucial. In sum, more reforms are needed in the domains of product and professional markets to increase thoroughly efficiency. Below one can find a few suggestions on this issue.

More Reforms for Efficiency

- Complete licensing, business park legislation, etc.
- Complete professional services deregulation
- Rationalizing the complex system of third-party payments
- Rationalizing regulation in the energy sector
- Rolling out the privatization program in a more aggressive way, emphasizing assets crucial for competitiveness like ports and not gambling
- Structural reforms that will improve the business environment and that are yet to be identified
- Emphasis at last in deregulating product and professional markets
- Last (but not least), a (meaningful) tax reform which will not include today's overtaxation

References

- Ackerman SA (ed) (2006) International handbook on the economics of corruption. Yale University Press, New Haven
- Bassanini A, Duval R (2006) Employment patterns in OECD countries. Reassessing the role of policies and institutions, vol 486, ECO WP. OCED, Paris
- Bassanini A, Ernst E (2002) Labour market institutions, product market regulation and innovation. Cross country evidence, vol 316, ECO WP. OCED, Paris
- Bassanini A, Scarpetta S (2001) Does human capital matter for growth in OECD countries? vol 282, ECO WP. OCED, Paris
- Bassanini A, Scarpetta S, Visco I (2000) Knowledge, technology and economic growth: recent evidence from OECD countries, vol 259, OECD ECO WP. OECD, Paris
- Conway P, Janod V, Nicoletti G (2005) Product market regulation in OECD countries: 1998 to 2003, vol 419, OCED ECO WP. OECD, Paris
- Conway P, de Rosa D, Nicoletti G, Steiner F (2006) Regulation, competition and productivity convergence, vol 509, OECD ECO WP. OECD, Paris
- Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A (2002) The practice of justice. World Bank Development Report 2002. Washington, D.C.
- EC (2006) Measuring administrative costs and reducing administrative burdens in the European Union. Commission Working Document COM (2006) 691 final, 14.11.2006, london
- EU (2002) Benchmarking the administration of business start-ups. European Commission Final Report, Brussels

- Gibson HD (2007) The contribution of sectoral productivity differentials to inflation in Greece. Bank of Greece, WP, vol 63
- Hajkova D, Nicoletti G, Vartia L, Yoo K-Y (2007) Taxation, business environment and FDI location in OECD countries. OECD ECO WP, vol 501. Also published in OECD economic studies no. 43/1 2007
- Kaufmann D, Kray A, Mastruzzi M (2005) Governance matters IV. The World Bank, Washington, DC
- Mitsopoulos M, Pelagidis T (2007a) Rent seeking and ex-post acceptance of reforms in higher education. J Econ Policy Reform 10(3):219–44
- Mitsopoulos M, Pelagidis T (2007b) Does staffing affect the time to serve justice in Greek courts? Int Rev Law Econ 177–192
- Mitsopoulos M, Pelagidis T (2010) Greek Appeals Courts' Quality Analysis and Performance. Eur J Law Econ 30 (1):17–39
- Nicoletti G, Scarpetta S (2003) Regulation, productivity and growth. OECD evidence, vol 347, OECD ECO WP. OECD, Paris
- Nicoletti G, Scarpetta S (2005) Product market reforms and employment in OECD countries, vol 472, OECD ECO WP. OECD, Paris
- Nicoletti G, Scarpetta S (2006) Regulation and economic performance: product market reforms and productivity in the OECD, vol 460, OECD ECO WP. OECD, Paris
- Nicoletti G, Bassanini A, Ernst E, Jean S, Santiago P, Swaim P (2001) Product and labour markets interactions in OECD countries, vol 312, OECD ECO WP. OECD, Paris
- OECD (2003) The sources of economic growth in OECD countries. OECD, Paris
- OECD (2007a) Going for growth. OECD, Paris
- OECD (2007b) Economic surveys: Greece. OECD, Paris
- Paterson I, Fink M, Ogus A (2003) Economic impact of regulation in the field of liberal professions in different member states, regulation of professional services, final report-part3, Jan 2003, study by the Institut fuer Hohere Studien, Wien for the European Commission, DG Competition
- Pelagidis T, Mitsopoulos M (2009) Vikings in Greece. Kleptocratic interest groups in a rent-seeking society. Cato J 29(3):399–416
- Pelagidis T, Toay T (2007) Expensive living: the Greek experience under the euro. Intereconomics Rev Eur Econ Policy 42(3):167–176
- Scarpetta S, Tresselt T (2002) Productivity and convergence in a panel of OECD industries. Do regulations and institutions matter? vol 342, OECD ECO WP. OECD, Paris
- Scarpetta S, Hemmings P, Tresselt T, Woo J (2002) The role of policy and institutions for productivity and firm dynamics: evidence from micro and industry data, vol 329, OECD ECO WP. OECD, Paris
- Shavell S (1993) The appeals process as a means of error correction. J Leg Stud 12(2):379–426

Eighteenth-Century Piracy

► Piracy, Old Maritime

Electoral Systems

Guido Ortona

Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università del Piemonte Orientale, Alessandria, Italy

Definition

Electoral system. An electoral system is the set of rules employed to summon a representative body on the basis of the preferences expressed by a group of electors, through a given voting procedure. The chapter compares electoral systems with reference to the election of a Chamber of a democratic Parliament.

Introduction and Overview

An *electoral system* is the set of rules employed to summon a representative body on the basis of the preferences expressed by a group of electors, through a given voting procedure. This definition applies to many settings, from parents' circles to supernational organizations; in this survey we will deal only with the election of a Chamber of a Parliament. We will also limit the discussion to the context that is usual in present-time democracies; hence we will assume (a) that all voters express the same number of votes (very often one, but there are many exceptions) and (b) that within the set of allowed alternatives, there are no constraints on the choice made by the voters. The votes are cast in an *election*, and for obvious reasons the election usually takes place in different *districts*; there are however several countries, not all of them minuscule, where there is a unique nationwide district. This is the case, for instance,

of Israel and the Netherlands. The number of the Members of Parliament (normally shortened to MPs) to be elected in each district is the *district magnitude*.

A meaningful classification of electoral systems may be obtained through the crossing of two dual dimensions, *nominality* and *plurality/proportionality*. A district may elect one MP, in which case it is *uninominal*, or more, and it is *plurinominal*; and the Seats may be assigned either *proportionally* to the valid votes obtained by each list of candidates (usually a *party*) or according to the number of votes received by each candidate, in which case the system is *majoritarian*. A uninominal election is obviously majoritarian, while a plurinominal one needs not being proportional but usually is, hence the frequent identification in the everyday political debate of uninominal elections with the use of the majority rule and of plurinominal ones with proportionality.

The electoral law plays a fundamental role in the functioning of a democracy, and it is a main determinant of the perspectives, and hence of the strategies, of both incumbent MPs and challenging candidates. No surprise, then, that the electoral law may be subject to changes that may easily be dramatic and frequent. Comprehensibly, this is particularly true in times of political and economic turbulence; this entry is being written in such a period, hence the reader is reminded that the data quoted in it are those of the last months of 2016.

An authoritative source (IDEA 2005) proposes the following classification of the electoral systems actually employed in at least one country at the time of its publication:

Plurality/Majority systems: Plurality (also labeled first-past-the-post), two-round system, alternative vote, block vote, and party block vote

Proportional representation: List proportional (open or closed), single transferable vote

Mixed systems: Parallel, mixed-member proportional

Other: Single nontransferable vote, limited vote, and Borda count

The Mixed member proportional system is the system in use in Germany; the author of this entry

suggests that it should be considered a member of the proportional representation family. More on this in section “[Proportional Systems](#).” The list is not exhaustive, more so if one considers also the systems that have been defined by the doctrine, those employed at a sub-parliamentary level, and the variants within a given system. What is remarkable is the absence, among the plurality systems, of the Condorcet system. The Condorcet system qualifies as the best of the family because if the choice is not the Condorcet winner, then there exists another one that a majority prefers to it. (The *Condorcet winner*, i.e., the alternative chosen when the Condorcet procedure is in use, is the one that is preferred by a majority in dual contests against each of the others in turn. It may not exist, but in large electoral settings this is a very rare occurrence, and it may be taken into account via a suitable tie-breaking rule.) The doctrine assumed the Condorcet system as the benchmark to be employed to assess comparatively other plurality systems, since the very beginning of the quantitative comparison of electoral systems up to now (see Merrill 1984, and Diss and Doghmi 2016). Not surprisingly, normally the best system results being the *Borda rule* (see below). The implementation of the Condorcet rule requires computing facilities not available until relatively recently, yet the increasing introduction of electronic voting devices makes it a reasonable option. No country, however, is presently considering its adoption.

In this entry we will describe the aforementioned systems in some detail (sections “[Majoritarian Systems](#),” “[Proportional Systems](#),” and “[Mixed and Other Systems](#)”), and then we will move to the problem of the choice of the electoral system (sections “[Majority or Proportionality? Theory](#),” “[Majority or Proportionality? Evidence](#),” and “[Electoral Reforms and Public Opinion](#)”).

Majority, Proportionality and History

The most relevant difference among electoral systems is the fundamental principle upon which they are based, that is whether they are *majoritarian* or *proportional*. Before proceeding, it is necessary to

make it clear a very basic point, which is not always duly considered in the literature, let alone in the political debate. A majoritarian system assumes that the citizens of a particular constituency have common interests and/or opinions and that they are called to choose a delegate to represent at the best their interests and/or opinions, in the spirit of the *theorem of Condorcet*. (The theorem states that the probability of making the best choice increases with the number of deciding subjects, provided – inter alia – that there is a choice that is actually the best for each of the subjects.) A proportional system assumes instead that the community of interests does not concern a given location but a given *social group* throughout the territory. At the district level, the communality of interests may not exist, and if it exists, it is inferior to other, nonlocal communalities. It follows that the representation must be divided (proportionally) among the various groups, in order to reproduce the articulations of the society in a manageable scale. This is why Parliaments with an ancient origin are typically majoritarian, like those of the UK and of the USA, while modern ones are typically proportional (for a discussion of this point based on historical evidence, see Ahmed 2013). Also, this is why the majority of present-time scholars favor proportionality (see Bowler and others 2005).

Worldwide, in 2016, 85 Parliaments were proportional, 62 were majoritarian, and 34 were mixed (including Germany); the use of other systems was not uncommon, as they summed to 69, divided among nine different systems (data from the ACE Electoral Knowledge Network, <http://aceproject.org/>). Proportional systems are most largely employed in Europe, where 35 Parliaments (including Germany, see below) out of 46 belonged to the family.

Majoritarian Systems

The simpler and most used majoritarian system is uninominal *plurality*, also labeled *first-past-the-post* (FPTP); it is adopted in the UK, USA, India, and in many other countries of British tradition. There are as many districts as MPs to be

elected, and in each district the candidate who obtains a plurality of valid votes is elected. The *two-round system* is adopted in France and in several countries formerly in the French empire. The districts are uninominal, but if no candidate obtains a *majority*, there will be a second round involving a subset of the candidates (in France those who got at least 12.5% of valid votes, or the first two if none exceeded this threshold). The other majoritarian systems are much less in use. With the *alternative vote* (also known as *instant-runoff vote* or *transferable vote*), the districts are uninominal too. The voter is requested to rank all the candidates according to her/his preferences or some of them. In the first round only the first preferences are considered, and a candidate is elected if she/he obtains a majority. If none gets it, the candidate with less votes is excluded, and each of her/his votes is transferred to the second preference of the voter. The procedure is iterated until the majority arises. The system is employed in Australia and in other countries in Oceania; in 2011 a referendum rejected its adoption in the United Kingdom. *Block vote* and its variant *party block vote* are majoritarian but plurinominal. In the first version each voter may vote for several candidates, and those with more votes are elected; in the second each voter votes for a party, and the party with more votes wins all the seats of the district. The first variant is in use in Lebanon and Mauritius, but the most extreme application occurs in Monaco, where 16 MPs out of 24 are elected with this system, and each voter may cast up to 24 votes. The second variant is in use in Cameroon and Chad.

Proportional Systems

A *pure proportional system* (one national district, no seat assignment threshold) is in force only in 10 countries, that is Colombia (Senate only), Fiji, Israel, Mozambique, Namibia, the Netherlands, Paraguay (Senate only), Serbia, Slovakia, and South Africa. In any case, the seats are assigned to each party according to the share of valid votes obtained. With proportionality there is a problem with the rounding of the results, that

may be serious if the district magnitude is small. Most countries employ the *D'Hondt rule*, a member of the family of *highest average methods*. The number of valid votes of each party in each district are divided by successive integers, starting with 1, and the M largest ratios, where M is the district magnitude, individuate the elected candidates. There are several variants; the one most in use is the *Sainte-Laguë rule*, where only odd integers are considered, what favors the smaller parties. *Largest remainder methods* are much less employed: their principle is that the remainders determine the seats to be assigned to each party after that each one obtained the number of seats given by its integer share of votes. The party list may be either *closed*, when the voter may only choose the party, or (partially or totally) *open*, if she/he may also indicate one or more preferences within the list.

The real proportionality depends crucially from the *district magnitude*. Lijphart suggested that the following formula:

$$T = 75 / (M + 1)$$

where T is the implicit percent threshold and M is the district magnitude may roughly indicate the implicit threshold that the magnitude introduces. For instance, a district magnitude of 9 would produce (roughly) the same results of a district with a threshold for participating to the assignment of seats of 7.5%.

There is a commonsense wisdom that a trade-off exists between *proportionality* and *governability*: a not-constrained proportional vote would produce a plethora of parties, what would make it difficult the summoning of a stable and effective government. This assumption has been criticized by several authors, both on empirical and theoretical foundations, and actually it does not enjoy a sound theoretical basis (see f.i. Migheli et al. 2014). The reason is that a majoritarian system tends to produce few large parties, usually two, that include interests and opinions that may be highly different if not contrasting, in the logic of the so-called *Law of Duverger*; what may induce a bargaining *within* a party that may be as engaging as that *among*

parties in proportionality. (The Law of Duverger is not a law but an empirical regularity, not very often respected in full, according to which plurality produces a two-party system.) Nevertheless most countries adopt some correction to pure proportionality; these corrections are of three types, a *participation threshold*, by far the most common, a *majority prize*, and a *reduced district magnitude*.

Participation thresholds (that establish that a party will not get seats unless its share of valid votes is greater than the threshold) range typically around 3–5%, with Liechtenstein (8%) and Turkey (10%) outlying. As stated above, a threshold is in use in most proportional countries.

The majority prize is present in Greece and Italy (and in san Marino). The party (or the coalition of parties) that obtain a given share of votes obtains additional seats. There are some historical precedents, like the *Law Acerbo* (1924, from the name of its author), that opened the way to the electoral triumph of Mussolini and the so-called *Loi Scélérate* in France and *Law Truffa* in Italy (“scélérate” meaning wicked and “truffa” meaning fraud), respectively, in 1951 and 1953. Both the historical precedents (and the names they were popularly labeled in the last two cases), and the inability to resolve a political stalemate in the two countries concerned at present, justify clearly enough the limited consensus that this system enjoys among the scholars. In Spain the districts are usually small, what, as we saw, reduces the proportionality; again the results in terms of governability are apparently poor.

The *single transferable vote* aims at preserving the vote-for-a-person feature of majoritarian systems while at the same time assigning the seats proportionally. This is why it is highly appreciated by the English and American literature; however, by its very nature, the system requires small districts, what makes its proportionality limited. It is in use only in Australia (Senate), Ireland, and Malta. Omitting minor technicalities, its typical structure is the following. Suppose a district magnitude of M , say, 5, and N valid votes, say, 1,200. Each voter must rank all the candidates according to her/his order of preference. All the candidates that are ranked first by at least 201 voters are

elected; this is because at most 5 candidates may obtain so many first preferences (and in general $1 + M/(N + 1)$ first preferences). Suppose now that a candidate obtains 301 first preferences. 201 are “used” to win her/his seat, while 100 (chosen at random) are transferred (whence the name of the system) to the *second* preference, and the procedure is iterated until 5 MPs are summoned.

Mixed and Other Systems

A *parallel system* is in use when some MPs are elected with a system and the others with another one (or possibly more); the two systems are normally a majoritarian and a proportional one. A parallel system is supposed to guarantee a representation, albeit with a reduced share, to all the parties whose share of votes exceeds the established threshold, while at the same time producing an enhanced governability thanks to the majoritarian principle. Several countries adopt a parallel system, including Russia, Mexico, Hungary, the Philippines, Thailand, Venezuela, Japan, and South Korea. Instead, a *mixed-member proportional system* maintains the proportionality while requesting also a uninominal vote. The most relevant example is Germany; voters vote for a candidate in a uninominal majoritarian district and at the same time for a party in a large proportional one. The seats to be assigned to each party are determined by its share in the proportional vote, but the candidates obtaining a plurality in the uninominal districts are elected anyway. If a party obtains, say, 200 seats in the proportional vote but it wins in 205 uninominal districts, the total number of MPs is increased correspondingly. (In Germany a 5% national threshold is in force, but this threshold is not requested for a party that wins in at least 3 uninominal contests.) As a result, the system is not that different from an open list proportional one. New Zealand and Bolivia also employ it.

The *single nontransferable vote* is the application of the majority rule to plurinominal districts. Each voter has one vote, and the M candidates with more votes are elected, M being the district magnitude. It is in use in some Asian countries,

among which Indonesia (Senate). The *limited vote system* is analogous, but the voter may express more than one vote, albeit less than the district magnitude; it is adopted in Spain for the election of the Senate. Finally, the *Borda count*, despite the appealing theoretical properties of the system, is employed only in Slovenia, and for two seats only, and in Nauru. (Each voter ranks the candidates, giving 1 point to the first, 2 to the second, etc.; the points are summed, and the M candidates with less points are elected, M being again the district magnitude.)

Majority or Proportionality? Theory

The political doctrine made a lot of efforts to individuate the advantages and the flaws of the different electoral methods. We will consider only the comparison of the two grand families, that of the proportional systems and that of majoritarian ones. Here is a (non-exhaustive) list. For a more detailed discussion of the points listed below, see IDEA (2005), and the sub-site *Electoral Systems* of the site of ACE (Electoral Knowledge Network) or, obviously, the (enormous) scientific literature.

Majority/Plurality, Advantages

- (a) It tends to produce a one-party government, what enhances both stability and governability.
- (b) Symmetrically, it produces a strong and realistic opposition.
- (c) It encourages the parties to represent broad arrays of interests.
- (d) It tends to exclude extremist parties from the Parliament.
- (e) It forces the parties to choose credible and respected candidates and allows voters to choose individuals instead of parties.
- (f) It opens the way to independent candidates.
- (g) It is simple to understand.
- (h) It enhances the accountability of the government, as the responsibility for unpopular choices cannot be easily concealed.
- (i) It promotes the representation of the local interests.

Majority/Plurality, Flaws

- (a) Many voters will not be represented.
- (b) It may assign all the power to a party that may well have a limited consensus, what could create serious problems especially in divided societies.
- (c) It excludes small parties, and hence, easily, minority groups from obtaining a representation.
- (d) It spoils many votes, possibly even the majority of them.
- (e) As typically there will be two candidates, both close to the center of the political spectrum, many voters will not find a suitable candidate, mostly if the spectrum is large.

In addition, some of the advantages may actually be considered flaws.

- (f) A one-party government may exclude the possibility of the establishment of an effective “social contract” among contrasting groups of interest.
- (g) An extremist party is possibly more dangerous if it is excluded from the Parliament.
- (h) A candidate may be locally strong not due to her/his moral and intellectual standing but to her/his link with a lobby or a powerful pressure group (or person).
- (i) There will be a limited competition within the large partitions of the electorate, usually “left” and “right,” because each partition will candidate just one person in each district, what can reduce the quality of the candidates.
- (j) A perceived excess or responsibility may induce inaction.
- (k) Inaction may also be the consequence of the necessity of composing contrasting interests in a party forced to look for the sustain of different social groups (as suggested above in section “[Proportional Systems](#)”)
- (l) A proportional system adopting the D’Hondt rule, or even better the Sainte-Laguë rule, may propitiate independent candidacies not less than majority.

Proportionality, Advantages

- (a) It is the “natural” system, because it reproduces the overall society in a manageable

assembly (see above, section “[Majority, Proportionality and History](#)”).

- (b) It creates homogenous parties, thus avoiding the problems referred to above under letter *k*.
- (c) It puts the political transactions more in the open, as it will be more *among* parties than *within* parties.
- (d) It propitiates the representation of minorities.
- (e) It offers the voter a broader choice, thus enhancing the competition among parties.
- (f) It allows more stability and less traumatic changes, as the governments will typically be made by coalitions.
- (g) It strengthens the link between citizens and politicians, as it offers the voters a larger choice.

Proportionality, Flaws

- (a) Coalition governments may induce a stalemate in the activity of the government.
- (b) A too fragmented system of parties can make the political bargaining cumbersome.
- (c) Small centrist parties will be given an excess of power, what reduces the actual proportionality of the system.

Again, some advantages may actually turn into pitfalls.

- (d) The stability may correspond to an excessive entrenchment in power of large centrist parties.
- (e) Extremist parties may enjoy a undeserved status.

The preceding summary indicates that the typical features of the two meta-systems may easily be either an advantage or a flaw, according to the circumstances, what, arguably, makes it difficult to assess their nature on a general basis. Consequently, the support for either system is usually lexicographic. Those who prefer majority claim that the most relevant advantage of the system, i.e., the enhanced governability given by a reduction of the deciding subjects, is sufficiently important to make it preferable despite its flaws, while those who favor proportionality claim the same with reference to representativeness.

Majority or Proportionality? Evidence

A sound quantitative comparative assessment is made difficult, if not impossible, by the lack of sufficient data. The econometric literature is vast, yet its results are contradictory, as they rest too strongly on the choice of data and on unavoidable simplifying assumptions. However, on a less sophisticated basis data offer only a limited support to the basic argument in favor of majority. Already a couple of decades ago, some authors, like Lijphart (1999) and Farrell (1997), noticed that there is no convincing evidence of the superiority of plurality in term of governability. In Europe, at the beginning of 2017, the two large nonproportional countries (UK and France) apparently faced serious problem of governability, and this is also the case of the three less proportional countries among the proportional ones, that is Spain, Italy, and Greece. Actually, two main tenets in favor of the argued greater governability of majority/plurality systems are not confirmed by facts. The first is the blackmailing power of small centrist parties: in the two more proportional countries of Europe, Italy (1948–1993) and the Netherlands (since 1946), the possibility for a small party to blackmail a majority through the menace of leaving has been a very rare occurrence. The likely reason is that a power rent induces the birth of further small centrist parties, until the rent is dissipated, as first suggested by McGann, Enschedé, and Moran in 2009 and as confirmed by data.

As a result, quite a consensus has been reached among scholars that proportionality is preferable, as we noted in section “[Majority, Proportionality and History](#)” above, yet this consensus is far from being unanimous, and the debate is far from been conclusive on what proportional system should be deemed the best one. Probably a plurality of scholars would suggest a list proportional system with a threshold. (A recent contribution pointing to this result is Raabe and Linhart 2017; albeit, as we noticed, cross-country comparisons are risky, the result is strengthened by having the authors taken into account a very large set of data, 590 elections in 57 countries.) As a general conclusion, however, it is very likely that no system may be assumed to

be the best one for all the existing democracies, because the specific features play a very relevant role. Probably there are systems which are out of date or inferior to others in many if not most cases, yet Katz (1997, p. 308) has surely a point when he claims that what is the best electoral system depends on “who you are, where you are, and where you want to go.”

Electoral Reforms and Public Opinion

Broadly speaking, an electoral reform may be either the result of an attempt to make the system closer to the expectations of the citizens or on the contrary of an attempt of the political élite in power to escape a greater control by the citizens, be it to defend some privileges or to enhance the governability. For instance, the first is the case of New Zealand, that in 1994 moved from plurality, as in the UK, to mixed-member proportionality, as in Germany; the second is that of Italy, where two successive electoral laws, in 2014 and in 2016, have been rejected by the Constitutional Court. Dissatisfaction with the political establishment makes it often appealing for the public opinion the proposal of a change in the electoral system. Survey data show that in the UK a majority would prefer to change to proportionality, but in Italy in 1993 a referendum introduced a parallel system, with a larger share for plurality, in lieu of the nearly perfect proportionality in force. (The proportionality was reintroduced in 2005.) Not surprisingly, the electoral system is more likely reformed in times of economic or social change. A source (Renwick 2011) lists 47 modifications of the electoral law across Western Europe from 1945 to 2008; only 5 occurred between 1960 and 1980.

It is difficult to disentangle the satisfaction (or the dissatisfaction) with overall politics and politicians from that with the electoral system, not to speak of the low number of observations available; hence the data on the attitude of the citizens are inconclusive. Yet, if we admit that the turnout may be a proxy for the relative approval of the citizens for their electoral system, there are hints that proportionality is more appreciated than plurality. This is what is suggested by the last

elections (at the end of 2016) in comparable countries of Europe. The turnout was 75% in Italy in 2013 and in the Netherlands in 2012, 71.5% in Germany in 2013, 66% in the UK in 2015, and 57% in France in 2012. The first three countries are proportional; the last two are not. That voter turnout is higher in proportional elections is also claimed by IDEA (2016, p. 36), as an overall result of the literature. Among others, Dalton (2008) confirms this guess at the textbook level. In addition, Renwick (quoted above) reports that out of the 84 changes of the electoral law made in Europe (including non-Western countries) between 1945 and 2008, 37 aimed at increasing the proportionality and 23 at reducing it (the remaining ones did not affect this side of the electoral rule).

References

- Ahmed A (2013) *Democracy and the politics of electoral systems choice*. Cambridge University Press, New York
- Bowler S, Farrell DM, Pettit RT (2005) Expert opinion on electoral systems: so which electoral system is “best”? *J Elections Public Opin Parties* 15:3–19
- Dalton RJ (2008) *Citizen politics*. CQ Press, Washington, DC
- Diss M, Doghmi A (2016) Multi-winner scoring election methods: Condorcet consistency and paradoxes. *Public Choice* 169:97–116
- Farrell DM (1997) *Comparing electoral systems*. Macmillan, Basingstoke
- IDEA (International Institute for Democracy and Electoral Assistance) (2005) *Electoral system design: the new international IDEA handbook*. IDEA, Stockholm
- IDEA (International Institute for Democracy and Electoral Assistance) (2016) *Voter turnout trends around the world*. IDEA, Stockholm
- Katz RS (1997) *Democracy and elections*. Oxford University Press, New York
- Lijphart A (1999) *Patterns of democracy. Government forms and performance in thirty-six countries*. Yale University Press, New Haven
- McGann A, Enschedé J, Moran T (2009) The surprisingly majoritarian nature of proportional democracy. Paper presented in the annual meeting of the APSA, Toronto
- Merrill S III (1984) A comparison of efficiency of multi-candidate electoral systems. *Am J Polit Sci* 28:23–48
- Migheli M, Ortona G, Ponzano F (2014) Competition among parties and power: an empirical analysis. *Ann Oper Res* 215:201–214
- Raabe J, Linhart E (2017) Which electoral systems succeed at providing proportionality and concentration? Promising designs and risky tools. *Eur Polit Sci Rev*. Published on line Jan, 17
- Renwick A (2011) Electoral reform in Europe since 1945. *West Eur Polit* 34:456–477

Further Reading

- Farrell DM (2011) *Electoral systems: a comparative introduction*, 2nd edn. Palgrave Macmillan, Houndmills, Basingstoke
- Gallagher M (ed) (2005) *The politics of electoral systems*. Oxford University Press, Oxford, UK
- Norris P (2004) *Electoral engineering*. Cambridge University Press, Cambridge, UK

Emissions Trading

Edwin Woerdman

Faculty of Law, Department of Law and Economics, University of Groningen, Groningen, The Netherlands

Abstract

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights. The current international dissemination and intended linking of emissions trading schemes underlines the growing relevance of this instrument. There are three basic design variants of emissions trading: cap-and-trade (allowance trading), performance standard rate trading (credit trading), and project-based credit trading (such as domestic offsets, JI, and the CDM). These design variants are analyzed in terms of effectiveness, efficiency, and acceptance. It is also explained why emissions trading schemes may become inefficient hybrids of such design variants.

Synonyms

[Allowance trading](#); [Cap-and-trade](#); [Credit trading](#); [Permit trading](#); [Tradable emission rights](#); [Transferable discharge permits](#)

Definition

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights.

Introduction

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights. The term “emissions trading” is an umbrella concept for several design variants that differ substantially both in theory and in practice.

In one system, trade in emissions is carried out by firms subject to an absolute emissions cap, which is referred to as “cap-and-trade” or “allowance trading.” Examples are the US sulfur dioxide (US SO₂) emissions trading program, the European Union Emissions Trading Scheme (EU ETS), and the California Cap-and-Trade Program to cost-effectively reduce greenhouse gases, such as carbon dioxide (CO₂), as well as some of the Chinese CO₂ emissions trading pilots, including the one in the province of Guangdong.

The other system is based on trade in emissions carried out by firms subject to a relative emissions standard, which is referred to as “performance standard rate trading” or sometimes “credit trading.” Think of the early emissions trading schemes in the USA to flexibly maintain air quality standards, the (recently terminated) nitrogen oxide (NO_x) emissions trading scheme in the Netherlands, or the CO₂ emissions trading pilot in the Chinese city of Shenzhen.

Again another system is project-based credit trading, where an investor receives credits for achieved emission reductions at a domestic or foreign host. These emission reductions are measured from a baseline that estimates future emissions at a project location if the project had not taken place. Think of domestic offsets or the international projects of Joint Implementation (JI) or the Clean Development Mechanism (CDM) under the Kyoto Protocol on climate change.

This article is organized as follows. Section “[Emissions Trading: Theory and Practice](#)” discusses the concept and international dissemination of emissions trading. Section “[Emissions Trading Variants](#)” describes the aforementioned three basic emissions trading variants in more detail. Section “[Analysis of Emissions Trading Variants](#)” analyzes these emissions trading variants in terms of effectiveness, efficiency, and acceptance. Section “[Emissions Trading Hybrids: the Case of the EU ETS](#)” focuses on hybrids of the emissions trading variants by taking the EU ETS as an example. A summary is presented in section “[Summary](#)”.

Emissions Trading: Theory and Practice

In the previous century, law and economics scholar Ronald Coase (1960) postulated that trading emissions would improve the cost-effectiveness of environmental regulation. The economic concept of emissions trading was developed further by John Dales (1968). To explain this concept in an easy way, one could say that emissions trading resembles a waterbed. Suppose you would like to raise the water level at the head of a waterbed, then you would have to push down the water level at the foot of the bed. This is also how emissions trading works. One company can buy emission rights and may emit more, but the company selling the emission rights first has to reduce its emissions by an equal amount in order to make these rights available for sale. As a result the government can be certain, provided there is adequate monitoring and enforcement, that the emissions of all the companies together will not exceed the number of emission rights allocated in the emissions trading scheme. As a consequence, emissions trading is an effective tool for achieving emissions targets. Emissions trading is also efficient in the sense that companies can look for the cheapest way to fulfill their emission reductions obligations. The specific design of an emissions trading scheme determines its environmental effectiveness, economic efficiency, and political acceptability.

Emissions trading has gone truly global. Emissions trading schemes have emerged in North America, including the Regional Greenhouse Gas Initiative (RGGI) and the Western Climate Initiative (WCI) in the USA as well as the Québec Cap-and-Trade System in Canada. In Europe, there is the EU Emissions Trading Scheme (EU ETS) for greenhouse gases as well as the Swiss Emissions Trading Scheme. Oceania holds Australia's Carbon Pricing Mechanism (AUS CPM) and the New Zealand Emissions Trading Scheme (NZ ETS). Asia has the Tokyo Cap-and-Trade Program, seven emissions trading pilots carried out in China (for instance, in Beijing and Shanghai), a forthcoming Korea Emissions Trading Scheme, as well as a pilot in Kazakhstan. Moreover, Turkey, Ukraine, and the Russian Federation are considering the adoption of emissions trading schemes, as well as countries such as Thailand and Mexico (for an overview, see, for instance, ICAP 2014).

An interesting, recent development is the intended linking of some of these emissions trading schemes (Weishaar 2014). In North America, the California Cap-and-Trade Program has been linked to the Québec Cap-and-Trade System in 2014. The EU and Australia agreed on a pathway for linking the EU ETS and the AUS CPM in 2018 (although the latter scheme has been under fire in domestic Australian politics). The EU is also negotiating with Switzerland on linking the EU ETS with the Swiss ETS.

Emissions Trading Variants

There are three basic design variants of emissions trading:

- Cap-and-trade
- Performance standard rate trading
- Project-based credit trading

Cap-and-Trade

The cap-and-trade system (or allowance trading) imposes a cap on the annual emissions of a group of companies for a number of years. The emission rights are allocated to established companies for

the entire period either for free or through annual sale by auction (a combination is also possible). Newcomers and companies seeking to expand must purchase rights from established companies or from a government reserve, while a company closing down a plant can sell its emission rights.

Some examples are the US sulfur dioxide (US SO₂) emissions trading program initiated in 1995, the European Union Emissions Trading Scheme (EU ETS) for greenhouse gases set up in 2005, and some of the Chinese CO₂ emissions trading pilots, for instance, in the province of Guangdong, erected in 2013.

To introduce and start a cap-and-trade scheme, the legislator has to put in place the following design elements (Tietenberg 1980; Nentjes et al. 2002):

- Set a cap on total emissions per year for a group of emission sources, in advance, and for a range of successive years.
- Create allowances, entitling emissions equal to the total emissions cap.
- Distribute the allowances among group members, either by auctioning the allowances at predetermined dates or by handing out the allowances free of charge.
- Allowances can be traded freely.
- Monitor emissions per source and track the allowances.
- Check compliance over a past budget period (e.g., a book year) by comparing monitored emissions with the number of allowances handed over at the end of the period.
- Penalize noncompliance if emissions exceed the allowances, with a fine per lacking allowance that is a multiple of the market price of an allowance.

Performance Standard Rate Trading

Performance standard rate trading (or credit trading) is different from cap-and-trade. A system of performance standard rate trading is based (not on an emissions cap but) on a mandatory emissions standard adopted for a group of companies. The emissions standard dictates permitted emissions per unit of energy consumption or per unit of added value. In this system emission reduction

credits can be earned by emitting less than what is prescribed by the emissions standard. These credits can then be sold to companies that can use them to compensate their emissions in excess of the emissions standard which applies to them.

If the economy grows, the supply of credits also increases because companies do not operate under an absolute emission ceiling but have to observe a relative emissions standard. An energy-intensive company that expands production, or a newcomer entering the industry, therefore has a right to new emissions, as long as it obeys the emissions standard. This means that absolute emissions will grow. To prevent that from happening, the emissions standard can be strengthened (Deweese 2001; Weishaar 2007).

This system has been developed since the 1970s in the US Environmental Protection Agency (EPA) emissions trading program. A system of tradable nitrogen oxide (NO_x) emission reduction credits has been in place for energy-intensive companies in the Netherlands between 2005 and 2013. A system of tradable reduction credits can also be found in the CO₂ emissions trading pilot in the Chinese city of Shenzhen, launched in 2013.

Project-Based Credit Trading

Project-based emissions trading, like Joint Implementation (JI) and Clean Development Mechanism (CDM) projects under the Kyoto Protocol, can be seen as a variant of credit trading. JI relates to emission reduction projects in East European countries, whereas the CDM refers to such projects in developing countries. Both credit trading and emission reduction projects allow for the transfer of credits, but projects usually require pre-approval to check the environmental integrity of the project baseline. This is not necessary under credit trading where the baseline is existing environmental policy (like an emissions standard), so that compliance can be checked by the end of the year. Moreover, the firm which funds the reductions under credit trading is also the firm where the reductions are realized, but in the case of project-based emissions trading, three design options are available (Dutschke and Michaelowa 1999):

- Multilateral approach
- Bilateral approach
- Unilateral approach

In the multilateral approach, an international fund would be created in which private and/or public entities from industrialized countries are required to pool their investments. The bilateral model places more emphasis on private investment and market forces since project selection and implementation are left to the participants. In the unilateral model, a (legal entity within the) host government generates the credits on its own without foreign direct investment.

If such projects are not applied internationally, but domestically, the literature talks about “domestic offsets.” A domestic party then invests in a domestic project in order to generate credits that can be used by the former to meet certain emission requirements.

Project-based credit trading generates credits on the basis of the difference between baseline emissions and predicted (or actual) emissions at the project site. The baseline is an estimation of future emissions at the project site in the absence of the project. This baseline is thus a counterfactual that will never materialize.

Analysis of Emissions Trading Variants

What are the similarities and differences between cap-and-trade, performance standard rate trading, and project-based credit trading, and how do they perform in terms of effectiveness, efficiency, and acceptance? This is explained and discussed in the next subsections.

Effectiveness

Effectiveness refers to reaching the emissions targets in an emissions trading scheme. Cap-and-trade is environmentally effective because it imposes an absolute limit on total emissions. If monitoring and enforcement are in order, the emissions cap will be achieved. This is a very strong and important property of cap-and-trade. Effectiveness will only be a problem if emissions are not adequately monitored or if noncompliance

measures are not enforced. Emissions will then be higher than intended by the legislator. However, if the institutional capacity of a country is solid, cap-and-trade will reach effectiveness.

Performance standard rate trading may have a problem in reaching an absolute emission level for the industry to which the relative standard applies. As production and energy consumption rise, the emissions of companies bound by the emissions standard rise proportionally. This will not change if tradable reduction credits are added to the emissions standard. The only way to circumvent this problem is by strengthening the emissions standard. Firms could lobby against such a sharper emission requirement, unless the adjustment to the standard is automatically made.

Project-based credit trading is likely to invoke even stronger effectiveness concerns. There may be several plausible ways to calculate the baseline for emission reduction projects, in particular in host countries where domestic climate change policy is being developed or where such policy is absent. The problem is that the baseline emissions which would have occurred without the project will never be known when the project is implemented. Effectiveness can be undermined if future emissions are overestimated by inflating the baseline to claim more credits. This incentive is strongest for the investor and host under the CDM in developing countries that do not have a nationwide emissions cap. Even if credits are generated on the basis of genuine emission reductions achieved at the project location, emissions may still increase in the CDM host country outside this location (Jepma and Munasinghe 1998). The incentive to inflate the project baseline also exists for legal entities involved in a JI project in Eastern Europe, but not for the JI host party government with an assigned amount of emissions under the Kyoto Protocol since this government would run the risk of being in noncompliance by transferring too many credits.

Efficiency

Efficiency refers to the pricing of emissions so that consumers internalize the environmental damage. Sometimes efficiency is interpreted

more modestly as cost-effectiveness, which is about reaching the environmental targets at the lowest possible cost.

Cap-and-trade has “superior” efficiency properties (Tietenberg et al. 1999). In an emissions trading system based on emissions caps, each emission unit has a price and reductions in these units are profitable. From an efficiency point of view, it does not matter whether the emission allowances are allocated for free or sold at auction. If the company uses the free rights to cover its emissions, so-called opportunity costs are associated with them (Nentjes et al. 2002; Woerdman et al. 2008). The company then foregoes the opportunity to sell its allowances and misses sales revenues. The opportunity costs constitute a part of the cost price of the product. In a cap-and-trade system, each unit of emission has a price. Consequently, emission reductions in a cap-and-trade system are profitable regardless of the method by which they are achieved. It makes no difference whether such a reduction is achieved through cleaner exhaust gases, through cuts in energy consumption, or by limiting production. In a cap-and-trade scheme, there is an incentive to examine all emission reduction possibilities and to apply the least-cost option.

In a system of performance standard rate trading, companies that would have to make high costs to achieve the emissions standard will instead buy emission reductions from companies able to comply with the emissions standard at a lower cost. This improves cost-effectiveness, but when one compares such tradable reduction credits to cap-and-trade, then credit trading contains an important inefficiency. Although the emissions standard limits the emissions, the emissions within the limits set by the emissions standard remain without a price (Nentjes and Woerdman 2012). When selling credits, the received amount of money is equal to the sum paid by companies that exceed the emissions standard to purchase the credits. Consequently, for the group of companies as a whole, the cost of the permitted emissions is nil. Pollution not exceeding the relative standard is for free, and absolute emissions are allowed to rise if production or energy consumption increases.

As a consequence, there is no incentive in a performance standard rate trading system to reduce emissions by economizing on fuel input or by slowing production, because this does not earn emission reduction credits. Such credits can only be earned through reducing emissions per unit of energy or output. Mandated emissions released in producing the good have no price, and therefore, their costs are not included in the price of the product. The price of the product is then too low, which leads to overconsumption. The wider range of reduction possibilities in a cap-and-trade system leads to lower total emission reduction costs than in a performance standard rate trading system (Deweese 2001; de Vries et al. 2013).

An alternative, perhaps more simple way of explaining the difference in efficiency properties between cap-and-trade and credit trading is by looking at the supply and demand of emission rights (Woerdman 2004). If the economy grows in a cap-and-trade scheme, the demand for allowances increases, but the supply of allowances remains constant as a result of the emission ceiling. The emissions target will be achieved, and the emissions scarcity is reflected in a higher price for carbon-intensive products. If the economy grows in a credit trading scheme, however, not just demand but also supply of credits will increase since companies do not have an emission ceiling but have to observe an emissions standard. If an energy-intensive company wants to expand production, or if a newcomer enters the industry, it thus has a right to new emissions. The consequence is that the social costs of the extra emissions are not fully reflected in the costs per unit of product. Carbon-intensive products are therefore priced too cheaply.

Project-based credit trading improves the cost-effectiveness of reaching the emissions targets of the investor, but certainly does not have the full-blown efficiency properties of cap-and-trade where each unit of emissions has a price. The project-based variant of trading emissions also suffers from relatively high transaction costs, such as information costs, contract costs, and enforcement costs. Projects such as those under the JI or CDM framework usually require

pre-approval to check the environmental integrity of the project baseline, thereby raising transaction costs. Baseline standardization could improve this, by means of developing business-as-usual scenarios for several project types and regions, so that it will not be necessary anymore to construct a baseline for each individual project.

Acceptance

In order to be implemented, an emissions trading scheme needs to be politically acceptable. If an emissions trading scheme is not accepted by a political majority (or by a powerful minority), there will be no emissions trading scheme in the first place.

With respect to acceptance, the emissions trading literature has mainly focused on acceptance by companies (Dijkstra 1999). Performance standard rate trading produces cost savings for both buyers and sellers. The introduction of these credits, to supplement an emissions standard, is therefore likely to receive broad support from companies. The sale by auction of emission rights under an emissions cap is likely to be less acceptable for companies, in particular for those firms that compete on an international product market (Grubb and Neuhoff 2006). Allocating allowances for free basically has the same efficiency properties as selling the allowances at auction (Hahn and Stavins 2011).

Established companies that grow relatively fast are aware that they will not be permitted to produce higher emissions under a cap-and-trade scheme as they expand their production capacity. For this reason, they do not only prefer free rights above sale by auction, but they would rather opt for performance standard rate trading. However, companies that are able to cut emissions more cheaply prefer cap-and-trade because the demand for emissions in this system is higher than in a credit trading scheme.

There is some literature that has also focused on acceptance by politicians (Woerdman 2004). Despite the efficiency advantages of allowance trading, politicians are usually still somewhat inclined to opt for credit trading. A political economy explanation is that credit trading has advantages for certain interest groups, such as the

energy-intensive industries which do not have to purchase extra emission rights if they want to expand their production. However, there are also some advantages of credit trading for politicians themselves. Allowance trading sets emission ceilings via a rather “complex” process of allocating environmental property rights, whereas credit trading more “simply” uses existing environmental policy to calculate the tradable emission reductions. The “political transaction costs” (or start-up costs) of allowance trading are relatively high since it comes to replace existing environmental policy, while credit trading builds increasingly on extant policy, ineffective and inefficient as it may be. Moreover, under allowance trading, a choice must be made between auctioning allowances or giving them away for free, whereas emissions are always handed out for free under credit trading.

The acceptance of project-based emissions trading depends on a number of things, including the political desirability of an emissions cap, the monitoring and enforcement capacities of the administrative infrastructure, and the political merit of using market-based instruments in the first place. Countries that do not want to impose an emissions cap on their industries now, for instance, because they do not want to restrain economic growth in order to fight poverty, will not implement a cap-and-trade scheme. They are also likely to prefer project-based emissions trading to performance standard rate trading, the latter of which is more systematic in approach and thus more demanding in terms of administrative infrastructure. Governments that are critical of neoliberal market approaches are less likely to adopt any of those emissions trading design variants, although their views could change if they discover that they can earn money from selling emission rights. This attitude change has been witnessed for many poor developing countries in the case of the CDM under the Kyoto Protocol, for instance, in Africa. Being opposed first to rich countries “buying their way out” of their reduction obligations, poor African countries actually wanted to attract more CDM projects later, as soon as it became clear to them that such projects are lucrative and that most projects were going to countries like China where transaction costs are lower.

Emissions Trading Hybrids: The Case of the EU ETS

The European Union Emissions Trading Scheme (EU ETS) is a hybrid of the emissions trading design variants discussed above (Nentjes and Woerdman 2012). This regional carbon market has been up and running since 2005 and targets the power sector, a number of industrial sectors, and the aviation sector. It covers 30 countries in Europe and caps about 40% of their greenhouse gas emissions. The EU ETS has a number of implementation problems (e.g., Faure and Peeters 2008; Jaraitė et al. 2013; de Perthuis and Trotignon 2013; van Zeven 2014), for which there is not enough space here to treat them at length (for a recent overview and analysis, see Woerdman 2015). One of the causes of those implementation problems (but certainly not the only one) is that the EU ETS is a hybrid in at least two ways:

- By combining cap-and-trade with elements of performance standard rate trading
- By allowing a limited import of credits generated from CDM projects

The first hybrid element of the EU ETS concerns its new entrant and closure provisions, which resemble those of a performance standard rate trading scheme. In the textbook model of cap-and-trade, new and growing firms have to buy allowances from established companies or from a government reserve. The EU ETS, however, allocates allowances free of charge to newcomers as well as to industries expanding their production capacity (in excess of 10%). This is primarily the result of industry lobbying. Moreover, in the textbook model of cap-and-trade, a company closing down a plant can sell the allowances that remain. In the EU ETS, however, allowances need to be surrendered in case of installation closure or in case of significant decline in production capacity. This was mainly desired by politicians considering it unfair if firms would keep their allowances in such cases. Unfortunately, these credit-like rules lead to the following inefficiencies (e.g., Ellerman 2007; Nentjes and Woerdman 2012).

If companies would keep their allowances in case of closure, it would be more attractive to shut down old, inefficient plants since the allowances could then be sold. Since a company loses its allowances under the EU ETS, however, the closure of dated, climate-unfriendly installations is made less attractive, which is inefficient. Moreover, if (variable) production costs cannot be covered anymore, a company would normally shut down its installations and leave the market. Since a company will lose its allowances under the EU ETS, however, an incentive is provided to companies that make losses to maintain production capacity in order to continue receiving free allowances which can then be sold. Maintaining capacity which is not deployed for production purposes is inefficient. In addition, when newcomers or expanding firms make calculations preceding their capacity investment decision, they do not have to take the market value of their allowances into account because they get them for free. No opportunity costs are attached to this because allowances are surrendered in the event of plant closure or decline in production capacity. Here the allowances allocated for free act as credits when deciding on production capacity, like in a system of performance standard rate trading. Carbon therefore remains unpriced when production capacity is expanded, which is inefficient.

The second hybrid element of the EU ETS concerns its possibility to import relatively cheap credits generated from CDM projects to further enhance cost-effectiveness. CDM credits can be traded for EU allowances. To safeguard the effectiveness of the scheme, however, the import of CDM credits is quantitatively restricted for companies (to 11% of their allocation in the period 2008–2012 or, for newcomers, to 4.5% of their verified emissions during the period 2013–2020). The reason for this quantitative restriction is twofold: the import of such credits increases the overall emissions cap and the environmental integrity of project-based credits may be weaker than that of allowances in a cap-and-trade system. There are also qualitative restrictions as CDM credits are not allowed to be imported from nuclear energy projects, reforestation activities, or projects involving the destruction of industrial gases.

Summary

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights. There are three basic design variants of emissions trading: cap-and-trade (allowance trading), performance standard rate trading (credit trading), and project-based credit trading (such as domestic offsets, JI, and the CDM). In practice, emissions trading schemes can be hybrids of these design variants.

Cap-and-trade is effective in reducing absolute emission levels because it operates under an absolute emissions cap. Cap-and-trade is also efficient, since each unit of emissions has a price. When emissions start to rise in a performance standard rate trading scheme, however, an absolute emission level can only be achieved if the emissions standard is strengthened. An inefficiency of performance standard rate trading is that not every unit of emissions has a price. Moreover, as opposed to cap-and-trade, there is no incentive to reduce emissions by economizing on fuel input or by slowing production. Project-based credit trading is less systematic and more ad hoc due to its project focus. This also entails relatively high transaction costs, although standardizing baselines for several project types and regions helps to reduce those costs.

Firms will generally lobby in favor of free emission rights based on performance standard rate trading. Various politicians also tend to favor such tradable credits if this builds upon already existing direct regulation. As a result, emissions trading schemes sometimes become inefficient hybrids of the design variants discussed above, such as the EU ETS.

Nevertheless, those politicians that want to reach an absolute emission level at the lowest possible cost favor cap-and-trade. The current international dissemination and intended linking of cap-and-trade schemes underlines the growing relevance of emissions trading. Future research on the design elements and implementation problems of real-life emissions trading schemes could help to improve their effectiveness, efficiency, and acceptance.

Cross-References

► Transferable Discharge Permits

References

- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Dales JH (1968) *Pollution, property and prices: an essay in policy-making and economics*. Toronto University Press, Toronto
- de Perthuis C, Trotignon R (2013) *Governance of CO₂ markets: lessons from the EU ETS*. Working paper series no 2013–07, CDC Climat, Climate Economics Chair, Paris
- de Vries FP, Dijkstra BR, McGinty M (2013) On emissions trading and market structure: cap-and-trade versus intensity standards. *Environ Resour Econ*. <https://doi.org/10.1007/s10640-013-9715-2>
- Deweese DN (2001) Emissions trading: ERCs or allowances? *Land Econ* 77(4):513–526
- Dijkstra BR (1999) The political economy of environmental policy: a public choice approach to market instruments. Edward Elgar, Cheltenham
- Dutschke M, Michaelowa A (1999) Creation and sharing of credits through the Clean Development Mechanism under the Kyoto Protocol. In: Jepma CJ, van der Gaast WP (eds) *On the compatibility of flexible instruments*. Kluwer, Dordrecht, pp 47–64
- Ellerman D (2007) New entrant and closure provisions: how do they distort? *Energy J* 28:63–78
- Faure M, Peeters M (eds) (2008) *Climate change and European emissions trading: lessons for theory and practice*. Edward Elgar, Cheltenham
- Grubb M, Neuhoﬀ K (2006) Allocation and competitiveness in the EU emissions trading scheme: policy overview. *Clim Policy* 6(1):7–30
- Hahn RW, Stavins RN (2011) The effect of allowance allocations on cap-and-trade system performance. *J Law Econ* 54:267–294
- ICAP (2014) International overview of emissions trading schemes. Available online at <https://icapcarbonaction.com/ets-map>
- Jaraitė J, Jong T, Kažukauskas A, Zaklan A, Zeitlberger A (2013) Matching EU ETS accounts to historical parent companies: a technical note. European University Institute, Florence. Available online at <http://fsr.eui.eu/CPRU/EUTLTransactionData.aspx>
- Jepma CJ, Munasinghe M (1998) *Climate change policy: facts, issues and analyses*. Cambridge University Press, Cambridge
- Nentjes A, Woerdman E (2012) Tradable permits versus tradable credits: a survey and analysis. *Int Rev Environ Resour Econ* 6(1):1–78
- Nentjes A, Boom JT, Dijkstra BR, Koster M, Woerdman E, Zhang ZX (2002) National and international emissions trading for greenhouse gases. Dutch National Research Programme on Global Air Pollution and Climate Change (NRP) Report No. 410 200 093, Bilthoven
- Tietenberg T (1980) Transferable discharge permits and the control of stationary source air pollution: a survey and synthesis. *Land Econ* 56(4):391–416
- Tietenberg T, Grubb M, Michaelowa A, Swift B, Zhang ZX (1999) International rules for greenhouse gas emissions trading: defining the principles, modalities, rules and guidelines for verification, reporting and accountability. UNCTAD/GDS/GFSB/Misc.6, United Nations Conference on Trade and Development (UNCTAD), Geneva
- van Zeven J (2014) The allocation of regulatory competence in the EU emissions trading scheme. Cambridge University Press, Cambridge
- Weishaar SE (2007) CO₂ emission allowance allocation mechanisms, allocative efficiency and the environment: a static and dynamic perspective. *Eur J Law Econ* 24:29–70
- Weishaar SE (2014) Emissions trading design: a critical overview. Edward Elgar, Cheltenham
- Woerdman E (2004) The institutional economics of market-based climate policy. Elsevier, Amsterdam
- Woerdman E (2015) The EU greenhouse gas emissions trading scheme. In: Holwerda M, Roggenkamp MM, Woerdman E (eds) *Essential EU climate law*. Edward Elgar, Cheltenham (forthcoming)
- Woerdman E, Arcuri A, Clò S (2008) Emissions trading and the polluter-pays principle: do polluters pay under grandfathering? *Rev Law Econ* 4(2):565–590

Empirical Analysis

Alexander J. Wulf

SRH Hochschule Berlin, Berlin, Germany

Synonyms

Empirical law and economics; Empirical legal research; Empirical legal science; Empirical legal studies

Definition

Empirical analysis uses empirical research methods from economics and the social sciences in an attempt to provide answers to research questions in the field of law, i.e., in order to investigate the operative and functional aspects of law and legal consequences. The goal of empirical legal

research is to make a contribution to all subjects and phenomena that are of interest to legal scholars and for which no methods have previously been available. The use of empirical methods in legal science can therefore lead to results that cannot be achieved by traditional law research with the methods at its disposal. The ultimate aim of the approach is to contribute to a systematic understanding of our legal system based on empirical data.

In empirical analysis research methods from the social sciences are used to examine research questions in the legal sciences in order to study the operative and functional aspects of the law and their effects (Baldwin and Davis 2003, p. 880). The increasing trend toward testing hypotheses and questions has been described as “the next big thing” in legal science (George 2005, p. 142), even as a revolution (Ho and Kramer 2013) whose significance has as yet been underestimated (Gordon 1993, p. 2085).

In the USA the research approaches of the social sciences first began to take on greater significance in connection with legal realism. Legal realism grew up in the 1930s and 1940s, mainly in the USA, and was characterized especially by the use of interdisciplinary methods borrowed from the social sciences (Eisenberg 2011, p. 1720). In legal realism it was, for example, assumed that judges’ decisions were determined not only by laws, precedents, and general legal principles but also by the judges’ social backgrounds and political convictions (Baldwin and Davis 2003, p. 882). For example, social scientific methods were needed to be able to analyze judges’ verdicts from this viewpoint. Since these methods are often empirical, legal realism introduced the empirical paradigm into the legal sciences for the first time. However, these early empirical research efforts remained few and far between, and thus legal realism fell short of its empirical potential. Because legal realism brought the empirical approaches into the legal sciences and thus prepared a basis for their acceptance, it is seen as the trailblazer of empirical research in the legal sciences (George 2005, pp. 144–145).

Since the beginning of this century, the main country in which empirical legal research has

been practiced to any great extent is the USA (George 2005, pp. 141–142). Today it has a wide forum (Eisenberg 2011, pp. 1713–1714) in a number of journals with international repute (e.g., the *Journal of Empirical Legal Studies*, the *Journal of Legal Studies*, the *Journal of Law and Economics*, and the *European Journal of Law and Economics*), at conferences (e.g., the Annual Conference on Empirical Legal Studies and the Annual Conference of the European Association of Law and Economics), academic societies (e.g., the Society for Empirical Legal Studies, the American Law and Economics Association, the European Association of Law and Economics), and university research centers (e.g., the University of California at Berkeley – Center for the Study of Law and Society; University of California, Los Angeles – School of Law Empirical Research Group; and Hamburg University – Institute of Law and Economics, Germany). It appears likely that its initial rapid growth will continue in the future. The success of empirical methods in the legal sciences is partially due to the fact that it brings together scholars from several different disciplines who have been working independently of each other on different aspects of the legal system, e.g., the sociology of law (“Law and Society”) and the economic analysis of law (“Law and Economics”). This promotes interdisciplinary legal research (Eisenberg 2011, pp. 1719–1720 and 1722–1724).

The common goal of these different disciplines is to develop a systematic understanding of the legal system based on empirical data (Eisenberg 2011, p. 1720). The issues covered by this research are broad. The main focus is on analyzing how legislators enact and implement legal regulations and how these regulations affect the behavior of those to whom they are applied (Ulen 2008, p. 74). Empirical legal research strives as far as possible to make a contribution to all issues and phenomena which, while they are of importance for lawyers, have not previously been empirically investigable because of the lack of methods (Suchman 2006, p. 2).

The development of empirical legal research as an independent field of legal science is sometimes seen as a response to the fact that traditional

scholars of legal doctrine neglected to do empirical research on the actual functioning of the legal system (Eisenberg 2011, pp. 1734–1735). On this view, as a result of the lack of empirical data on our legal systems, many of the theories, assumptions, and prognoses on which traditional legal research has been based, either implicitly or explicitly (Ho and Kramer 2013, p. 1202), have been left without empirical support and therefore remain vague. Moreover, as a result of this neglect, the actors involved in the legal system such as the courts, parties to actions, lawyers, political decision-makers, and society in general are only insufficiently informed about the functional aspects of the legal system (Eisenberg 2011, pp. 1736–1737). Empirical legal research can be seen as an attempt to remedy this state of affairs.

The Contribution of Empirical Analysis to Legal Science

The use of empirical methods in the legal sciences can lead to insights that traditional legal research is unable to achieve with its methods. In this section the limitations of empirical research will be discussed. For reasons of space only a few central aspects will be highlighted.

Objectivity and Value Freedom

Economists and empiricists working in the legal sciences frequently adopt the view, which originated in the natural sciences, that empirical research is objective and can thus be value-free (Friedman 1953, p. 3; see also Sen 1981). However, it is highly doubtful whether it is really possible in the context of economics and the social sciences to explain, analyze, and interpret data in a value-free way. If we assume that it is not possible, this empiricist view reduces the value of empirical research for institutions, decision-makers, and politics (Hausman and McPherson 2006). This may also apply to the legal sciences since this view may result in a neglect of moral and ethical issues. Empirical research data would then be of little use to legal scholars (Hausman and McPherson 2006, pp. 291–308). Thus the support provided by empirical research for finding

solutions to normative issues, which requires a weighing up of philosophical and ethical arguments, is limited (Lawless et al. 2010, p. 21).

Qualitative and Quantitative Research Approaches

Since empirical legal science is mainly US based, it uses almost exclusively quantitative research methods, i.e., methods which are used to analyze statistical data (Suchman 2006, p. 2; Chambliss 2008, pp. 25–26). To date qualitative research methods, which are widespread in the social sciences in Europe and are employed to evaluate data from sources such as texts and audio and video and other forms of visual material, have not played such an important role (Baldwin and Davis 2003, pp. 891–892). One reason for this neglect is that in the USA empirical legal science is closely allied to the economic analysis of law (“Law and Economics”). This movement originated in economics which, unlike other social sciences, does not use qualitative research methods on principle, since they are seen as being too “soft,” i.e., not sufficiently objective. Since in Europe the sociology of law has traditionally played an important role, its quantitative orientation may have constituted an additional barrier to the spread of empirical analysis in Europe (Posner 1997, pp. 5–6).

Methodological Problems and Oversimplification of the Complexity of Legal Issues

Like all interdisciplinary research approaches, research in empirical legal science requires a high level of specialized knowledge of the various scientific disciplines. The fact that legal scholars need to acquire the methodological know-how of the social sciences (Eisenberg 2011, pp. 1728–1729) and social scientists, the detailed knowledge that legal scholars have of the functioning of the legal system (Baldwin and Davis 2003, p. 883) results in two main problems for empirical analysis. These are the methodological problems of empirical research and the problem of reducing legal complexity. Both problem areas are briefly described below.

Methodologically, empirical legal science is criticized for being of lower-than-average quality, although it can be doubted whether the standards of other forms of (interdisciplinary) research are actually any higher. The reasons for such methodological problems vary. They frequently arise from the incorrect use of methods adopted from the social sciences or from flawed research designs. In order to counteract these problems, empirical legal science needs to develop its own methodological discourse. In the long term this could promote a sustainable awareness of methods and thus improve the quality of empirical analysis (Eisenberg 2011, pp. 1730–1731). Some of the traditional publications of legal science, e.g., in the USA the student-edited law journals and in Europe-edited volumes, may not really be suitable for ensuring the methodological quality of the empirical studies submitted for publication (Chambliss 2008, pp. 26–28).

Another criticism frequently leveled at the empirical legal approach is that they reduce the complexity of legal issues (Baldwin and Davis 2003, p. 883). Such reduction is foreign to classical legal scholars, since it is not required for their traditional methods. However, if the behavior of relevance to law is to be investigated with sufficient methodological rigor in empirical studies, it is often unavoidable. Thus the decision as to what extent basic legal conditions should be taken into account in an empirical study is of substantial importance. If the legal facts are oversimplified in the interests of methodological stringency, the results of such research can have little relevance for legal science. In order to avoid this, it is necessary first to present the legal issue to be examined in its full complexity. Only then can the complexity be reduced as required. Care must be taken to fully inform the recipient of the research results of this reduction (Faust 2006, pp. 849–850). If the complexity is reduced in accordance with these requirements, this usually satisfies the demands of legal science, particularly since it is rare for the results of empirical research to replace legal discourse; as a rule they merely provide an additional perspective.

Finally, it is important to be aware that there are areas of legal science in which such reductions in complexity are not possible and which cannot

therefore be investigated by means of quantifying methods. Even technically well-implemented empirical projects must fail when their subject matter is of an ethical and moral nature. Here it can be perturbing if issues that cannot be quantified are more or less uncritically functionalized. It is the task of discourse in legal science to critically assess the value of such research.

Future Directions

In continental Europe empirical legal research is currently less widespread than in the English-speaking countries, and many obstacles to its gaining more acceptance still need to be overcome. Legal scholars who have no basic knowledge of the empirical research methods required may initially find it disturbing that the utility and scope of such research efforts in the legal sciences have often not been sufficiently clearly defined (Hull 1989, p. 915). Moreover, since it is often seen as being closely linked to economics, legal scholars sometimes think that empirical legal science is theoretical and abstract, naively functionalist, or politically conservative. Today these preconceptions can no longer be considered justified (Klerman 2002, p. 1167). Empirical analysis can contribute to a systematic, empirically based understanding of our legal system. It can therefore be assumed that it will play an increasing role in the future, not only in the English-speaking countries but also in continental Europe and other non-English-speaking countries.

Cross-References

- Causation
- Impact Assessment

References

- Baldwin J, Davis G (2003) Empirical research in law. In: Cane P, Tushnet MV (eds) *The Oxford handbook of legal studies*. Oxford University Press, Oxford, pp 880–900
- Chambliss E (2008) When do facts persuade. Some thoughts on the market for empirical legal studies. *Law Contemp Probl* 71(2):17–39

- Eisenberg T (2011) The origins, nature, and promise of empirical legal studies and a response to concerns. *Univ Ill Law Rev* 2011(5):1713–1738
- Faust F (2006) Comparative law and economic analysis of law. In: Reimann M, Zimmermann R (eds) *The Oxford handbook of comparative law*. Oxford University Press, Oxford, pp 837–865
- Friedman M (1953) The methodology of positive economics. In: Friedman M (ed) *Essays in positive economics*. University of Chicago Press, Chicago, pp 3–43
- George T (2005) An empirical study of empirical legal scholarship. The top law schools. *Indiana Law J* 81(1):141–161
- Gordon RW (1993) Lawyers, scholars, and the “middle ground”. *Mich Law Rev* 91(8):2075–2112
- Hausman DM, McPherson MS (2006) How could ethics matter to economics. In: Hausman DM, McPherson MS (eds) *Economic analysis, moral philosophy, and public policy*, 2nd edn. Cambridge University Press, Cambridge, pp 291–307
- Ho DE, Kramer L (2013) The empirical revolution in law. *Stanf Law Rev* 65(6):1195–1371
- Hull NEH (1989) The perils of empirical legal research. *Law Soc Rev* 23(5):915–920
- Klerman D (2002) Statistical and economic approaches to legal history. *Univ Ill Law Rev* 2002:1167–1176
- Lawless RM, Robbennolt JK, Ulen T (2010) *Empirical methods in law*. Aspen Publishers, Alphen aan den Rijn
- Posner RA (1997) The future of the law and economics movement in Europe. *Int Rev Law Econ* 17(1):3–14
- Sen A (1981) Accounts, actions and values. Objectivity of social science. In: Lloyd C (ed) *Social theory and political practice*. Clarendon, Oxford, pp 87–107
- Suchman M (2006) Empirical legal studies. *Sociology of law, or something ELS entirely? AMICI. Newsl Sociol Law Sect Am Sociol Assoc* 13(1):1–4
- Ulen T (2008) The appeal of legal empiricism. In: Eger T et al (eds) *Internationalisierung des Rechts und seine ökonomische Analyse. Internationalization of the law and its economic analysis, Festschrift für Hans-Bernd Schäfer zum 65. Geburtstag*, Gabler, Wiesbaden, pp 71–88

Empirical Analysis of Judicial Decisions

Luciana Yeung
Insper Institute of Education and Research,
São Paulo, Brazil

Synonyms

[Economic analysis of judicial decisions](#); [Judicial behavior](#); [Judicial decision-making](#)

Definition

Evidence shows that, contrary to what believed traditional legal theorists, judges when making decisions are not merely making a pure exercise of interpretation of the law. They are influenced by factors such as panel composition, material and nonmaterial benefits, pressure groups, etc., and may follow distinct trends. This chapter presents some models that explain judicial behavior or judicial decisions. Recent (and not so recent) empirical literature has been providing a rich debate in this still recent discussion.

Introduction: Why Care? Judicial Outcomes and Economic Performance

If judges have a substantial amount of discretion in deciding cases, then it is important to know the motives and the value systems which influence their exercise of discretion. (C. Hermann Pritchett 1968)

Coase (1960) taught us that courts and court outcomes – i.e., judicial decisions – impact the economy. Several other authors have also shown empirically that while well-functioning courts provide proper environment for productive activities, guarantee contract enforcement, and reduce uncertainties in the economy, malfunctioning courts may deter growth, investments, job creation, and increase insecurity (e.g., Weder 1995; Sherwood 2004). Since judicial decisions are courts’ main “product,” evaluating “how judges judge” and “what explains judges’ decision making” becomes a crucial task.

Although this is a complex job, which has recently been aided by studies in fields such as cognitive sciences and empirical psychology, scholars of different backgrounds (political sciences, law, economics, sociology, etc.) have been debating about this topic for several decades. Pritchett (1968) accounts Charles G. Haines as being one of the pioneers in the study of judicial behavior, with the publication of “General Observations on the Effects of Personal, Political, and Economic Influences in the Decisions of Judges” in a 1922 edition of the *Illinois Law Review*. Pritchett himself is considered by many, one of the original creators’ of this field (Epstein 2016).

From the View of Judicial Decisions as Pure Interpretations of Law to Richard Posner's Nine Theories of Judicial Behavior

One frequent concern among judicial behavior scholars is whether judges are influenced by their preconceptions or ideologies. A long debate over this issue – and not yet finished – puts *legalists* in one side and *realists* in the other. Adepts of legalism would argue that when judges judge, they are purely interpreting the law, in the best manner they can; therefore, bringing law into life is the main judicial job. Realists, on the other hand, do not believe there is a unique and certain manner to interpret the law. Each judge, when deciding, is unavoidably influenced by prior beliefs, prior personal and/or professional experiences – even if they try to strictly follow legal rules. This all may be called one's *ideology*. Realist scholars then are mainly interested in creating good and accurate measures of judges' ideology.

Adept of the realist view of judicial behavior, Posner (2008) would explain judges' personal attitudes in terms of "Bayesian preferences," as defined by Bayes statistical theorem: it shows that future probabilities of a certain kind of behavior – or decision – may be explained by former probabilities of that same behavior. Thus, in order to predict the occurrence of a certain type of behavior/decision by a judge, researchers should evaluate how was his/her decision in the past.

Besides that, Posner summarized and further developed years of theorization in judicial behavior. The author categorized nine theories of judicial behavior: Attitudinal, Strategic, Sociological, Psychological, Economic, Organizational, Pragmatic, Phenomenological, and Legalist.

Attitudinal: The attitudinal theory explains that judges' decisions are mainly reflections of their political preferences or what it is called political ideology.

Strategic: The strategic theory argues that judges' decisions reflect their preoccupations with external factors, such as opinions and pressures coming from peers, other political powers, and the rest of society (public opinion, media, interest groups, etc.).

Sociological: This theory is focused on smaller judging groups and explains why factors such as panel and voting compositions do affect judicial decision.

Psychological: The psychological theory focuses on explaining how one's preconceptions influence decision-making in circumstances under uncertainty. As Posner poses, legal systems, especially (but not only) the one in the USA, is fundamentally characterized by uncertain events, facts, and information.

Economic: The economic theory presents judges as rational, utility maximizers, who constantly behave in response to incentives and constraints. In this case, utility maximizing may be related to desire for leisure, promotion, good reputation, or even, internal feeling of satisfaction. Thus, the economic model may also encompass the strategic and the sociological theories of judicial behavior.

Organizational: The so-called agent-principal problem is the basis of the organizational theory. Mainly, this model considers judges as agents of a principal (the government) and seeks to explain judicial decisions under this perspective.

Pragmatic: The pragmatic model explains that judges are concerned with and do consider the consequences of their decisions. It is also known as the *consequentialism* approach of judicial decision-making.

Phenomenological: Posner explains that "phenomenology studies first-person consciousness – experience as it presents itself to the conscious mind" (p. 40), so this theory relates to the self-consciousness of judges when judging.

Legalist: As briefly explained above, legalism views judicial decision-making as a "pure" interpretation of the law. Legalists believe that judges when deciding at courts are solely impacted by their effort to apply the law, not being disturbed by other influences, especially, any personal preferences or preconceptions of any kind.

Due to the complexity of the phenomenon of judicial decision-making, Posner recognizes that there is no single theory able to explain it entirely. Therefore, the theories above are, actually, complementary and not substitutes.

This entry adopts some of these perspectives (e.g., attitudinal, economic) and less others (e.g., phenomenological, legalist). In subsequent sections, we provide references for empirical evidence corroborating some of these theories, especially the Attitudinal, the Strategic, and the Sociological.

Common Law Versus Civil Law: Unified or Different Models for Judicial Decision?

Early literature on judicial decisions was based on common law systems, as it happened with most of the literature on economic analysis of law. Yet, the analysis of judicial decisions has no absolute boundaries across different legal systems. The reason is simple. Judges judge lawsuits, no matter where, or under which system. In modern days, even common law judges follow constitutions and statutes, and civil law judges, not rarely, adhere to precedents (some instances in mandatory, or quasi-mandatory manner – as lower court judges following higher courts – and some other instances voluntarily). As Schneider (2005) poses, it is unreasonable to believe, as orthodox legal theory assumes, that civil-law courts only apply the law coming from the legislature. Because of that, “propositions asserting substantial differences between common-law and civil-law systems may be overstated” (p. 139).

Thus, one can infer that judicial decision-making may be analyzed as one single phenomenon, independently of the locus where it happens. This does not mean that one might not eventually find out that the outcomes of judicial decisions differ from one country to another or from one system to the other. It may also be possible to find out that judges do have different tendencies in their judicial making depending on where they are. For instances, one might conclude that judges in some countries tend to be more sensitive to “social issues,” such as inequality and wealth distribution and that their decisions reflect these concerns. Yet, these results are related to the *different variables affecting the process of judicial decision*; it does not mean that there are different

models of judicial decision, one for each legal system. It is also not true that judges, when faced to similar incentives or constraints, behave differently according to their legal origin. As the literature shows, different modes of judicial behavior are caused by different constraints and incentives which judges are faced to. The different models used to explain judicial behavior, as proposed by Posner’s “Nine Theories of Judicial Behavior,” described above, illustrate this idea.

Some Early Literature on Judicial Behavior and Decision-Making

Focusing on the US Supreme Court, and beginning in 1940s, Pritchett developed several empirical methodologies for the analysis of judicial behavior, among them, “bloc analysis” and “attitude analysis.” While later the author himself considered the first one “a rather primitive device” (1968, p. 499), it was a very useful tool for the evaluation of Justices’ voting patterns in non-unanimous decisions. Based on observations for more than 20 years, Pritchett found evidence of persistent divergences between Justices based on ideological differences. Along with “bloc analysis,” the author employed “attitude analysis.” It was one of the first attempts ever in the literature to include personal characteristics – or “personal policy attitudes” – as determinants of stronger than average voting patterns by the Justices. Pritchett’ main hypothesis was that judges are influenced by their personal ideologies, and their decisions at courts are not mere interpretations of the “letters of the law.” This methodology, which employed a “box score” to register Justices’ attitude, became a classical reference in the literature of empirical analysis of judicial decisions.

Tate (1983) created a compendium of different methodologies applied in different studies. His analysis encompassed qualitative and quantitative approaches and was complemented with critical comments. At that time, the author showed optimism in the potentials of “standard statistical methods such as regression analysis” for the study of judicial behavior (p. 74).

Fon and Parisi (2006) developed a dynamic theoretical model to explain judicial decision making in civil law systems. Their focus was not on judges as individuals, but on the result of their action, i.e., decisions and precedents. The authors evaluated the circumstances under which legal rules may be consolidated or may be corroded at courts. In their model, external shocks usually push the system into a dynamic trend, pulling it away *or* towards the point of consolidation of rules. In turn, institutional threshold of *jurisprudence constante* played a crucial role in the definition of this dynamic path.

Factors Impacting Judicial Decision: Empirical Evidence

Several factors may affect judicial decision, besides the manner by which judges interpret the law (as legalists would argue). Those would include internal factors (such as one's ideology) and external factors (such as pressures from public opinion); some may change throughout one's career (again, ideology would be an example, though here we are speaking of ideology in a broad sense), and others are constant for a certain individual (such as gender or race). The outcome – i.e., how judges effectively judge – is a combination of all those factors, and no single one explains it all, all the time. For this reason, empirically measuring the impact of a certain factor is not an easy task. Luckily, much has been advanced in the last decades. Let us review some literature on these.

Ideology

“How much ideology affect judicial decisions” is one of the most common themes in empirical analysis of judicial behavior. Since the 1st half of the twentieth century, Pritchett succeeded in finding evidence that US Supreme Court Justices were influenced by political ideologies when judging (as before). This is what Posner called “the attitudinal theory,” which “deploys... highly developed apparatus to determine the extent to which actual judicial decisions reflect the attitudes or ‘ideology’ of the judge rather than the

dictates of the ‘law’” (Cameron and Kornhauser 2017, p. 536).

As proponents of this theory, Epstein, Landes, and Posner (ELP 2013) point to the fact that there are *ex ante* and *ex post* measures of judges' political ideology. For Supreme Court Justices, the most common *ex ante* measure is the party of the nominating President. But since nominations must be approved by the Senate (and this is true for many countries), some scholars have also used measures of senatorial influence to capture Justices' ideology. Others have used editorials of influential newspaper to capture US Supreme Court nominees' *ex ante* ideological profile. *Ex post* measures, as one would expect, are based on evaluations of Justices' and judges' votes, speeches, and articles. Researchers usually combine qualitative and quantitative analysis of written and oral material produced to infer one's ideological inclinations. ELP (2013) show that “judicial self-restraint [on personal ideology] has long been in decline [since the 1960s]” for US Supreme Court Justices. In other words, the impact of ideology has been growing over time. For other courts in the USA, the authors indicate that ideology also plays a role, although in weaker magnitudes. This means that ideology has proportionally more influence at the top of the judicial hierarchy (Cameron and Kornhauser 2017).

Supreme Courts outside the USA seem also to be influenced by ideology in their judicial decisions. Yeung and Azevedo (2015) use a set of approximately 1700 decisions of the STJ (Superior Tribunal de Justiça), one of the two supreme courts in Brazil (being STJ decisions eventually appealable to STF, the Federal Supreme Court). Their original objective was to evaluate whether this court tended to favor small debtors in contractual disputes involving financial institutions. There is widespread anecdotal evidence that Brazilian courts favor weaker parties and disfavor banks. As an overall result, the authors did not confirm this popular belief. Yet, when analyzing specific variables, evidence of ideological decision emerges. For instances, Justices at the STJ decide differently depending on who stands as the defendant party: if an individual appear as defendant, the debtor tend to be favored

in a significant manner, as compared to the instances in which an enterprise appear as defendant. Justices seems to believe that, faced with financial institutions, individuals need more protection by the law than enterprises do.

Another evidence of ideology was assessed indirectly in this study. Being an appellate supreme court, STJ receives cases from 2nd degree state courts from all over the country. The authors measured whether there was any “regional factor” affecting STJ decisions. The only state with a significant result was the Brazilian southernmost state, Rio Grande do Sul: cases with that origin were consistently reformed by the STJ Justices, and in the direction of *disfavoring* debtors. It was clear that they used their power to “correct” some pro debtor trend by the Southern judges. Curiously, there is a long lasting theoretical discussion behind this fact: historically, the state of Rio Grande do Sul has been known as the birthplace of a judicial movement called *Associação dos Juizes para a Democracia* (“Association of Judges for Democracy”), whose main goal is to promote “social justice,” or more precisely, wealth redistribution, by means of the Judiciary. Judges from that state are known to be adepts to this movement and, for this reason, to be more sympathetic to “social issues,” and less favorable to the “big capital,” as banks and big enterprises. Ideological influences on their decisions are not only undisguised, but in fact, a clear statement. In this sense, Yeung and Azevedo’s results find evidence of two ideological impacts: one by judges in the state of Rio Grande do Sul, which is *pro debtor*, and the other, by the Justices at STJ, overall regarding the southern judges as “biased.” Both results were consistent and statistically significant.

Further effects have been observed in Justices’ ideologic profiles. ELP (2013) distinguish two phenomena: the *ideological drift*, and the *ideological divergence*. The former captures changes in a Justice’s *ex ante* ideology since his/her appointment to the Supreme Court. The second refers to changes in the Justice’s ideology compared to the ideology of the nominating President. This effect explains why Justices may, sometimes, vote different to what is expected. Ideological divergence

may happen gradually (i.e., Justices start at the Supreme Court much aligned with their nominating President, but then divergence increase due throughout the time), or it may exist since the beginning of a nominee’s career at the Supreme Court. In this case, it may be caused by the President’s lack of information about the nominee’s profile, or it may happen because the President had other political compromises when nominating the Justice, which were unrelated to ideological positions.

Abundant are theories and empirical evidence of ideology impacts on judicial decision-making; it is not the only factor explaining judges’ behavior. It would be naïve to believe that judges, even Supreme Court Justice, would be able to use their discretionary power to entirely (and only to) pursue their personal ideologies. Several other constraints curtail these motives.

Gender

The impact of a judge’s gender on judicial decision has also been studied in several studies; as before, we bring a short selection.

Peresie (2005) finds judges’ gender as an impacting factor on decisions at USA appellate courts on disputes over sexual harassment and sex discrimination. Gender acts as a direct impact factor – i.e., female judges favor more frequently the victims of discrimination – and as an indirect factor, through peer effect on panels – i.e., female judges do influence their male counterparts when judging such cases. Peresie finds that panels with female judges tend to favor alleged victims twice as often as panels with only male judges. In this study, gender was more impacting than ideology on judicial decisions.

Farhang and Wawro (2004) find a strong panel effect by women, i.e., female judges tend to influence their male colleagues in the panels. Yet, they find that a second woman on the panel does not have the same effect as the first one. These authors also try to find evidence of race impact, but – differently to the gender factor – they find none, although they are cautious to interpret this latest result.

Boyd et al. (2010) employ propensity score matching and also find significant gender impact

in sex discrimination disputes. Here, as in Peresie (2005), impacts occur both directly (on individual female judges) and indirectly (on male counterparts in panels). Although the authors analyzed 13 types of judicial disputes, only those on sex discrimination were significantly impacted by the judges' gender. Within their list of analysis, there were other types of disputes that are usually gender sensitive, such as abortion and sexual harassment, but on no other issue had gender significant impact.

Other studies bring similar evidence for courts outside the USA. King and Greening (2007) analyzed decisions by the International Criminal Tribunal on cases of sexual assault in former Yugoslavia. They found that female judges tend to punish more severely defendants who assaulted women – in what constitutes a “gender solidarity” between judges and victims. This solidarity seems also to be present when all-male panels analyzed cases involving male victims: sentences for these cases were more than 100 months lengthier than those in which there was at least one female judge on the panel. In France, under the context of child support orders, Bourreau-Dubois et al. (2014) find that when the parents' average offer is reasonable, female judges tend to be more generous than male judges.

Grezzana and Ponczek (2012) analyzed more than 90,000 labor disputes at the Brazilian Superior Labor Court. Overall, the authors find no evidence of gender impact. However, once they control for the object of dispute, impact is evident for cases such as “wage equalization” and “employment and union link.” Under these circumstances, female judges tend to favor female litigants (workers), whereas male judges tend to favor male ones. Again, there seem to be some kind of “gender solidarity” between judges and litigants in the Brazilian Superior Labor Court.

Why would judges' gender have impact on the manner by which they decide? Based on previous literature, Boyd et al. (2010) draw 4 accounts affecting judging, either individually or in groups (panels). First, *different voice* is the manifestation of the differences that male and female individuals see and analyze the world and society; basically, this is the individual

perspective of things, particularly here, of the cases disputed at courts. *Representational* account is the manifestation of female judges seeing themselves as representatives of all female individuals in society, and specifically, of female litigants in disputes. Female judges would decide in favor of women in cases where there are particular interests for the whole class of women in society. Third, *informational* account posits female judges as having more information that would be valuable for resolving the dispute. Under such circumstances, their male counterparts would benefit from this privileged information, and the effect will be channeled through the panel voting. Last, *organizational* account will actually oversee the gender impact on judicial decisions; the view here is that professional training and institutional rules in the Judiciary are clear and similar enough to minimize any significant differences between male and female judges. All these accounts have been explored, tested, and analyzed by a rich literature on this topic Boyd et al. (2010) provide detailed references on each of these approaches.

Besides gender, there are other factors affecting judicial decisions, which are related to minority groups such as race, ethnicity, religious group, and social background, among others (e.g., an interesting piece of work by Schwartz and Murchison 2016, on the impacts of ethnicity-nationality at the Constitutional Court of Bosnia-Herzegovina). Due to the limitations of this chapter, we will leave these topics undiscussed, despite their undoubted importance as the empirical literature on judicial behavior has already shown.

Voting Panels, Composition, and “Peer Effect”

Before, we have seen some accounts of how panel composition at courts affect judges' voting patterns. Social psychologists and behaviorists have long studied the effects of peer pressure in organizations, mostly enterprises, and one would expect – and actually observe – the same happening in public and political organizations, such as courts (and the Congress, etc.).

Epstein, Landes, and Posner (ELP 2013) have a theoretical explanation for the occurrence of panel composition effect, and they test it. Panels

may decide unanimously – when there is no dissenting vote – or non-unanimously – when there *is* dissent. ELP explains that there are costs and benefits of dissenting, and not rarely the former supersede the latter. Costs of dissenting include writing the dissenting opinion, disagreeing with colleagues, and reputational costs infringed on the other members of the panel. All these create *dissent aversion* and, consequently, one will avoid disagreeing on minor issues – specifically technical ones – and dissent more frequently will be caused by ideological disagreements, which are more difficult to resolve by discussion and compromise. ELP (2013) show evidence of this effect for the US Supreme Court from years 1953 to 2008. For other lower courts, the authors predict more dissent when an appeal is being reversed (dissenter has ally in the district court), and less dissent in smaller courts of appeals (judges sit together more frequently, making dissent much costlier). The authors also predict that dissent will be inversely proportional to the court's workload, i.e., the busier judges are, the less they will dissent. Historical evidence from the US Supreme Court and the Courts of Appeals corroborates the authors' predictions. ELP also find that the presence of dissent impact on the length of written voting: majority opinions are longer if there is one dissenting member on the panel, and still longer if there is more than one. Apparently, more words are necessary to justify an opinion when it is faced with opposition. Finally, the authors link the frequency of dissent to one's career: federal judges in the USA tend to dissent more during the first half of their active lives.

Voting panels may also *potentialize* some other factors, for instances, political ideology. As observed by Sunstein et al. (2006, *apud* ELP 2013), judges appointed by Republican Presidents decide against affirmative action cases more often than those appointed by Democratic Presidents; yet, the frequency is much higher for full-Republican panels and much lower for full-Democratic ones. For the gender factor, similar peer effect is observed: all-female panels tend to favor female litigants and the opposite for all-male panels on male litigants (as shown above). However, sometimes, panels may *attenuate* or

even *reverse* ideology impact. Studies have found evidence of liberal judges deciding in more conservative manner on panels than they would otherwise do, and the contrary to conservative judges (check on ELP 2013, chapter 2).

Outside the USA, Smyth (2005) studied the pattern of dissent on the Australian High Court. He finds evidence for dissent caused by diverging political ideology, but no evidence of relation between caseload and dissent rate. As to a judges' active career, Smyth finds evidence of *increasing* dissent rate throughout the time, a diverging result to what ELP have shown for the USA.

Related to peer effects, some interesting literature on the influence of social norms on judicial behavior may also be accounted for. Using an analytical approach, Harnay and Marciano (2004) build up a model in which they show that judicial behavior is not entirely a product of individual calculation; instead, it reflects the interactions in a system where judges do care about what other professionals in the community think and do and have desires for certain degree of conformity. In other words, they show evidence that judicial behavior cannot be explained solely by the economic theory from Posner, but also by elements of the strategic theory. This "tendency to conform" to precedents by a particular judge is a result of a cost-benefit analysis: he/she analyzes the expected gains of deviation versus the gains of compliance with the precedents and does that at the private and professional levels. This, according to the authors, explains why judges behave in conformity to precedents under certain circumstances and deviate under others.

External Pressures: Media, Popular Opinion, and Interest Groups

Besides the effect exerted by peers in voting panels as discussed above, there are other sources of external factors that might influence judicial decisions. Media and popular opinion have always restrained, somehow, public agents' behavior; yet, the intensity of this impact has grown exponentially with the modernization of telecommunication technology. In some countries, Supreme Court voting sessions are broadcast live on TV channels. Although average citizens can

rarely comprehend the matters discussed at courts – and especially high courts – due to their complexity and technicality, from times to times Justices' decisions are in the spotlight, highlighted on the front pages of newspapers, in the evening TV shows, and discussed by lay citizens. Thus, even judges who are not directly elected feel, somehow, constrained by what society has to say about the outcomes of their work. As Epstein and Kobyłka (1992) pose: “Most modern Court decisions reflect public opinion. When a clear-cut poll majority or plurality exists, over three-fifths of the Court's decisions reflect the polls. By all arguable evidence the modern Supreme Court appears to reflect public opinion as accurately as other policy makers” (p. 24).

Organized interest groups may also exert considerable impact on judicial decisions; although not new – at least in the North-American context – this is a phenomenon that has grown recently, and that is still not well understood (Epstein and Kobyłka 1992).

Empirical literature on the effects of media, public opinion, and interest groups on judicial decision is also vast and growing. Due to its higher exposure and its greater impact on the rest of society, studies of this kind have mainly focused on supreme courts. Casillas et al. (2011) find significant influence of public opinion on the US Supreme Court decisions and more impacting in nonsalient cases (because in salient cases, there are high stakes in following legalistic considerations and/or personal ideologies). The authors measure the costs that the Supreme Court incurred by ignoring public opinion in nonsalient cases during the period of 1970s to year 2000. Giles et al. (2008) follow the same direction and, even though they are more cautious about the existence of direct impacts of public opinion on Supreme Court outcomes, they do assert that there is evidence of causality on Justices' voting.

Epstein and Martin (2010) also find evidence that Supreme Court decisions are, to a certain degree, aligned with public opinion. Besides the usual explanation that Justices care about their reputation and society's approval, Epstein and Martin argue that the alliance may occur because Justices *are themselves* part of society and of the public. Thus, in this case, they are actually

deciding based on their personal ideologies, and not only as a reflection of external preferences. It would not be easy to empirically separate these two effects, and the authors leave the analysis for future studies.

With regards to interest groups, Collins and Martinek (2010) find evidence of their impact on the probability of success by appellants, but not by appellees, on the US courts of appeals. For the Supreme Court, Collins and Solowiej (2007) also show that interest groups, represented by *amici curiae* briefs presented to the Court, do affect the amount of information that would be important for judges' decisions. A positive outcome the authors find at the US Supreme Court is that it is accessible to a wide variety of interest groups, which reflect a democratic and pluralistic society. However, despite their clear influence, at least on provision of information, the authors were not able to explain how and how much do these groups affect judicial decision-making.

A notable remark about the involvement of interest groups in judicial decisions is made by Epstein and Kobyłka (1992). They differentiate this type of activity with those of other political pressure groups: “Unlike the more traditional arenas of group lobbying, . . . [the import of interest groups in the Judiciary] is not so much derived from their numbers but is more a function of the kinds of arguments they present to a Court” (p. 306). These authors believe this the manner by which they influence judges' decision.

Pressure coming from the Executive and the Legislature Powers (i.e., the Congress) also exerts significant impacts on judicial decisions. The interplay between judges and those actors has long been discussed by legal scholars and is a never-ending object of study. Especially in the case of Supreme Courts, due to the Presidential nomination of its Justices, the quest for a better understanding of this relationship is related to the crucial matter of independence of powers.

Lopes and Azevedo (2017) find evidence in this direction for Brazil. Under that system, there are two high courts: STJ (Superior Tribunal de Justiça) – for civil appeals – and STF (Supremo Tribunal Federal) – a constitutional court, which is de facto above STJ. Although Justices at both

courts are formally nominated by the President, STF is more political, with closer links to the Executive and Legislative. Presidents also have more discretion in the nomination of STF Justices. Perhaps because of that, this study finds that STF is significantly more impacted by political influences than STJ.

Garoupa et al. (2013) try to understand these influences at the Spanish Constitutional Court. They conclude that, due to the limitations imposed by civil law systems, impacts in this Court do not strictly follow the patterns predicted by the attitudinal model, specifically on the dichotomy liberal x conservative, or left x right.

As of the relations between courts (especially the Supreme Court) and Congress, Epstein and Kobylka pose that “[they] are hardly random. The political composition of the legislature vis-à-vis that of the Court plays a major role in determining the course of those relations, be they antagonistic or amiable” (p. 24). The fact that Supreme Court Justices must be approved by the Senate also poses a constraint to the first. Yet, since policies created by the Congress, if questioned by unsatisfied litigant groups and individuals from society, may be overruled by the Court, the power factor and impact is not a one-way path. Again, due to the limitations of this chapter, we will leave detailed discussions apart. (Further references may be found on Epstein et al. 2013, chapter 2).

Another interesting type of external pressure may affect judicial behavior, specifically at the context of international courts. Under such regimes, in which no strict hierarchy exists between international and local courts, the problem of compliance by these ones is intensified. Dyevre (2016) designs a theoretical model in which he predicts the circumstances under which domestic judges have incentives to accommodate with the European Court of Justice (ECJ). For this purpose, the author employs the concept of *crisis costs*, or costs for domestic noncompliance. The level of these costs is the determinants of the behavior by local judges. His model shows that, in legal regimes as integrated as the EU, non-compliance (or defiance) makes local courts worse off. Because of the magnitude of such

costs, there are incentives for the ECJ and these courts, even strong ones – such as the German Federal Constitutional Court – to seek accommodation through dialogue. For smaller domestic courts, there is no substantial benefit to defy the ECJ generating conflicts. Thus, the extent to which pressure may be exerted upon judicial decision may go beyond national borders.

Indeed, the analysis of international courts, such as the ECJ, renders several possibilities to test hypotheses on judicial behavior. (Other insightful research about judicial decision-making at the ECJ may be found at: Josselin and Marciano 1997; Tridimas and Tridimas 2002; Portuese 2012; among others). In this short chapter, it will not be possible for us to dwell extensively on the multiple works on this theme; yet, we close with one another intriguing work: Vaubel (2009), by analyzing data from 42 countries, shows that the ECJ is indeed biased towards political centralization. Furthermore, and most importantly, this bias is a consequence of independent courts (not the contrary), which are not subordinated to the European Parliament and the European Commission.

Benefit Maximization (or Satisfaction)

Another set of factors that might significantly affect judicial decision-making is the desire that judges have to maximize self-benefits. In Posner’s terms, this is the so-called economic theory of judicial behavior. One example of benefit that judges normally try to maximize is the chances to succeed in their career. By observing the Italian Constitutional Court, Melcarne (2017) corroborates the hypothesis that judges’ career concerns affect their behavior; the reason is the reputational impacts of their conduct. Furthermore, these concerns are independent to judges’ personal characteristics. In line with the discussion in the previous item, Melcarne also observed that judges are sensitive to external pressures, especially interests of the Executive Power.

Schneider (2005) also confirms career incentives in judges’ behavior; yet, his observations of the German Labor Courts go further, by analyzing theories of internal labor markets (or tournaments) in the judicial structure. As a combination of career concerns and external pressures

(as discussed in the previous item), the author shows that “[j]udges are likely to decide in a way that conforms to the policies or opinions of those agencies who influence their appointment. These agencies can be higher-level courts, parliaments, and governments” (p. 140). He also claims that these results might be generalized to other jurisdictions, due to the similarity of the structure of German labor courts and that found in many other civil-law countries.

However, some researchers are cautious about viewing judges as benefits maximizers, whatever benefits might be translated to. As Christmann (2014) poses “[m]odels that treat judges as rational maximizing men commonly find a stronger distortion in the resolution of disputes” (p. 411). Because of this skepticism, and with the influential developments of behavioral economics, a new derivation of the benefit-maximization theory on judicial decision-making is emerging. Authors following this line see judges not as benefit-maximizers, but as Tsoussi and Zervogianni (2010) pose, “satisficers” “who make decisions within real-world constraints, such as imperfect information and uncertainty, cognitive limitations and erroneous information” (p. 333). The concept of “satisficer” comes from Herbert Simon’s notion of “bounded rationality”: instead of fully rational maximizers – who seek the “best” outcome, satisficers make decisions based merely on the choice of “good enough” outcomes. Under this view, judges’ goal “is not optimize but to render opinions that are merely satisfactory” (p. 333), and because of that, they often engage in “improper” behavior, especially under nasty environmental conditions, such as overloaded court dockets and increasingly complex lawsuits.

The theory of judges as satisficers is a promising one, however, still needs to be corroborated by empirical analysis. Hopefully, the ever growing literature on behavioral law and economics will, very soon, bring promising results in this direction.

Above, we have shown several factors that, according to a rich empirical literature, may impact judicial behavior. Due to the limitations of this chapter, the discussion was brief and non-exhaustive. However, some authors argue that

there is still a long way to go, in the study of the factor impacting judicial decision. They claim that several qualitative factors have been entirely left out the discussion up to now. For instances, Richards (2017) points to “background variables, such as [judges’] education, prior experience, training, socio-economic background, personality,” which could “contribute to a more universal understanding of judicial behavior, crossing the boundaries of both space and time” (p. 558). Indeed, much still needs to be done.

Future Directions

What can one infer from this brief discussion and review of the empirical studies on judicial decisions? What have the studies so far evidenced?

At the beginning, the studies had a more *positive* approach: attempts to capture trends in judicial decisions, sources of personal ideology, and other factors that might impact court outcomes.

Throughout the time, as theories and methodologies developed (although, as seen above, there is still much to be pursued), authors of judicial decisions had increasingly (although sometimes not explicitly) adopted a *normative* perspective. Resulting evidence shows that judges are influenced by factors such as panel composition, material and nonmaterial benefits, pressure groups. Based on that, one might ask how institutions should be better designed to provide the “right” incentives for judges to behave in the “desired” manner, by making decisions that are more democratic, more inclusive, more efficient, . . . ? (And each one has his/her own desired outcome for judicial decisions.)

Empirical analysis of judicial decisions has helped to answer these questions, too.

References

- Bourreau-Dubois C, Doriât-Duban M, Ray JC (2014) Child support order: how do judges decide without guidelines? Evidence from France. *Eur J Law Econ* 38(3):431–452
- Boyd CL, Epstein L, Martin AD (2010) Untangling the causal effects of sex on judging. *Am J Polit Sci* 54(2):389–411

- Cameron CM, Kornhauser LA (2017) Rational choice attitudinalism? *Eur J Law Econ* 43:535–554
- Casillas CJ, Enns PK, Wohlfarth PC (2011) How public opinion constrains the U.S. Supreme Court. *Am J Polit Sci* 55(1):74–88
- Christmann R (2014) No judge, no job! Court errors and the contingent labor contract. *Eur J Law Econ* 38(3):409–429
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Collins PM, Martinek WL (2010) Friends of the circuits: interest group influence on decision making in the U.S. courts of appeals. *Soc Sci Q* 91(2):397–414
- Collins PM, Solowiej LA (2007) Interest group participation, and conflict competition, the U.S. Supreme Court. *Law Soc Inq* 32(4):955–984
- Dyevre A (2016) Domestic judicial defiance and the authority of international legal regimes. *Eur J Law Econ* 44:453. <https://doi.org/10.1007/s10657-016-9551-2>
- Epstein L (2016) Some thoughts on the study of judicial behavior. *William & Mary Law Review* 57(6):2017–2073
- Epstein L, Kobylka JF (1992) The Supreme Court and legal change: abortion and the death penalty. Thornton H. Brooks series in American Law & Society. The University of North Carolina Press, Chapel Hill
- Epstein L, Martin AD (2010) Does public opinion influence the Supreme Court? Possibly yes (but We're not sure why). *J Constit Law* 12(2):263–281
- Epstein L, Landes WM, Posner RA (2013) The behavior of federal judges. Harvard University Press, Cambridge
- Farhang S, Wawro G (2004) Institutional dynamics on the U.S. court of appeals: minority representation under panel decision making. *J Law Econ Org* 20(2):299–330
- Fon V, Parisi F (2006) Judicial precedents in civil law systems: a dynamic analysis. *Int Rev Law Econ* 26:519–535
- Garoupa N, Gomez-Pomar F, Grembi V (2013) Judging under political pressure: an empirical analysis of constitutional review voting in the Spanish Constitutional Court. *J Law Econ Org* 29(3):513–534
- Giles MW, Blackstone B, Vining RL (2008) The Supreme Court in American democracy: unraveling the linkages between public opinion and judicial decision making. *J Polit* 70(2):293–306
- Grezzana S, Ponczek V (2012) Gender bias at the Brazilian superior labor court. *Braz Rev Econom* 32(1):73–96
- Harnay S, Marciano A (2004) Judicial conformity versus dissidence: an economic analysis of judicial precedent. *Int Rev Law Econ* 23:405–420
- Josselin JM, Marciano A (1997) The paradox of leviathan: how to develop and contain the future European state? *Eur J Law Econ* 4(1):5–22
- King KL, Greening M (2007) Gender justice or just gender? The role of gender in sexual assault decisions at the international criminal tribunal for the Former Yugoslavia. *Soc Sci Q* 88(5):1049–1071
- Lopes F, Azevedo PF (2017) Government appointment discretion and judicial independence: preference and opportunistic effects on Brazilian Courts. Working paper – Insper. Available on <https://www.insper.edu.br/working-papers/working-papers-2017/government-appointment-discretion-and-judicial-independence-preference-and-opportunistic-effects-on-brazilian-courts/>
- Melcarne A (2017) Careerism and judicial behavior. *Eur J Law Econ* 44:1–24
- Peresie JL (2005) Female judges matter: gender and collegial decision-making in the federal appellate courts. *Yale Law J* 114(7):1759–1790
- Portuese A (2012) Law and economics of the European multilingualism. *Eur J Law Econ* 34(2):279–325
- Posner RA (2008) How judges think. Harvard University Press, Cambridge
- Pritchett CH (1968) Public law and judicial behavior. *J Polit* 30:480–509
- Richards D (2017) Lee Epstein, William M. Landes, Richard A. Posner: the behavior of federal judges: a theoretical and empirical study of rational choice. *Eur J Law Econ* 43:555–558
- Schneider MR (2005) Judicial career incentives and court performance: an empirical study of the German labour courts of appeal. *Eur J Law Econ* 20(2):127–144
- Schwartz A, Murchison MJ (2016) Judicial impartiality and independence in divided societies: an empirical analysis of the constitutional court of Bosnia-Herzegovina. *Law Soc Rev* 50(4):821–855
- Sherwood RM (2004) Judicial performance: its economic impact in seven countries. Paper presented in the 8th annual conference da international society for new institutional economics (ISNIE), Tucson. Available on <http://www.isnie.org>
- Smyth R (2005) The role of attitudinal, institutional and environmental factors in explaining variations in the dissent rate on the High Court of Australia. *Aust J Polit Sci* 40(4):519–540
- Sunstein CR, Schkade D, Ellman LM, Sawicki A (2006) Are judges political? An empirical analysis of the federal judiciary. The Brookings Institution, Washington, DC
- Tate CN (1983) The methodology of judicial behavior research: a review and critique. *Polit Behav* 5(1):51–82
- Tridimas G, Tridimas T (2002) The European Court of Justice and the annulment of the tobacco advertisement directive: friend of national sovereignty or foe of public health? *Eur J Law Econ* 14(2):171–183
- Tsaoussi A, Zervogianni E (2010) Judges as satisficers: a law and economics perspective on judicial liability. *Eur J Law Econ* 29(3):333–357
- Vaubel R (2009) Constitutional courts as promoters of political centralization: lessons from the European Court of Justice. *Eur J Law Econ* 28(3):203–222
- Weder B (1995) Legal systems and economic performance: the empirical evidence. In: Rowat M et al (eds) *Judicial reform in Latin America and the Caribbean – proceedings of a World Bank Conference*. World Bank technical paper number 280. The World Bank, Washington, DC
- Yeung L, Azevedo PF (2015) Nem Robin Hood nem King John: testando o viés anti-credor e anti-devedor dos magistrados brasileiros. *Econ Anal Law Rev* 6(1):1–12

Empirical Law and Economics

► [Empirical Analysis](#)

Empirical Legal Research

► [Empirical Analysis](#)

Empirical Legal Science

► [Empirical Analysis](#)

Empirical Legal Studies

► [Empirical Analysis](#)

Endowment Effect

Ben Depoorter^{1,2} and Sven Hoeppe²

¹Hastings College of the Law, University of California, San Francisco, CA, USA

²Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium

Abstract

A vast body of experimental studies in psychology and economics finds that individuals tend to value goods more and demand higher prices when they own the goods than they would be willing to pay for the good when they do not already own it. Although research on the endowment effect has been done for more than three decades, its theory, empirical methodology, results, and implications continue to be topics of intense discussion among economists, lawyers and psychologists. In this entry, we review the theoretical framework and

empirical evidence on the endowment effect and highlight some implications for law and economics research.

Definition

The term “endowment effect” describes observations of a gap between the willingness-to-pay for a good not owned already and the willingness-to-sell a good in one’s possession. In other words, the term describes the tendency that people value things higher merely because they own them.

Introduction

A major tenet of rational choice theory is that individual preferences are endogenous: although individuals may value things differently, preferences exist independently of situational circumstances. This principle, most commonly referred to as “preference exogeneity” (Korobkin 1998a, p. 611), “source independence” (Issacharoff 1998, p. 1735), and “status irrelevance” (Korobkin 2003, p. 1228), plays a crucial role in the economic analysis of law. For instance, when economists analyze how the legal system should distribute entitlements among competing claimants or how to regulate consensual exchanges of entitlements after their initial allocation, it is generally assumed that legal rules do not influence the value of the entitlement itself.

Behavioral research demonstrates however that the assumption of preference exogeneity does not hold in many circumstances. A vast body of experimental studies in psychology and economics finds that individuals tend to value goods more and demand higher prices when they own the goods than they would be willing to pay for the good when they do not already own it (compare literature reviews and extensive references in: Horowitz and McConnell 2002; Korobkin 2003; Morewedge et al. 2009). The observed gap between the willingness-to-pay (WTP) for a good not owned already, on the one hand, and the willingness-to-sell (WTS) a good in one’s possession, on the other hand, was termed

“the endowment effect” in a seminal paper by Richard Thaler (1980, p. 44).

Although research on the endowment effect has been done for more than three decades, its theory, empirical methodology, results, and implications continue to be the topics of intense discussion among economists, lawyers, and psychologists (cf. Plott and Zeiler 2005, 2007, 2011; Arlen and Talley 2008; Isoni et al. 2011; Klass and Zeiler 2013). This entry briefly reviews the theoretical framework and empirical evidence on the endowment effect. Moreover, it highlights a few implications for the economic analysis of law.

Empirical Evidence on the Endowment Effect

Early demonstrations of the endowment effect at work led to most striking results. In their pioneering study, Knetsch and Sinden (1984) randomly divided experimental participants into two groups. One group received a lottery ticket for a \$50 cash prize. Participants in the other group received \$3 in cash. Then the researchers offered to sell or buy the lottery tickets for \$3. A striking majority of the ticket holders, specifically 82%, kept their lottery tickets. This suggests that the WTA of these subjects was greater than \$3. Moreover, only 38% of the cash owners chose to buy tickets, which suggests a WTP of less than \$3.

In another early study that received much attention, Knetsch (1989) endowed some participants with coffee mugs and later offered each of these participants to trade a large Swiss chocolate bar for a mug. Other participants received a large Swiss chocolate bar and later had the chance to trade the chocolate for one of the same mugs. Surprisingly, not only did merely 11% of the mug owners choose to trade them for the chocolate bar. This would be intuitive if the mugs had a higher value. However, also merely 10% of the other participants, who were endowed with the chocolate bar, were willing to trade it in for a mug. In fact, Knetsch (1989) used a third group to provide a baseline measure. Participants in this group had a choice between a mug and a chocolate bar without any initial endowment. The results

suggest that participants perceived the two goods as roughly equivalent on average: 56% chose the mug and 44% opted for the chocolate bar.

While the endowment effect appears to be robust across ages – at least across ages 5 (children in kindergarten) to 20 (college students) (Harbaugh et al. 2001) – the endowment effect varies across different *types of entitlements*. Due to the constraints of the experimental method, endowment effects are mostly observed for small and tangible consumer items [see, however, a study by Dubourg et al. (1994) on endowment effects for small changes in safety features in the purchase of cars]. Meanwhile, no endowment effects appear to occur when involving tokens with a certain value (van Dijk and van Knippenberg 1996). Conversely, this study observed a significant WTA/WTP disparity when the value of the token was uncertain. In line with this finding, the endowment effect is observed to be more pronounced when it is difficult for individuals to compare the two commodities involved (Bar-Hillel and Neter 1996; van Dijk and van Knippenberg 1998). In a similar vein, Shogren et al. (1994) found that the endowment effect is stronger for endowments for which no close substitutes are available. Curiously, endowment effects were absent in experiments involving items with close substitutes (candy bars), although endowment effects appear robust when concerning many other fungible items (coffee mugs, ballpoint pens). A meta-study on the endowment effect finds that endowment effect is strongest for nonmarket goods, next highest for ordinary private goods, and lowest for experiments involving forms of certain value tokens such as money (Horowitz and McConnell 2002). Similarly, the strength of the endowment effect has also been found to vary across various *circumstances*. Findings in an experiment by Loewenstein and Issacharoff (1994) suggest that the endowment effect is stronger when the commodity is obtained as a result of skill or performance as opposed to chance or random luck.

The endowment has also been linked to *status quo bias* (Samuelson and Zeckhauser 1988), i.e., a preference for the current state that biases people against both selling and buying goods. This

explanation finds some support in a series of field experiments by Hartman et al. (1991).

Critics have challenged endowment effect findings as resulting from flawed experimental designs and “strategic heuristics” (i.e., subconscious applications of usually sensible bargaining habits). Plott and Zeiler (2005) claim that the observed WTA/WTP gap disappears when implementing experimental procedures that control for subjects’ misconceptions about the value of the traded commodity, for instance, by preventing that participants receive a signal that increases the value of the endowment when it is presented to them as a gift by the experimenter. The underlying generalization that the endowment effect is merely an artifact of experimental methodology is a hotly debated issue. As a response to generalizations from Plott and Zeiler’s (2005) study, Isoni et al. (2011) find that the WTA/WTP gap is reduced for mugs, but remains for lotteries even when employing Plott and Zeiler’s (2005) method. To date, no credible conclusion can be drawn from the debate (see also Plott and Zeiler 2011; Klass and Zeiler 2013).

It has also been remarked that the observed gap between willingness-to-pay and willingness-to-sell might not occur in market settings that involve experienced traders and learning opportunities (cf. Knez et al. 1985; Coursey et al. 1987; List 2003). For instance, Coursey et al. (1987) reported results that indeed suggest that the WTA/WTP gap may diminish in such settings. [For a critical discussion of these findings, see Knetsch and Sinden 1987.] Kahneman et al. (1991) ran a set of experiments comparing endowment effects with regard to tokens and consumer goods traded in market setting with repeated interactions that provided learning opportunities. While there was no significant endowment effect on the token market, as soon as the commodity was changed to coffee mugs or ballpoint pens, the market did not clear even in repeated trials. On the one hand, the results indicate that there is no similar effect for money of other value tokens as long as the embodied value is certain (similarly van Dijk and van Knippenberg 1996). On the other hand, strong endowment effects emerged when the traded commodity consisted of consumer goods. In both the market

for ballpoint pens and the market for coffee mugs, selling prices were about twice as high as buying prices. Specifically for the mugs, the median mug holder was unwilling to sell below \$5.25, while the median buyer was unwilling to pay more than \$2.25–\$2.75. Experiments involving repetition (cf. Shogren et al. 1994) and studies that employed a subject pool of very experienced traders (List 2004) did not reliably reduce the endowment effect and, therefore, further dispel the notion that the WTA/WTP gap merely results from inexperience.

Recent research has further expanded the psychological foundation of the endowment effect and, in general, retains the central idea that potential sellers contemplate a large loss, whereas potential buyers contemplate a small gain (Nayakankuppam and Mishra 2005; Zhang and Fishbach 2005; Johnson et al. 2007).

Explaining the Endowment Effect

To date, the cause of the endowment effect remains a topic of contention. Experiments have ruled out many of the potential explanations derived from traditional consumer theory, including income effects, transaction costs, information problems, and market mechanisms (cf. Brown 2005; also compare, Knetsch and Wong 2009).

Richard Thaler (1980) initially proposed that the endowment effect results from the different mental treatment of out-of-pocket costs and opportunity costs, which he argued to be mentally coded as loss and foregone gains, respectively. Through this underweighting of opportunity costs, he linked the endowment effect to the asymmetry in value that Kahneman and Tversky (1979) describe as *loss aversion*: the observation that the negative value of giving up an object is greater than the value associated with acquiring it. The fact that losses are weighted more strongly than objectively commensurate gains might help explain endowment effect observations since “if a good is evaluated as a loss when it is given up and as a gain when it is acquired, loss aversion will, on average, induce a higher dollar value for owners than for potential buyers, reducing the set of mutually accepted trades” (Kahneman et al. 1990, p. 1328).

To clarify the impact of loss aversion, one study specifically investigated whether the gap and low trading volume resulted from a reluctance to sell or a reluctance to buy. The experiment involved three groups. Only the “sellers” received a coffee mug and were asked whether they would be willing to sell the mug at each of a series of prices between \$0.25 and \$9.25. The “buyers” were asked whether they would be willing to buy a mug at the same prices. Finally, the “choosers” were asked to choose at each of the prices between receiving a mug or the money. Recognizing that the “sellers” and “choosers” are in objectively identical situations – deciding at each price between the mug and the money – reveals a surprising pattern: the median reservation prices were \$7.12 for the “sellers,” \$3.12 for the “choosers,” and \$2.87 for the “buyers.” The results suggest that the observed patterns are “produced mainly by owner’s reluctance to part with their endowment, rather than buyers’ unwillingness to part with their cash” (Kahneman et al. 1991, p. 196).

Loss aversion, however, is merely a descriptive construct employed by Kahneman and Tversky (1979) to explain the asymmetry in prospect theory’s value function. It begs the further question as to why individuals experience it. Psychologists have stated that loss aversion may stem from a preference to avoid *regret* and/or to remain *consistent* in decision-making. Since giving up an endowment has a higher potential of causing future regret than not obtaining the endowment, loss aversion may be the result of a regret-avoidance strategy (cf. Gilovich and Medvec 1995; Landman 1987). The desire for consistency (cf. Cialdini et al. 1995) comes into play after the endowment is assigned and a sense of entitlement is established. Being confronted with the contradictory idea to sell the endowment might cause cognitive dissonance, i.e., excessive mental stress or discomfort (Festinger 1957). In this regard, when a seller increases the reservation price (lower WTA), this can be regarded as an attempt to reduce the dissonance created by selling something that had just been acquired and associated with oneself.

Martinez et al. (2011), Lin et al. (2006), and Zhang and Fishbach (2005) have identified

emotions as a moderating variable on the magnitude of the endowment effect. In their studies, when people are induced to positive emotional states rather than negative emotional states, the WTA/WTP disparity is bigger.

Finally, there is some discussion to what extent ownership itself may account for the disparity between the WTP and the WTA or, alternatively, whether tangent explanations such as loss aversion are involved. On the basis of their analysis and experiments, Plott and Zeiler (2005, 2007) argue that ownership in and of itself does not induce a gap between WTA and WTP. However, Morewedge et al. (2009) disentangle ownership from loss aversion in a series of experiments where brokers buy and sell mugs that they do not own. The endowment effect was observed to disappear when buyers were owners and when sellers were not, suggesting that ownership played a role in causing the WTA/WTP difference. One interpretation is that ownership, by way of association and self-identification with the object in possession, may induce endowment effects.

Implications for the Economic Analysis of Law

It has been noted that the endowment effect is “the most significant single finding from behavioral economics for legal analysis to date” (Korobkin 2003, p. 1227). Although the endowment effect is relevant in many areas of law (for a review of the various legal applications of the endowment effect in the law review literature to date, see Klass and Zeiler 2013), the implications for the economic analysis of property law and contract law are especially noteworthy.

Economic Analysis of Property Law: Initial Allocation of Property Rights, Coasian Bargaining, and Remedies

The endowment effect and the associated behavior anomalies of loss aversion and status quo bias have implications for the Coase Theorem. In the presence of endowment effects, the initial assignment of property rights may affect the final outcome, even if transaction costs are insignificant (Korobkin and Ulen 2000). By making a right

holder more reluctant to trade away the item in an efficient transaction, the endowment effect complicates the efficient reallocation of property rights through bargaining. This alters some of the traditional normative implications in the economic analysis of property rights. First, policy-makers might consider allocating property rights to the highest-valuing user – even if transaction costs are low. Second, the endowment effect increases the appeal of damage remedies as compared to property rule protection. If parties who litigate contested entitlements and who are awarded injunctive relief by a court are prone to exhibit endowment effects, they may be particularly unlikely to bargain away their court-approved entitlement, even when the opposing party places a higher value on it (cf. Jolls et al. 1998). Indeed, a series of survey experiments by Rachlinsky and Jourden (1998) indicates that injunctive remedies induce an endowment effect by way of ownership. For instance, when considering the protection of an environmental resource, participants were more insistent on protecting the conservation of the resource than was the case when the same environmental resource had liability rule protection. In contrast to injunctive relief, damage remedies allow higher-valuing individuals to take the entitlement and then pay the market price via the remedy (Korobkin and Ulen 2000). Thus damage remedies may circumvent endowment effects that would otherwise complicate an efficient allocation of rights.

Economic Analysis of Contract Law: Default Versus Mandatory Rules, Explicit and Implicit Opt-in and Opt-out Mechanisms, Standard Form Contracts, and Remedies

First and foremost, the status quo bias interpretation of the endowment effect has implications for the regulation of contract default rules. The endowment effect helps explain the observed “stickiness” of legal default. Even if parties would not likely select an undesirable contract provision, they often neglect to remove this provision when it is assigned to them by default rule. Contracting around undesirable defaults appears to be difficult even when transaction costs are low enough to make it efficient to do so (cf. Korobkin 1998a).

Likewise, the legal endowment effect helps explain why individuals find it hard to opt out of defaults (Johnson et al. 1993). In an experiment, two groups of participants had to choose between alternative automobile insurance policies. One group was presented with a default involving a cheaper policy that restricted the right to sue for pain and suffering resulting from a car accident. Participants in this group were offered the option to instead acquire the full right to sue at an 11% increase of the premium (opt-in). Another group was given a default insurance policy that contained no initial restriction on the right to sue. At the same time, however, participants in this group had the option to forego the right to sue in exchange for a reduction of their insurance premium equivalent to the increase of premiums presented to group one (opt-out). Although standard economic reasoning would predict that the choice between the two options should be the same, the applicable default affected the selection of the policy by participants. While only 23% of participants of the first group elected to opt in to acquire the full right to sue, 53% of the second group elected to retain the full right to sue. Moreover, participants with the restricted right default were willing to pay an 8% average increase of the premium to acquire the full right. When the full right was already the default, however, participants indicated an average willingness to pay 32% more for full coverage than for limited coverage. Overall then, the applicable default affected the preferences of participants in both groups. Note in this regard that also mandatory rules – like withdrawal rights in contract law – can lead to implicit opt-in or opt-out scenarios (Hoeppner 2012). Based on these findings, the endowment effect deserves consideration whenever rule-makers contemplate making use of default rules and opt-out or opt-in mechanisms.

Second, terms and provisions embodied in standard form contracts may likewise serve as reference points for contracting parties in negotiating and drafting contracts. If standard contract form an expectation baseline, they may trigger loss aversion. As a result, even when provided the opportunity to bargain for more efficient terms, contract terms may be biased toward the provided standard forms (cf. Kahan and Klausner

1996). A study by Korobkin (1998b) confirms that when default terms and preexisting form terms offered potential conflicting reference points, negotiation outcomes were biased toward the terms provided in the standard form contracts.

Third, endowment effects have been found to affect contract breach and enforcement decision by way of the applicable remedy. In a laboratory experiment, Depoorter and Tontrup (2012) observed that participants were more likely to reject efficient contract breaches and enforce contracts – even if it is against their financial interests – when the legal default is specific performance, as opposed to damages. In other words, possession of one remedy over the other framed the moral intuition such that the performance of the original contract was experienced more severely to contract promises.

Overall, endowment effects suggest that the statutory framework of contract law is likely to affect the substantive content of contracts even if contract law is based on the notion of private autonomy (freedom of contract).

Conclusion

The endowment effect has been heralded as one of the most important findings of behavioral research (Korobkin 2003). As the empirical findings, methodology and theoretical explanation remain a topic of intense discussion among economists and psychologists, the concept has already been widely applied in both law and economics and mainstream legal scholarship (cf. Klass and Zeiler 2013). As the premises and foundations of the endowment effect continue to be explored, the concept will likely find its way in both theoretical and empirical economic analysis of law.

Cross-References

- [Behavioral Law and Economics](#)
- [Bounded Rationality](#)
- [Coase Theorem](#)
- [Consumer Bias](#)
- [Experimental Law and Economics](#)

References

- Arlen J, Talley E (2008) Introduction: experimental law and economics. In: Arlen J, Talley E (eds) *Experimental law and economics*. Edward Elgar, Northampton, pp xv–lxi
- Bar-Hillel M, Neter E (1996) Why are people reluctant to exchange lottery tickets? *J Pers Soc Psychol* 70:17–27
- Brown TC (2005) Loss aversion without the endowment effect, and other explanations for the WTA-WTP disparity. *J Econ Behav Organ* 57:367–379
- Cialdini RB, Trost MR, Newsom JT (1995) Preference for consistency: the development of a valid measure and the discovery of surprising behavioral implications. *J Pers Soc Psychol* 69:318–328
- Coursey DL, Hovis JL, Schultze WD (1987) The disparity between willingness to accept and willingness to pay measures of value. *Q J Econ* 102:670–690
- Depoorter B, Tontrup S (2012) How law frames moral intuitions: the expressive effect of specific performance. *Ariz Law Rev* 54:673–717
- Dubourg WR, Jones-Lee MW, Loomes G (1994) Imprecise preferences and the WTP-WTA disparity. *J Risk Uncertainty* 9:115–133
- Festinger L (1957) *A theory of cognitive dissonance*. Stanford University Press, Stanford
- Gilovich T, Medvec VH (1995) The experience of regret: what, when, and why. *Psychol Rev* 102:379–395
- Harbaugh WT, Krause K, Vesterlund L (2001) Are adults better behaved than children? Age, experience, and the endowment effect. *Econ Lett* 70:175–181
- Hartman R, Doane MJ, Woo CK (1991) Consumer rationality and the status quo. *Q J Econ* 106:141–162
- Hoepfner S (2012) The unintended consequence of doorstep consumer protection: surprise, reciprocity, and consistency. *Eur J Law Econ* 38:247–276
- Horowitz JK, McConnell KE (2002) A review of WTA/WTP studies. *J Environ Econ Manag* 44:426–447
- Isoni A, Loomes G, Sugden R (2011) The willingness to pay-willingness to accept gap, the “endowment effect”, subject misconceptions, and experimental procedures for eliciting valuations: comment. *Am Econ Rev* 101:991–1011
- Issacharoff S (1998) Can there be a behavioral law and economics? *Vanderbilt Law Rev* 91:1729–1745
- Johnson EJ, Hershey J, Meszaros J, Kunreuther H (1993) Framing, probability distortions, and insurance decisions. *J Risk Uncertainty* 7:35–51
- Johnson EJ, Haubl G, Keinan A (2007) Aspects of endowment: a query theory account of loss aversion for simple objects. *J Exp Psychol Learn Mem Cogn* 33:461–474
- Jolls C, Sunstein CR, Thaler RH (1998) A behavioral approach to law and economics. *Stanford Law Rev* 50:1471–1550
- Kahan M, Klausner M (1996) Path dependence in corporate contracting: increasing returns, herd behavior and cognitive biases. *Wash Univ Law Q* 74:347–366

- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47:263–292
- Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the coase theorem. *J Polit Econ* 98:1325–1348
- Kahneman D, Knetsch JL, Thaler RH (1991) The endowment effect, loss aversion, and status quo bias. *J Econ Perspect* 5:193–206
- Klass G, Zeiler K (2013) Against endowment theory: experimental economics and legal scholarship. *UCLA Law Rev* 61:2–64
- Knetsch JL (1989) Endowment effect and evidence of nonreversible indifference curves. *Am Econ Rev* 79:1277–1284
- Knetsch JL, Sinden JA (1984) Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. *Q J Econ* 99:507–521
- Knetsch JL, Sinden JA (1987) The persistence of evaluation disparities. *Q J Econ* 102:691–695
- Knetsch JL, Wong WK (2009) The endowment effect and the reference state: evidence and manipulations. *J Econ Behav Organ* 71:407–413
- Knez P, Smith VL, Williams AW (1985) Individual rationality, market rationality, and value estimation. *Am Econ Rev* 75:397–402
- Korobkin RB (1998a) The status quo bias and contract default rules. *Cornell Law Rev* 83:608–687
- Korobkin RB (1998b) Inertia and preference in contract negotiation: the psychological power of default rules and form terms. *Vanderbilt Law Rev* 51:1583–1651
- Korobkin RB (2003) The endowment effect and legal analysis. *Northwestern Univ Law Rev* 97:1227–1293
- Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. *Calif Law Rev* 88:1051–1144
- Landman J (1987) Regret and elation following action and inaction: affective responses to positive versus negative outcomes. *Pers Soc Psychol Bull* 13:524–536
- Lin CH, Chuang SC, Kao DT, Kung CY (2006) The role of emotions in the endowment effect. *J Econ Psychol* 27:589–597
- List JA (2003) Does market experience eliminate market anomalies? *Q J Econ* 118:47–71
- List JA (2004) Neoclassical theory versus prospect theory: evidence from the market place. *Econometrica* 72:615–625
- Loewenstein GF, Issacharoff S (1994) Source dependence in the evaluation of objects. *J Behav Decis Mak* 7:157–168
- Martinez LF, Zeelenberg M, Rijsman JB (2011) Regret, disappointment and the endowment effect. *J Econ Psychol* 32:962–968
- Morewedge CK, Shu LL, Gilbert DT, Wilson DT (2009) Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect. *J Exp Soc Psychol* 45:947–951
- Nayakankuppam D, Mishra H (2005) The endowment effect: rose-tinted and dark-tinted glasses. *J Consum Res* 32:390–395
- Plott CR, Zeiler K (2005) The willingness to pay-willingness to accept gap, the “endowment effect”, subject misconceptions, and experimental procedures for eliciting valuations. *Am Econ Rev* 95:530–545
- Plott CR, Zeiler K (2007) Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory? *Am Econ Rev* 97:1449–1466
- Plott CR, Zeiler K (2011) The willingness to pay-willingness to accept gap, the “endowment effect”, subject misconceptions, and experimental procedures for eliciting valuations: reply. *Am Econ Rev* 101:1012–1028
- Rachlinsky JJ, Jourden F (1998) Remedies and the psychology of ownership. *Vanderbilt Law Rev* 51:1541–1582
- Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. *J Risk Uncertainty* 1:7–59
- Shogren JF, Shin SY, Hayes DJ, Kliebenstein JB (1994) Resolving differences in willingness to pay and willingness to accept. *Am Econ Rev* 84:255–270
- Thaler RH (1980) Toward a positive theory of consumer choice. *J Econ Behav Organ* 1:39–60
- van Dijk E, van Knippenberg D (1996) Buying and selling exchange goods: loss aversion and the endowment effect. *J Econ Psychol* 17:517–524
- van Dijk E, van Knippenberg D (1998) Trading wine: on the endowment effect, loss aversion, and the comparability of consumer goods. *J Econ Psychol* 19:485–495
- Zhang Y, Fishbach A (2005) The role of anticipated emotions in the endowment effect. *J Consum Psychol* 15:316–324

Energy Governance: EU-Russia Gas Exchanges

C. Locatelli

GAEL, CNRS, Grenoble, INP, INRA, University of Grenoble-Alpes, Grenoble, France

Definition

Strong and lasting interdependencies exist between EU and Russia concerning natural gas exchanges. They imply some questions about energy security from both parties. One could have expected the setting up of more or less institutionalized governance structures – as an institutional system of rules – allowing the management of risks and externalities linked to this interdependence.

On both sides there is a willingness to cooperate. However, to date, the institutional gap between the supply and demand of the cooperation is a constraint to define a governance structure. The question is to know if international standards based on rules generated by the EU are consistent with Russia's institutional environment. The competitive logic and the regulation on which the EU energy policy is founded conflict with the institutional specificities of the Russian economy. This contradiction explains the failure in the European strategy of external governance with Russia. The "gas conflicts" between the EU and Russia can be viewed as the result of the confrontation of different models that structure the natural gas industries of these two countries.

Through this study case, the question of the internationalization of the rules, norms, and more generally an institutional system of regulation is clearly posed.

Introduction

The determinant factors of the relation between Russia and the EU concerning natural gas exchanges are the subject of an impressive literature. The strong and lasting interdependencies that exist between the two zones imply some questions about energy security from both parties. One could have expected the setting up of more or less institutionalized governance structures allowing the management of risks and externalities linked to this interdependence (Escribano 2015). This was the aim of the energy partnership and its prolongation in a strategic partnership, the Energy Charter – the first constraining multilateral treatise in the field of energy, notably in the matter of investments (Wälde 2008; Kustova 2016). It was an attempt to institutionalize a cooperative zone at the level of EU-Russia to respond to the stakes of energy security. However, their painful implementation bears witness to the constraints that weigh on the definition of this zone of cooperation, notably because of strong divergences concerning the preferences of those involved in this relation. The failures in this domain are an

illustration of the problems linked to all attempts to internationalize the norms, rules, and regulations defined in agreement with the national institutional framework.

Interdependences in the Natural Gas Exchanges Between Russia and the EU: Different Preferences and Interests

The interdependences in the natural gas exchanges between the EU and Russia are not questionable. The Russian gas company, Gazprom, is the sole Russian gas exporter to Europe because of its monopoly on exportations by pipelines. It supplies 30% of the EU imports. Since the 1990s, this represents more than 100 Gm³/year of natural gas exported. At the same time, the characteristics of the gas networks (specific assets, non-re-deployable assets according to Williamson's transaction cost theory 1985) founded the interdependence. Thus, Europe (70.8% of total Russian gas exports) is the first and foremost export market for Russia. Because of the infrastructures in place, in the short term, Russia is dependent on this market and cannot reorient its exports toward other destinations. This dependency is especially important for the State since the revenue linked to hydrocarbon exports (oil and gas) is an essential factor of the budget equilibrium and of economic growth.

In this relation of interdependence, the position of the Russian state (interests and preferences) is defined by its role as a hydrocarbon supplier (Romanova 2014). Its main aim is to have stable markets in terms of the quantities exported and the levels of prices. Therefore, it is to secure access to its main export market, the EU (Kratochvil and Tichy 2013). A double aim is imposed on Gazprom: to maintain its market share in the European gas market (the strategic aim of the State, its principal shareholder) and to maximize its income from exports (Boussena and Locatelli 2016; Yang et al. 2016). The result is a concept of demand security. For the EU, given its position as a major importer (the largest at the worldwide level), the question is one of securing its supply in terms of both price and volume. This defines a

concept of supply security in a specific context, a competitive natural gas market.

Perception of Risks Linked to Interdependence

Since the beginning of the 2000s, the perception of a natural gas security problem has emerged both in the EU and in Russia. It is linked to interdependence between the two zones. According to the analysis proposed by Cherp and Jewell (2014), problems of energy security derive from a price risk and a volume risk.

EU: The Risks Linked to the Gazprom Market Power

In the process of liberalization, taking into account the oligopolistic structure of the gas supply, competitive markets are an important condition of the supply security. One of the primary aims of the EU Energy Policy is to limit the market power of all dominant companies, i.e., those capable of distorting competition. From this point of view, Gazprom, which supplies more than 30% of the external requirement of the EU or over 60% in central Europe and in the Balkan countries, represents a specific risk in the eyes of the EU. This is especially important as more than 50% of Gazprom's shareholding is in the hands of the State. Furthermore, the different gas crises between Russia and Ukraine, and the current conflict, pose the question of reliable transit for Russian gas exports to Europe.

Russia: Uncertainties of the European Gas Market

Since the end of the 2000s, the EU gas market has become more volatile and uncertain, notably for its traditional suppliers. The process of liberalization of the gas industries, the surplus supply, and the weakness of the demand linked to the economic crisis, but also to European policies in favor of renewable energies and the drastic drop in the price of oil, have had serious effects. Firstly these factors all led to the emergence of more competitive gas markets, and secondly they resulted in a drop in the price of natural gas of

more than 50% since 2012. These factors are susceptible to put Gazprom's market share in Europe into question. This tendency will continue. In the medium-short term, new LNG suppliers (like the USA, Australia) and uncertainties concerning demand (economic growth, climate policy) mean that the EU is a risky market for Russia and for Gazprom.

These evolutions have led Gazprom to significantly modify its export policy and to adopt strategies that will be different from its traditional contractual framework. Its attachment to long-term take-or-pay (TOP) contracts is known. These contracts have organized its sales to Europe. They have allowed a share in the risks of the price and volumes between the buyer and the seller, notably by guaranteeing the quantities sold annually (removal clause), and guaranteed the development of the gas fields and gas pipelines to Europe. But in the current phase of volatile and uncertain gas markets, Gazprom found itself obliged to reassess this contractual logic. To answer competitive pressures, it had to proceed with more or less major modifications of certain clauses of its long-term TOP contracts (Boussena and Locatelli 2016).

Stakes and Failures of the Governance Structure for Gas Exchanges Between the EU and Russia

The risks linked to this interdependence asked questions about the definition of a governance structure as an institutional system of rules (Lavenex and Schimmelfenning 2009). The primary aim should be the possibility, through cooperation, to limit the extent of conflicts and to stabilize the relations of exchanges between the two zones. On both sides there is a willingness to cooperate. The different attempts to build partnerships, the Energy Charter proposed by the EU or the "Conceptual Approach to the New Legal Framework for Energy Cooperation" developed by A. Medvedev in response to the Charter are some examples. Nevertheless, the cooperation implies some convergence of the actors' preferences. However, to date, the institutional gap

between the supply and demand of the cooperation is a constraint to define a governance structure as demonstrated by the withdrawal of Russia from the process of the ratification of the Energy Charter treaty.

The EU Offer of Cooperation: Building the Markets by the Exporting the Competitive Model

The EU has attempted to manage its relation of dependence with the gas suppliers by the export of its liberal organizational model: governance by competition and EU energy regulation and norms. This form of external governance is an answer to the stakes of security and interdependence through the setting up of a system of common rules and of a high level of institutionalization (Lavenex and Schimmelfenning 2009). In particular, the objective is to overcome the incomplete nature of the competitive reforms concerning the European gas industries. In this way, it is a means to increase the efficiency and the problem-solving capacity of internal liberalization policy (Lavenex 2004). The process of liberalization necessitates opening up upstream of the gas producers and exporting the competitive rules of the EU gas markets. Particularly, the emergence of several gas companies in Russia following the disintegration of Gazprom, potential exporters and therefore in competition on the EU market, would be susceptible to increase the liquidity of the spot markets and render them credible. Next, the aim is to answer the *market failures* that stem notably from an imperfect competition, from the dominant position of certain companies and asymmetric information.

This is the philosophy of the Energy Charter. It is an attempt to build a supranational economic and political integration, to institutionalize an international market system (Andersen and Sitter 2016). Furthermore, it offers guarantees for international investments and allows a competitive principle of nondiscrimination for access to hydrocarbon resources (Haghighi 2007). This access represents a key stake for the “EU gas security” (Boussena and Locatelli 2013). Additionally, the transformation of energy governance of its main suppliers is the essential aim of the EU

energy security strategy (Keating 2012). Finally, on the EU market, the gas supply companies behave in compliance with the EU regulation. In this framework, the EU is capable of “normalizing” their behavior. This defines a power of the EU market according to the terminology of Damro (2012) and Goldthau and Siter (2015).

The Institutional Hiatus: Coherence and Complementarities of the EU Rules with the Russian Institutional Environment

The competitive logic and the regulation on which the EU energy policy is founded conflict with the institutional specificities of the Russian economy. This contradiction explains the failure in the European strategy of external governance with Russia. The questions of coherences and complementarities (Aoki 2001; Amable 2016) of organizational models issued from the economic practices of Western markets with the Russian institutional environment and of their transfer as such remains at the heart of analyses of Russian economic reforms (Hausner 1995; Stiglitz 1999; Murell 2001; Roland 2000; Locatelli and Finon 2003). According to this neo-institutionalist approach EU rules must be coherent or complementary with the Russian institutional and economic environment. It is not the case in the gas sector.

The Russian gas market is characterized firstly by the presence of a State company that is vertically integrated with a monopoly of transport and exports (Gazprom). Some competitive fringes exist on larger and larger segments of the Russian market. This dual market depends on dual prices (administrated and free) and on a strict control by the State of the access to hydrocarbon resources. This reform is that which is compatible with the Russian institutional environment. This latter, marked by the weakness of the market institutions (weakness of the *Rule of Law*, ownership rights that remain poorly defined, regressive fiscal system etc.), seems incompatible with the competitive model and de-integrated organization of the EU gas industries (Locatelli 2014).

This path of reform, and the contestation by the Russian state of the EU nominative power

(Godzimirski 2015; Newmann and Posner 2015), results in increasing divergences in the preferences and interests of these two actors. Their divergent approach to the organization of the gas industries is at the origin of major regulation conflicts. In this context, it is increasingly hard to define and to put in place a governance structure that will be susceptible to manage the gas security problem. Through this study case, the question of the internationalization of the rules, norms, and more generally an institutional system of regulation is clearly posed.

Cross-References

- [Conflict of Laws](#)
- [Economic Integration](#)
- [Institutional Complementarity](#)
- [Institutional Economics](#)

References

- Amable B (2016) Institutional complementarities in the dynamic comparative analysis capitalism. *J Inst Econ* 12(1):79–103
- Andersen S, Sitter N (2016) The external reach of the EU regulatory state: Norway, Russia and the security of natural gas supplies. In: Peters I (ed) *The European Union's foreign policy in comparative perspective beyond the 'actorness and power' debate*. Routledge, Contemporary European Studies, Oxford, pp 80–97
- Aoki M (2001) *Toward a Comparative Institutional Analysis* MIT Press, Cambridge; London, 560 p
- Boussena S, Locatelli C (2013) Energy institutional and organisational changes in EU and Russia: revisiting gas relations. *Energy Policy* 55:180–189
- Boussena S, Locatelli C (2016) Guerre des prix ou instrumentalisation de l'incertitude sur les prix: quelle stratégie pour un fournisseur dominant sur le marché gazier européen ? *Cahier de recherche EDDEN*, 1
- Cherp A, Jewell J (2014) The concept of energy security: beyond the four As. *Energy Policy* 75:415–421
- Damro C (2012) Market power in Europe. *J Eur Publ Policy* 19(5):682–699
- Escribano G (2015) Fragmented energy governance and the provision of global public goods. *Global Pol* 6(2):97–106
- Godzimirski J (2015) Russia-EU energy relations : from complementarity to distrust? In: Godzimirski J (ed) *EU Leadership in energy and environmental governance: global and local challenges and responses*. Palgrave Macmillan, London, pp 89–112
- Goldthau A, Sitter N (2015) Soft power with a hard edge: EU policy tools and energy security. *Rev Int Polit Econ* 22(5):941–965
- Haghighi S (2007) *Energy security: the external legal relations of the European Union with major oil- and gas-supplying countries*. Hart, Oxford
- Hausner J (1995) Imperative vs interactive strategy of systemic change in central and Eastern Europe. *Rev Int Polit Econ* 2:249–266
- Keating M (2012) Re-thinking EU Energy Security: The Utility of Global Best practices for Successful Transnational Energy Governance. in Kuzemko C, Belyi A, Goldthau A, Keating M. (eds) *Dynamics of Energy Governance in Europe and Russia* (Palgrave MacMillan), pp. 86–105
- Kratochvil P, Tichy L (2013) EU and Russian discourse on energy relations. *Energy Policy* 56:391–406
- Kustova I (2016) A treaty à la carte? Some reflections on the modernization of the energy charter process. *J World Energy Law Bus* 9:357–369
- Lavanex S (2004) EU external governance in 'wider Europe'. *J Eur Publ Policy* 11(4):681–700
- Lavanex S, Schimmelfenning F (2009) EU rules beyond EU borders: theorizing external governance in European politics. *J Eur Public Policy* 791–818
- Locatelli C (2014) The Russian gas industry: challenges to the 'Gazprom model'? *Post Communist Econ* 26(1): 53–66
- Locatelli C, Finon D (2003) Les limites à l'introduction des institutions de marché dans un secteur de rente. *Histoire des représentations du marché: Xe colloque, Grenoble, Association Charles Gide pour l'étude de la pensée économique*, 25–27 septembre
- Murell P (ed) (2001) *Assessing the value of law in transition economies*. University of Michigan Press, Ann Arbor
- Newmann A, Posner E (2015) Putting the EU in its place: policy strategies and the global regulatory context. *J Eur Publ Policy* 22(9):1316–1335
- Roland G (2000) *Transition and economics: politics, markets, and firms*. MIT Press, Cambridge
- Romanova T (2014) Russian energy in the EU market: bolstered institutions and their effects. *Energy Policy* 74:44–55
- Stiglitz J (1999) Whither reform? Ten years of the transition. In Pleskovic B, Stiglitz J (eds) *World Bank annual bank conference in development economics*, pp 27–56
- Walde T (2008) Renegotiating acquired rights in the oil and gas industries: industry and political cycles meet the rule of law. *J World Energy Law Bus* 1(1):55–97
- Williamson O (1985) *The economic institutions of capitalism: firms, markets, relational contracting*. Free Press, New York
- Yang Z, Zhang R, Zhang Z (2016) An exploration of a strategic competition model for the European Union natural gas market. *Energy Econ* 57: 236–242

Energy Regulation

Régis Lanneau

Law school, CRDP, Université de Paris Nanterre,
Nanterre, France

Definition

Energy is a fundamental input in modern economies. Its regulation is probably one of the most complex since it is related not only to economic and development issues but also to environmental, geopolitical, and social ones. This entry is using law and economics to inquire into energy regulation, to show the rationale of its regulation and the contemporary challenges it has to face.

Introduction

Man is probably the only animal which is using energy not only to maintain his bodily functions at work but, and by far, for his own comfort. At the end of the eighteenth century and before the industrial revolution, the world was consuming around 250 million tons of oil equivalent (toe); nowadays this figure is reaching 13 giga toe; this dramatic increase also occurred through the development of new energy sources which led to great transitions in energy mix from renewable energy which dominated the eighteenth and nineteenth century to the raise of fossil fuels, especially coal at the end of the nineteenth and oil since the mid-twentieth century. Energy is now a fundamental – not to say a central – input in all production, it is necessary to provide most services, and it is required by both consumers and producers for heating, cooling, cooking, lighting, and using machines and household appliances and of course transportation. Modern life – as we know it – is dependent on energy consumption; so is economic growth and development.

Energy regulation is probably one of the most complex to design, implement, and assess. It is of

course a sensitive political issue since it will not merely have local consequences (for the energy sector) but also powerful systemic ones (e.g., for the economy or for the environment). Moreover, its scope is broader than most other regulation: it not only focuses on production but also addresses questions regarding transportation, distribution, storage, zoning, and use. It can also be addressed indirectly through regulations concerning mines, buildings, cars, safety standards, working conditions, and, of course, the environment. The body of rules is typically huge, poorly systematized, and often cryptic for outsiders. It is probably where law and economics could help to cut some underbrush: the framework it is providing to understand regulations is especially powerful regarding energy regulation (1). Moreover, the economic approach could also help to address some contemporary issues in energy regulation (2).

Before exploring these two topics, it is noteworthy to mention that the idea of energy regulation in the European Union context is not as broad as what the word refers to in the “common language.” Indeed, only natural gas and electricity are concerned, oil benefiting from a special treatment. This could easily be explained using an economic approach of these regulations (infra; 1.1; see also Viscusi et al. 2001), but this restrictive understanding is not helping to stress the fact that energy regulation cannot be addressed energy by energy; it would be better to conceptualize it as an ecosystem in which energy are interlinked.

Why Regulate?

Economic analysis is providing a traditional rationale for regulating: market failures. In the case of energy regulation, it allows for a unifying approach of all regulations which is often missing. The economic rationale is of course not the only one. Political rationales are also powerful considering the central function of energy in modern life.

The Economic Rationale

From an economic point of view, regulations are required if markets are revealing “market failures”

(externalities, market power, imperfect information). This market failure approach to energy regulation reveals itself quite powerful to systematize existing regulations and could explain the “special treatment” reserved to electricity and natural gas. Imperfect information regulation being not fundamental to energy regulation will not be considered here.

Externalities

Producing and consuming energy entail external costs which are not fully considered by rational economic agents either because transaction costs are too high to internalize them through a market or because property rights are not well defined (which of course complexify the analysis to pinpoint “real” externalities). Burning coal, gas, or oil to produce electricity leads to the production of carbon dioxide, sulfur dioxide, and other pollutants; using uranium to fuel a nuclear plant creates radioactive waste. If discharging pollutants (or emitting waste) in the atmosphere is not perceived as a cost, overproduction is to be expected. Even worse, in such system, producers will not have the right incentive to substitute one source of production for another because they will only consider their private costs and not the social cost of their activity.

These externalities are probably the most famous, but many other sources exist:

1. Storage safety (e.g., of oil or gas)
2. Location (e.g., of plants, pipelines, transformer substations)
3. Transmission (because of the physical characteristics of electricity, a coordination is required between producers to power the grid; interconnection of networks also required some coordination)
4. Transportation (because energy fuels are hazardous materials)

Externalities in the energy sector are often addressed through environmental, safety, and zoning regulation. The label should not confuse the lawyer or the economists; these are ways to regulate energy.

Market Power

It is at this level that a “labeled” energy regulation exists. If it is true that market power could also be – and is – addressed *ex post* through competition law (e.g., through the doctrine of essential facilities), some specific characteristics of gas and electricity were considered to require specific forms of regulations.

First, because of their network structure requiring substantial physical infrastructure, this sector often exhibits the characteristic of “natural monopolies” at least for part of these infrastructures (inefficient to duplicate) and not necessarily for production or distribution. However, and before restructuring policies introduced in the 1990s in Europe, it was quite common to face a fully vertically integrated entity responsible for production, transmission, and distribution. Since the market discipline is not effective for natural monopoly, nothing is incentivizing the monopoly to improve its performance or services (for an alternative view and more development on the traditional approach, see Posner 1999). Because of this situation, nationalization and price regulation were often used with all their practical difficulties (for a detail of price regulation technics, see Decker 2015; see also Viscusi et al. 2001).

Second, with the restructuring policies of the 1990s (directives n^{os} 96/92/CE and 98/30/CE for gas and electricity in Europe), allowed by technology improvements, an independent system operator in charge of maintaining and operating the network typically appeared. This operator, quite often a non-profit entity, was and still is often supervised by an independent regulatory agency. These policies were supposed to introduce market discipline with the expected benefits of lowering prices, controlling costs of construction and stimulate innovation (Decker 2015). Their success is mixed (for electricity, see Pollitt 2007).

The Political Rationale

From a political point of view, two main rationales are often mentioned. First, it is required to insure some form of energy independence. Second, territorial cohesion and social justice consideration are also playing a key role in energy regulation.

Energy Independence and Security of Energy Supply

Since energy is the lifeblood of economic systems and especially crucial for defense matter (especially transport of troops and dissuasion), geopolitical consideration often entered in energy regulation. Being, like Ukraine, at the mercy of an energy supplier is perceived as a real geopolitical risk; moreover, limiting risks of an economic paralysis due to a price shock or preventing supply disruption are also considered by energy regulations. The development of the nuclear industry in France proceeds from this logic; the General De Gaulle wanted to insure an energy independence of France. This logic also exists at the level of the European Union. Article 194 TFEU especially mentioned the “security of energy supply” (see also Helm 2002). Council directive 2009/119/EC is “imposing an obligation on Member States to maintain minimum stocks of crude oil and/or petroleum products” (typically 90 days of the average daily consumption of a country).

This security of energy supply could lead to specific regulation in hospital and other public services to make sure that the service will not be interrupted.

Territorial Cohesion and Social Justice

Energy being central to modern life, its provision is often considered a matter of public interest. Reliable access to energy in all parts of the territory is considered as an aim in itself, and the regulators could impose some obligations network operator to make sure that basic provision of energy (electricity especially) is ensured and all parts of the country or a regional zone (typically eastern countries in Europe) are connected. In general, these obligations are leading to compensation (which should not provide an advantage or a disadvantage to the addressee of these obligations). This type of regulation typically does not exist for oil because of its physical characteristics (it can be transported and stored) and relatively secondary importance for basic energy needs.

As a basic need, social fares are also quite common for part of the consumption of energy (the one considered as a “need”). Since these social fares result from a political will, providers

are compensated for the difference between the social fare and the regular one. Moreover, some regulations are introduced to limit the possibility of suppliers to withdraw access to “poor” families unable to pay their bills. Once again, a compensation should then be provided.

Contemporary Issues in Energy Regulation

Energy regulations are still facing numerous challenges from pushing toward a new energy mix to conciliating development and climate change and adapting infrastructures to new needs. It would also have been possible to mention the difficulties of achieving efficient regulations in a domain where lobbying and rent-seeking are considered as a norm. In this section, it will not be possible to address all these issues. Two important challenges will only be quickly considered: first, insuring a real competition between energy producers, so that energy will be supplied by efficient methods, and second, regulations are now trying to promote energy efficiency and are more and more addressing not only suppliers but also consumers.

The Challenge of Insuring Real Competition Between Energy Producers: A Way to Develop Green Energies?

The challenge of insuring an effective competition does not only exist between producers but also among producers of a certain type of energy. We already noted that, in the domain of electricity (for both retail and generation, Decker 2015), restructuring policies achieved mixed results. Ensuring the effectiveness of competition law would already be a first step, and regulated prices in this domain should probably be removed following this logic.

However, the most significant challenge is to ensure an effective competition between energy suppliers, a competition based not on private costs but on social costs. Typically, when regulations are not trying to handle negative externalities resulting from some producers and are also not compensating some others for the positive externalities they are generating, prices are no longer

playing their coordination function. Some costlier energies appear cheaper for consumers which will then not adopt optimal behavior. The energy mix transition (from fossil fuel to renewable energies) which is required by climate change is then stifled (OECD 2012). Even worse, according to OECD, support for the production and consumption of fossil fuel amounted to about USD 55–90 billion every year during the period 2005–2011 for OECD countries (OECD 2013). If such a policy could be easily explained through public choice, it appears clearly inefficient from a macroeconomic point of view. Removing these support measure would certainly spur innovation in and consumption of green energy by lowering incentives to produce and use fossil fuels. It would also allow to reduce and improve the efficiency of subsidies to green energy and green technologies (Harris 2006). Note that it would also be required to compute the environmental cost of each and every technology to assess the “real” price which, considering valuation methodology (see Markandya et al. 2002), will lead to lots of contention.

Because of some form of path dependency, it is nevertheless not certain that such radical evolution of regulation is to be expected in the near future or will happen smoothly (Unruh 2002).

Energy Efficiency and Regulation of Its Use

If consumers and producers were entirely rational, they would switch to energy-efficient technology when this technology is also cost effective. If it is not the case, they will not necessarily invest in energy-efficient machinery, appliances, cars, or even buildings. This situation could arise because of implicit discount rate. The consumer will choose an appliance or a car or consider some improvement of his house if and only if she is perceiving it as a good investment. Nevertheless, energy-efficient technologies are often costlier than other technologies, and integrating the private future benefits on energy costs is not sufficient to lead consumer to make such a choice. Indeed, these benefits should be discounted at their present value, and if the implicit discount rate is too high compared to a “normal” discount rate, consumers will not adopt an optimal

behavior. Imperfect information could also explain this situation. From this point of view, building codes and fuel efficiency standards could make sense if and only if they are designed adequately. A less paternalist approach would lead to efficiency labeling which is not without its own side effects.

This type of regulation has only indirect effects on energy generation or distribution but could lower the demand for energy and hence the externalities generated (see Mitchell and Woodman 2010). Considering the difficulties to achieve an optimal regulation relative to energy production, this option should not be disregarded.

Conclusion

From a law and economics point of view, energy regulation should be addressed with the idea that what is to be regulated is a full ecosystem and not some sector-specific problems. The market failure approach of regulation provides an interesting starting point to describe and criticize energy regulations. It also enlightens the fact that in this domain of essential complexity, information requirement for an efficient regulation is gigantic. It is nevertheless well accepted that direct subsidies to fossil fuel should be avoided.

Cross-References

- [Efficient Market](#)
- [Energy Governance: EU-Russia Gas Exchanges](#)
- [Market Failure: Analysis](#)
- [Public Goods](#)
- [Public Interest](#)

References

- Decker C (2015) Modern economic regulation. Cambridge University Press, Cambridge
- Harris J (2006) Environmental and natural resource economics, a contemporary approach. Houghton Mifflin Company, New York
- Helm D (2002) Energy policy, security of supply, sustainability and competition. *Energy Policy* 30:173

- Markandya A, Harou P, Bellù L, Cistulli V (2002) Environmental economics for sustainable growth, a handbook for Practitioners. Edward Elgar Publishing, Cheltenham
- Mitchell C, Woodman B (2010) Regulation and sustainable energy systems. In: Baldwin R, Cave M, Lodge M (eds) The Oxford handbook of regulation. Oxford University Press, Oxford
- OECD (2012) OECD environmental outlook to 2050. OECD Publishing, Paris
- OECD (2013) Inventory of estimated budgetary support and tax expenditures for fossil fuels – 2013. OECD Publishing, Paris
- Pollitt M (2007) Liberalization and regulation in electricity systems: how can we get the balance right? EPRG working paper, no. 0724. Available online: <https://www.repository.cam.ac.uk/handle/1810/194737>
- Posner R (1999) Natural monopoly and its regulation. CATO institute, Washington, DC
- Unruh G (2002) Escaping carbon lock-in. Energy Policy 36:4299
- Viscusi K, Vernon J, Harrington J (2001) Economics of regulation and antitrust. MIT Press, Cambridge, MA

Entrepreneurship

Marta Podemska-Mikluch
Faculty of Economics, Gustavus Adolphus
College, Saint Peter, MN, USA

Definition

While one could search in vain for a definition of entrepreneurship acceptable to all entrepreneurship scholars, dominant approaches share a number of characteristics. Typically, an entrepreneur is portrayed as an agent of change who operates in a world of incomplete and dispersed knowledge while facing uncertainty about the future. What differs in the various conceptualizations is the mechanism of overcoming uncertainty and knowledge problems. The three leading approaches portray entrepreneurs as dealing with these challenges through either alertness, creativity, or judgment, with new streams of literature adding hypothesis testing and effectuation to the list. While disagreements on the particulars of what entrepreneurship is and what role it plays in economic theory continue, economists now

agree on the strong relationship between entrepreneurship and institutions. In particular, it is now well accepted that whether entrepreneurship produces economic growth depends on the institutional context. The other side of this relationship – impact of entrepreneurship on institutions – only begins to receive attention, as does the role of entrepreneurship outside of the conventional market setting.

Introduction

Economists have yet to agree on a single definition of entrepreneurship. Thinkers interested in entrepreneurship originate from diverse perspectives and evoke entrepreneurship to solve distinctive theoretical puzzles. As a result, there is no agreement in the literature on what entrepreneurship is and what role it should play in economic theory. This being the case, anyone interested in entrepreneurship theory should not feel discouraged. The different conceptualizations of entrepreneurship share a number of characteristics. For example, in all cases, entrepreneur is portrayed as an agent of change who operates in a world of incomplete and dispersed knowledge while facing uncertainty about future. It is the mechanism of dealing with uncertainty and knowledge problems that differs in the different conceptualizations of entrepreneurship. The leading approaches portray entrepreneurs as being able to overcome these challenges due to (1) being equipped with alertness to previously unnoticed profit opportunities (Kirzner 1973), (2) being creative and having leadership skills (Schumpeter (1911) 2008, (1943) 2013; Shackle 1972), (3) exercising entrepreneurial judgment (Knight (1921) 2012; Casson 1982; Foss and Klein 2012), (4) experimentation akin to scientific hypothesis testing (Harper 1996), and (5) effectuation (Sarasvathy 2008).

Interest in entrepreneurship theory has been on the rise for the last three decades (Elkjaer 1991; Venkataraman 1997). This is true not only in economics but also in management. For example, in 1987, Entrepreneurship Division was added to the Academy of Management, and now

entrepreneurship has its own code (L26) in the Journal of Economic Literature classification (Foss and Klein 2012, 23). As can be learned from the numerous sources offering an in-depth survey of the history of entrepreneurship thought (Blaug 1998; Praag 1999; Brouwer 2002), the tradition is not new; its origins can be traced all the way back to Richard Cantillon's 1755 distinction between entrepreneur, employee, and landowner with entrepreneur being characterized as carrying the risk (Cantillon (1755) 2015). Cantillon treated the economy as an interconnected whole where coordination and adjustment were secured through competition and entrepreneurship. Even though other thinkers, for example, Jean-Baptiste Say and John Stewart Mill, followed Cantillon's footsteps and also contributed to the theory of entrepreneurship, the concept disappeared from economic theory almost completely by 1870 (Blaug 1998). Undeniably, while the interest in entrepreneurship is now growing, it is yet to be incorporated into mainstream economic theory, and the disagreements on its relevance continue (Foss and Klein 2012, 24).

Despite their status as idiosyncratic outsiders, two twentieth-century thinkers are associated with entrepreneurship in economics and management literatures: Joseph Schumpeter and Israel Kirzner. Equally well known is the difference in their portrayals of entrepreneurship: while Schumpeter sees entrepreneurship as a disruptive force that pulls the economy away from equilibrium, Kirzner views it as a coordinating, equilibrating force. Partially because neither of these theories made significant in ways into mainstream economics, Casson (1982) and Klein and Foss (2012) suggest a third approach, that of Frank Knight. To understand key aspects of each approach, and how they differ, a closer look is warranted.

Entrepreneurship as Creativity and Leadership

Schumpeter's entrepreneur is an innovator, an agent of change, and a disrupter. Entrepreneurship is the force that breaks away from the routine and

drives economic change. Schumpeter first introduced the concept of an entrepreneur as an innovator in his groundbreaking *Theory of Economic Development* (Schumpeter (1911) 2008). While Schumpeter was fascinated by Walras and by the general equilibrium theory, he found the static analysis of neoclassical economics to be incapable of explaining economic change. In response, he aspired to offer a process theory akin to that of Marx. Schumpeter described the tendency toward the equilibrium – the key concept in neoclassical economics – as “the circular flow of economic life” where nothing new ever happened. He reasoned that in a theoretical apparatus of general equilibrium, where the circular flow of production and distribution was in balance, there is no room for novelty. Only when entrepreneur is introduced, novelty enters the model. Entrepreneurs break away from the circular flow by making nonroutine choices, by arranging resources into “new combinations.” Without entrepreneurship, economy remains in a state of circular flow. So entrepreneurship is necessary in order to give an endogenous account of economic change; it is the driving force of economic development.

In explaining how entrepreneurs innovate, Schumpeter listed five distinct categories of “new combinations”: (1) introduction of a new good, this encompasses introducing a good to unfamiliar customers or creating a higher-quality good, (2) creation of a new method of production, (3) the opening of a new market, (4) the capture of new sources of supply, and (5) a new organization of industry (Schumpeter (1911) 2008, 66). Time is the essence in Schumpeter's conceptualization because time turns innovation into routine. The distinction between innovation and routine, between new and old, is the distinguishing factor between entrepreneurs and other economic agents. Once entrepreneurs build up their business and stop carrying out new combinations, they stop being entrepreneurs and turn into managers. In creating new combinations, entrepreneurs perform a function that is separate from that of capitalist, landowner, laborer, manager, or even an inventor. A manager could also be an entrepreneur but not every manager is an entrepreneur and vice versa. In addition, as can be implied from the five

categories of new combinations, Schumpeter recognized that innovation and invention differ: the former requires leadership and will, the latter requires intellect. For economic analysis the one that matters is innovation because “as long as they are not carried into practice, inventions are economically irrelevant” (Schumpeter (1911) 2008, 88).

Leadership, which manifests itself as the willingness to break away from the routine, is a key characteristic of Schumpeter’s entrepreneur. It is the entrepreneurial leadership that in Schumpeter’s framework provides the creative energy that propels societies in new directions and breaks existing structures. Therefore, to Schumpeter, entrepreneurship is the locus of leadership in a free society. As later developed in “capitalism, socialism, and democracy,” leadership and entrepreneurship are the forces behind creative destruction – a force that creates and destroy the structures of the economy in an ongoing process of innovation and change, making it the essence of economic development (Schumpeter (1943) 2013).

Entrepreneurship as Coordination

While Schumpeter’s entrepreneur is usually considered a creative and disruptive force, Kirzner’s entrepreneur is a middleman, an arbitrager. Entrepreneurship to Kirzner is a force that drives the market toward equilibrium, not away from it (Kirzner 1973). The entrepreneurial function is first and foremost that of securing coordination among market participants. In his most cited book, *Competition and Entrepreneurship*, Kirzner offers a simple definition of the “pure” entrepreneur as observing “the opportunity to sell something at a price higher than that at which he can buy it” (Kirzner 1973, 16).

Kirzner’s entrepreneur operates in a Hayekian world of dispersed knowledge, knowledge that is specific to time and place. In such a world, coordination is a puzzle: how is it possible that despite being equipped with different knowledge, individuals are able to find opportunities for mutually beneficial transactions? (Hayek 1937, 1945). Clearly, in the world of perfect knowledge, people

would be aware of the full spectrum of options that might be available to them, and there would be no errors for entrepreneurs to capture. Only in the world of incomplete knowledge, some buyers pay prices that are too high, and some sellers accept prices that are too low. By remaining alert to these errors, the entrepreneur observes the differences in prices and profits from reconciling them.

Just like Schumpeter, Kirzner too is critical of mainstream economics and offers a critique of contemporary price theory, in particular the Robinson-Chamberlin model of monopolistic competition (1973). He argues that neoclassical economists are overly committed to the equilibrium framework and pay no attention to the market process, the process through which equilibrium and coordination are generated. While the neoclassical price theory pays no attention to entrepreneurship, Kirzner argues that entrepreneurship is the driving force of the market economy. Along with other economists of the Austrian tradition, Kirzner maintains that the neoclassical approach relies on overly restrictive assumptions. In their view, prices encountered in a market economy are not equilibrium prices, but rather they should be considered emergent exchange ratios that evolve as entrepreneurs seek to discover arbitrage opportunities. Through the process of discovering arbitrage opportunities, entrepreneurs capture profits, and the economy moves toward the equilibrium. In contrast, if the process of adjustment is viewed as automatic, there is no role for the entrepreneur in the model. What is puzzling and interesting to Kirzner is assumed to happen automatically in the price theory framework.

I will thus argue that the dominant theory, by emphasizing certain features of the market to the exclusion of others, has constructed a mental picture of the market that has virtually left out a number of elements that are of critical importance to a full understanding of its operation. (Kirzner 1973, 4)

While on the surface Kirzner’s approach might appear to be completely different from that of Schumpeter, a number of authors view them as complementary (Hébert and Link 1988; Boudreaux 1994; Choi 1995). For example, Boudreaux argues that the two approaches are

complementary in that they both capture the distinct aspects of the competitive market process. According to Boudreaux, if economists adopted a broader notion of equilibrium – one that include quality adjustments and technological and organizational improvements in addition to price adjustments – then Schumpeter's entrepreneurship could also be viewed as an equilibrating mechanism (Boudreaux 1994). In his own reconsideration of the two approaches, Kirzner argues that entrepreneurs are both alert and creative but that only alertness is relevant to the analysis of the market process (Kirzner 1999). Alertness renders entrepreneurship the driving force of the market process. In contrast, boldness, creativity, and self-confidence are psychological traits. While being bold and creative might contribute to entrepreneur's success, these characteristics are not relevant to entrepreneurship's equilibrating function. It is alertness that coordinates market activities, not boldness or creativity. According to Kirzner, Schumpeter's psychological portrayal of the entrepreneur might be an accurate reflection of real-world entrepreneurs. However, these are secondary features that simply give entrepreneurs an edge, whereas alertness is foundational. Moreover, Kirzner reasserts that entrepreneurship must be viewed primarily as coordinating, not disruptive. Without entrepreneurship, it is impossible to explain market coordination.

While disagreement between Schumpeter and Kirzner received significant attention in the literature, Manne notes that a more prominent and far-reaching disagreement occurred between Kirzner and Demsetz (Manne 2014). Demsetz defended price theory and that entrepreneurship is not a separate category of market activity (Demsetz 1970). Subsequently, Tyler Cowen notes that the discussion of entrepreneurship versus optimization is alike to that between philosophy and poetry (Cowen 2003).

Entrepreneurship as Judgment

Foss and Klein abandon the discussion of creativity versus alertness and instead suggest that Knight's conception of entrepreneurship, as

judgmental decision-making under uncertainty, provides a better explanation of the entrepreneurial function and can be most smoothly integrated with the economic literature on the firm (Foss and Klein 2012). The authors define judgment as "decisive action about the deployment of economic resources when outcomes cannot be predicted according to known probabilities" (Foss and Klein 2012, 38). They view judgment as different from alertness because alertness tends to be passive, whereas judgment is active. As the authors explain, alertness is the ability to react to existing opportunities, while judgment refers to the creation of new opportunities (Foss and Klein 2012, 39).

They focus on the conceptualization of entrepreneurship as judgment because they see it as a natural complement to the theory of the firm. To capture this relationship, they turn to Knight who linked profit and the firm to the existence of uncertainty (Knight [1921] 2012). Entrepreneurship requires making a judgment about the most uncertain events, and due to this utmost uncertainty, entrepreneurial judgment cannot be sold on the market, but rather becoming an entrepreneur requires starting a firm. Judgment, as considered by Foss and Klein, pertains employment of resources; therefore, entrepreneurial decision-making is ultimately a decision-making about the employment of resources. Interestingly, and to further separate themselves from Kirzner, the authors argue that opportunities for profit do not exist independently, they are not out there, waiting to be discovered. Rather, they need to be created and manifested in action.

Entrepreneurship in Action

Numerous authors have expanded on Knight, Kirzner, and Schumpeter in a manner that allows us to better understand the mechanism of how entrepreneurs actually deal with uncertainty. One of these authors is David Harper, who adopted the Popperian growth of knowledge approach to illuminate entrepreneurial learning (Harper 1996). Harper postulates that market entrepreneurship is akin to scientific hypothesis testing. Similarly to

Foss and Klein, Harper is not persuaded by the notion of passive alertness and believes entrepreneurial opportunities are not discovered but must be created. Due to uncertainty and human fallibility, there is no way of knowing whether the new opportunity is going to be a success or a failure. Therefore, each time entrepreneurs launch a new product or venture, they are testing a hypothesis. If businesses succeed, the hypothesis is unfalsified, and it remains unfalsified as long as there is sufficient interest to sustain the venture. As in science, rejected hypothesis are valuable: they contribute to the growth of knowledge.

Accordingly, Harper defines entrepreneurship “as profit-seeking activity aimed at identifying and solving ill-specified problems in structurally uncertain and complex situations. It involves the discovery and creation of new ends–means frameworks, rather than the allocation of given means in the pursuit of given ends” (Harper 1996, 3). By exploiting similarities between the market process and the process of scientific knowledge formation, Harper sheds light on how entrepreneurs acquire, use, and disseminate new, often tentative and conjectural, knowledge. Following Loasby, Harper notes that the relationship between market process and scientific progress is not a coincidental metaphor but both describe the same process: the growth of knowledge (Loasby 1989).

Harper argues that since scientists and entrepreneurs are problem-solvers, their activities must be described in terms of human ends and purposes. This systematic approach to uncertainty makes Harper’s entrepreneur significantly different from Schumpeter’s and Kirzner’s conceptualizations. While Schumpeter and Kirzner highlighted the nonrational and intuitive aspects of entrepreneurial behavior, Harper argues that rationality and critical methods of error elimination are crucial for acquiring new knowledge.

Saras Sarasvathy offers parallel insights into how entrepreneurs deal with uncertainty. Sarasvathy contributes to the entrepreneurship scholarship a notion of effectuation, which she defines as an inverse of causation. Causal thinking involves selecting means most suitable for the achievement of predetermined ends. Effectuation, in contrast, starts with given means and seeks to

create new ends. While economic agents are typically portrayed as employing only causal thinking, Sarasvathy’s entrepreneurs combine causal and effectual thinking. They act based on what they know about themselves and others, focus on the control of resources at hand, and on what kind of effects are within their reach. As they generate the alternatives, they simultaneously assess the qualities of different ends. To Sarasvathy, entrepreneurship and effectuation are crucial in opening economics to imagination and novelty, which in turn has significant analytical implications. For example, Sarasvathy argues that while economists have traditionally assumed existence of such artifacts as firms, organizations, or markets, the creation of these facts cannot be explained without effectuation (Sarasvathy 2001, 2008).

Entrepreneurship and Institutions

Given the interest in the relationship between entrepreneurship and economic growth (Aghion and Howitt 1990; Wennekers and Thurik 1999; Blanchflower 2000), impact of institutions on entrepreneurship begun to attract increasing attention. In his now classic paper, Baumol considers the impact of institutional environment on the relative payoffs to different forms of entrepreneurship (Baumol 1990). According to Baumol, not all forms of entrepreneurship are productive. For example, rent seeking, litigation, and warfare are entrepreneurial in nature but have a negative impact on economic growth. When institutional environments renders payoffs to unproductive or destructive activities higher than payoffs to productive entrepreneurial activities, entrepreneurship does not translate into economic growth. In other words, Baumol argues that changes in the extent of productive entrepreneurial spirits are not random but depend on the rules within which entrepreneurial activity takes place. In a similar spirit, Sautet notes that “what is generally missing in countries with lackluster economic performance is not entrepreneurship as such but the right institutional context for entrepreneurship to take place and to be socially beneficial” (Sautet 2005).

It is now well accepted that institutions determine whether entrepreneurship contributes to economic growth. However, as noted by Boettke and Coyne, to simply point out the set of institutions most conducive to productive entrepreneurship is not the same as actually implementing these institutions (Boettke and Coyne 2003). This opens the question of what drives institutional change. While various theories of institutional change can be found in the literature, Wagner places institutions along with market phenomena as an outcome of entrepreneurial action and suggests bidirectional, entangled relationship between entrepreneurship and institutions (Wagner 2009, 2010). In a similar vein, and building upon the examination of institutional entrepreneurship through the lenses of institutional theory and institutional economics (Pacheco et al. 2010), Kuchař uses the emergence of surrogate motherhood market to illustrate how entrepreneurs drive institutional change by challenging existing institutional legal ordering and common interpretations of social phenomena (Kuchař 2016).

Directions for Future Research

If Sarasvathy and Venkataraman are correct in suggesting that we might have miscategorized entrepreneurship as a subfield of management and economics when it is a social force akin to democracy or scientific method (Sarasvathy and Venkataraman 2011), then scholars interested in pursuing entrepreneurship research need not worry about a shortage of research streams. Sarasvathy and Venkataraman go as far as to argue that entrepreneurship is not a subject of social inquiry but a method of human action. There are reasons to believe that they might be correct in anticipating an entrepreneurship-centered revolution, as evidenced by the growing interest in entrepreneurship and the emerging literature on entrepreneurship in unconventional institutional setting, i.e., social, political, institutional, and cultural entrepreneurship (Dacin et al. 2010). Another approach to considering entrepreneurship's role beyond the market setting is that of Podemska-Mikluch and Wagner: entrepreneurial

coordination across divergent institutional frameworks, not only within them (Podemska-Mikluch and Wagner 2017). If this broader understanding of entrepreneurship were to take off, one area of economic theory prone to rethinking would be market failure literature because its conclusion relies on the absence of entrepreneurship in any form, be it conventional, institutional, political, or cross-institutional.

References

- Aghion P, Howitt P (1990). A model of growth through creative destruction. National Bureau of Economic Research. <http://www.nber.org/papers/w3223>
- Baumol WJ (1990) Entrepreneurship: productive, unproductive, and destructive. *J Polit Econ* 98(5):893–921
- Blanchflower DG (2000) Self-employment in OECD countries. *Labour Econ* 7(5):471–505. [https://doi.org/10.1016/S0927-5371\(00\)00011-7](https://doi.org/10.1016/S0927-5371(00)00011-7)
- Blaug M (1998) Entrepreneurship in the history of economic thought. In: Boettke PJ, Ikeda S (eds) *Advances in Austrian economics*. JAI Press, Stamford, pp 217–239. <https://ideas.repec.org/p/exe/wpaper/9515.html>
- Boettke PJ, Coyne CJ (2003) Entrepreneurship and development: cause or consequence? *Adv Austrian Econ* 6:67–87
- Boudreaux D (1994) Schumpeter and Kirzner on competition and equilibrium. In: Peter J. Boettke (ed) *The market process: essays in contemporary Austrian economics*. Elgar, Aldershot
- Brouwer MT (2002) Weber, Schumpeter and Knight on entrepreneurship and economic development. *J Evol Econ* 12(1–2):83–105
- Cantillon R ((1755) 2015) Essay on the nature of trade in general. Liberty Fund. <https://muse.jhu.edu/book/41671>
- Casson M (1982) The entrepreneur: an economic theory. Rowman & Littlefield. <https://books.google.com/books?hl=en&lr=&id=vo-0aXjiLcoC&oi=fnd&pg=PR10&ots=3-PE11FJXM&sig=G2-othMw9CbJWjZe9zAWx9jNU1g>
- Choi YB (1995) The entrepreneur: Schumpeter vs. Kirzner. *Adv Austrian Econ* 2(Part A):55–78
- Cowen T (2003) Entrepreneurship, Austrian economics, and the quarrel between philosophy and poetry. *Rev Aust Econ* 16(1):5–23
- Dacin PA, Tina Dacin M, Matear M (2010) Social entrepreneurship: why we don't need a new theory and how we move forward from here. *Acad Manag Perspect* 24(3):37–57
- Demsetz H (1970) The private production of public goods. *J Law Econ* 13(2):293–306
- Elkjaer JR (1991) The entrepreneur in economic theory. An example of the development and influence of a concept. *Hist Eur Ideas* 13(6):805–815

- Foss NJ, Klein PG (2012) *Organizing entrepreneurial judgment: a new approach to the firm*. Cambridge University Press, Cambridge
- Harper DA (1996) *Entrepreneurship and the market process: an enquiry into the growth of knowledge*. Routledge, London
- Hayek FA (1937) Economics and knowledge. *Economica* 4(13):33–54
- Hayek FA (1945) The use of knowledge in society. *Am Econ Rev* 35(4):519–530
- Hébert RF, Link AN (1988) *The entrepreneur: mainstream views & radical critiques*. Praeger Publishers, New York
- Kirzner IM (1973) *Competition and entrepreneurship*. The University of Chicago Press, Chicago
- Kirzner IM (1999) Creativity and/or alertness: a reconsideration of the Schumpeterian entrepreneur. *Rev Aust Econ* 11(1–2):5–17
- Knight FH ((1921) 2012) *Risk, uncertainty and profit*. Dover Publications, Mineola
- Kuchař P (2016) Entrepreneurship and institutional change. *J Evol Econ* 26(2):349–379. <https://doi.org/10.1007/s00191-015-0433-5>
- Loasby B (1989) *The mind and method of the economist*. Books. Edward Elgar Publishing. <http://econpapers.repec.org/bookchap/elgeebook/288.htm>
- Manne HG (2014) Resurrecting the ghostly entrepreneur. *Rev Aust Econ* 27(3):249–258
- Pacheco DF, York JG, Dean TJ, Sarasvathy SD (2010) The coevolution of institutional entrepreneurship: a tale of two theories. *J Manag* 36(4):974–1010
- Podemska-Mikluch M, Wagner RE (2017) Economic coordination across divergent institutional frameworks: dissolving a theoretical antinomy. *Rev Political Econ*. January 29(2):249–266. <https://doi.org/10.1080/09538259.2016.1269425>
- Sarasvathy SD (2001) Causation and effectuation: toward a theoretical shift from economic inevitability to entrepreneurial contingency. *Acad Manag Rev* 26(2):243–263. <https://doi.org/10.5465/AMR.2001.4378020>
- Sarasvathy SD (2008) *Effectuation: elements of entrepreneurial expertise*. Edward Elgar Publishing, Cheltenham
- Sarasvathy SD, Venkataraman S (2011) Entrepreneurship as method: open questions for an entrepreneurial future. *Entrep Theory Pract* 35(1):113–135. <https://doi.org/10.1111/j.1540-6520.2010.00425.x>
- Sautet F (2005) The role of institutions in entrepreneurship: implications for development policy. *Policy Primer No.1, Mercatus Policy Series*
- Schumpeter JA ((1911) 2008) *The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle*. Transaction Publishers, New Brunswick
- Schumpeter JA ((1943) 2013) *Capitalism, socialism, and democracy*. Routledge, London.
- Shackle GLS (1972) *Epistemics and economics: a critique of economic doctrines*. Cambridge University Press, Cambridge, UK
- van Praag CM (1999) Some classic views on entrepreneurship. *De Economist* 147(3):311–335. <https://doi.org/10.1023/A:1003749128457>
- Venkataraman S (1997) The distinctive domain of entrepreneurship research. *Adv Entrep Firm Emerg Growth* 3(1):119–138
- Wagner RE (2009) Property, state, and entangled political economy. In: Schafer W, Schneider A, Thomas T (eds) *Markets and politics: insights from a political economy perspective*. Metropolis, Marburg, pp 37–49
- Wagner RE (2010) *Mind, society, and human action: time and knowledge in a theory of social economy*. Routledge, London
- Wennekers S, Thurik R (1999) Linking entrepreneurship and economic growth. *Small Bus Econ* 13(1):27–56

Environmental Crime

Giuseppe Di Vita

Department of Economics and Business,
University of Catania, Catania, Italy

Abstract

To study environmental crime in a perspective of law and economics, it is necessary to identify the protected species from an economic point of view and at the same time to give a legal definition of this kind of criminal behavior. The list of sanctions and their effective deterrence effects in cases of environmental crime are addressed in the final part of this entry.

Definition

Environmental crime is a behavior harmful for the natural environment and its population that is punished with criminal sanctions, according to the nature of the protected species and the type and magnitude of present and future damage.

Environmental Crime

Interdisciplinary issues are hard to approach because they involve different kinds of knowledge with the use of specific terminology; this is particularly true in the case of environmental crime. The analysis moves from the concept of environment as the object of legislative protection and then follows with the definition of

environmental crime. This is not an easy task because even within the legal doctrine there is no unanimous view of the concept of environmental crime, and technical language is not the same for lawyers and economists. The analysis follows with a list of the criminal sanctions that may be applied as a consequence of public enforcement of criminal environmental law against the violators of such rules. In conclusion, the deterrence effects of criminal sanctions are studied in practice, considering the results of empirical studies performed. (See Gray and Shimshack (2011) for a recent survey of the literature on empirical analysis on enforcement of criminal sanctions.)

Environment

From the economic point of view, we should define the interest protected by the law and the reasons why the policy maker (or legislator) chooses to punish illegal behavior against the environment by applying criminal sanctions.

The problem of an optimal use of the environment and its protection originates from its nature of public good (Siebert 1992) and its characteristics of non-excludability and non-rivalry. The inability of the market to determine an equilibrium price may be due to the social costs associated with the private use of the good (as in the case of pollution emissions), or the collective nature of the benefits, that make the excludability cost extremely high and the price market equal to zero (e.g., the case of biodiversity preservation). In the first case, the self-interest of individuals determines an exploitation of the environment, while in the second hypothesis, the indivisibility of the benefits makes the individuals unwilling to pay for this kind of good. Whenever the existence of a source of market failure is present, the market produces an inefficient allocation of resources (Bator 1958), thus rendering necessary the adoption of some economic and legal policy, to bring the system as close as possible to a situation of Pareto-optimal market allocation. A wide spectrum exists of possible measures that it is possible to implement in order to correct the inefficiencies of the market, such as Pigouvian taxes, standards,

tradable permits, Coasian solution, and creation of an institution (Faure 2012; Tietenberg and Lewis 2012). This study considers only the cases where the inefficient allocation of environmental assets is punished by means of criminal sanctions.

The main reason for committing a crime against the environment is the saving gained by avoiding a costly compliance to the regulation implemented, adopted with the aim of preserving the environment. The individuals who follow their self-interest, thereby breaking the environmental law, worsen social welfare, overexploiting the environment with irreversible harm for future generations. The economic benefits for violators increase according to the degree of stringency of the legislation (Almer and Goeschl 2015).

From the legislative and court point of view, the kind of remedy is commensurate with the gravity of the behavior to the environment (Rousseau 2009). So the private remedies of repayment for damage or injunction are confined to the less important cases where the consequences of illegal behavior are limited to the parties involved (e.g., the noise in a private apartment). The sources of market failures render the use of standard civil remedies difficult (like using the strict liability regimes, see Sigman 2010) and the administrative fines inadequate.

In the case of private goods, it is sufficient to design the property rights properly in order to bring the system to the optimal path, allowing the victim of damage to file the case in the court and to be reimbursed for the damage, thus indirectly promoting the private enforcement of environmental law. This efficient mechanism cannot be used in the case of public goods because of the incomplete appropriability of the benefits and the multiple offensive environmental damages that go beyond the individual dimension and involve collective values and the welfare of future generations.

To protect environmental goods like air, soil, and water, it is necessary to adopt policies promoting the use of devices to abate the level of pollution and/or to implement more environmentally friendly technologies together with other measures to preserve biodiversity and to avoid climate changes, domestic and trans-boundary pollution, illegal waste disposal in landfill,

pollution emissions, and so on. Often, such kinds of good cannot be efficiently protected by using the standard private law remedies (like injunction to avoid future damages and the condemnation of the author of the offense to pay past damages), in consideration of their characteristic as public good, requiring stronger measures from the policy maker, like a public enforcement of the law, that may only indirectly protect the private rights involved. (About the “civil enforcement tools” of environmental law, see Glicksman and Earnhart 2007.)

Environmental Crime

The “enforcement pyramid” (Ayres and Braithwaite 1995), i.e., the graduation of penalties for illegal behavior that could be applied to protect the environment, is made up of civil sanctions, administrative fines, and criminal sanctions. Although all the kinds of sanctions have as common denominator the protection of the environment by means of punishment of the violator and reimbursement of the damage, the first two kinds of enforcement of environmental law have as their principal aim to deter the violator and compensate the offended individuals, in contrast with the criminal sanctions that have as their main goal to punish the violator.

Among the civil sanctions to implement environmental law, the “citizen’s suit” deserves a special mention; it may be defined thus “... Citizen suits are court proceedings brought by citizens who seek to enforce public rights. In the environmental arena they are cases brought to enforce the rights or obligations created by environmental laws. They are civil as opposed to criminal cases. While citizens’ suits are often brought against government authorities this is not a definitive feature ...” (Mossop 1993). This measure to implement environmental law has been extensively used in wealthy (Naysnerski and Tietenberg 1992) and less developed countries, precisely since 1958 in the Czech Republic (Earnhart 2000). This kind of law enforcement is based on the citizens’ cooperation to protect the environment and may be useful in minor cases of

law infringement. (Regarding the effects of alternative policies increasing reliance of self-reporting by polluters, see Farmer 2007.)

The step immediately next to the first in the enforcement pyramid consists of the administrative fines that may be applied without any judge’s intervention; their implementation is devoted to public bodies of the state. The problem is that the dissuasive capability of administrative sanctions could be inefficient to punish and deter the more dangerous behavior to the environment. (See on the optimal level of environmental fines the interesting paper by Rousseau and Kjetil 2010.)

In general, it is possible to note that criminal proceedings are, in contrast to civil proceedings, initiated by the government or by public authorities rather than offended individuals. This kind of public enforcement of environmental law requires a high standard of proof, involving fundamental individual rights constitutionally protected (like individual freedom) and moral-social stigma consequent to an application of a criminal sanction.

The environmental crime may be defined from the legal point of view as a crime against the environment or the violation of an environmental law (Clifford and Edwards 1998). (Clifford and Edwards (1998) moreover propose a broad philosophical definition of an “environmental crime” as an act committed with the intent to harm or with a potential to cause harm to ecological and/or biological system and for the purpose of securing business or personal advantage. See Eman et al. (2013) for an updated survey of the definition of “environmental crimes” in general and within the European Union. For a criminological analysis of environmental crime, see Huisman and Van Erp 2013 and Lynch and Stretesky 2014).

Although from the legislative point of view it is simple to say what is an environmental crime, among the international community of law scholars, there is no unanimous definition of this concept; nevertheless, the phrase “environmental crime” is used in several international agreements and national legislation.

Here, we may underline the increasing relevance of the trans-boundary dimension of environmental protection (e.g., air, waste, water, global warming; green crimes include air

pollution, water pollution, deforestation, species decline, the dumping of hazardous waste, etc.). This indicates the need for the international community to define this concept univocally. The main obstacle to this aim is that the identification of environmental goods protected by criminal sanctions depends on the socioeconomic condition of each single country. So you may expect countries at the same stage of development to be more prone to adopt similar or identical legislation and enforcement systems, because their relationship between per capita income and pollution is the same. (About the relationship between per capita income and polluting emissions and the relevance of legal family in explaining the environmental quality levels at different stages of growth, see Di Vita (2008).) Developing countries are not able to use their scarce resources to protect the environment fully. These considerations could be of help in explaining why the delimitation of environmental crime differs among countries.

Criminal sanctions for violation of environmental law constitute a novelty introduced in the 1980s of the last century, with an increasing incidence in recent years (Billiet and Rousseau 2014; Goeschl and Jürgens 2014). Since the 1970s criminal sanctions to protect the environment have become a measure extensively used in the United States and later on in the European Community (Almer and Goeschl 2010). (For an updated survey of “environmental crime” within the European Union, see Öberg (2013).) In the European Union, the criminalization of environmental law enforcement has been strengthened by the European Directive 2008/99/EC (Billiet and Rousseau 2014). More recently, since the mid-1990s, some European countries have introduced administrative sanctions as a complement of criminal fines (Faure and Svatikova 2012; Faure 2012). Administrative measures have the advantage of being more flexible than criminal proceedings so that the enforcement of environmental law may be pursued at lower cost. Even in the former Eastern European countries, the environmental laws are beginning to be protected by the introduction of criminal sanctions (Eman et al. 2013).

Australia is a country where the protection of the environment by means of criminal sanctions

has a long tradition and since 1979 has played a leading role in criminal justice processing of environmental offenses through the New South Wales Land and Environment, being probably the first nation to introduce environmental courts (Walters and Westerhuis 2013). Among the so-called BRICS countries, Russia seems to have experienced a reduction of crimes against the environment (according to domestic statistics), although some people suggest that together with a decrease in ecological crimes and offenses, there is some increase of latency within the field of ecological crime. Recent research suggests that one of the major profiles of environmental crime counteraction could be the nongovernmental ecological control system development adopted in Russia that is devolved to both municipal and social ecological control. This raises some doubt about the degree of enforcement necessary to achieve the optimal level of environmental protection (Anisimov et al. 2013).

Even China, one of the most dynamic economies among the developing countries, has recently levied criminal sanctions to punish violators and deter future infringements of environmental law (Zong and Liao 2012). In the African continent, protection of the environment from criminals is not so widespread for the economic considerations discussed previously. Nevertheless, this may represent a missed opportunity to preserve one of the largest green lungs of the world and an inexhaustible source of biodiversity.

The key point of view to consider the importance of environmental crime is the dynamic effects of this kind of sanction in bringing the economic systems toward a more sustainable path to preserve the environment and render the stock of environment useful for future generations.

Criminal Sanctions To Punish and Prevent Behavior Harmful to the Environment

Based on the previous considerations, it is clear that the state has to choose whether to punish some illegal behavior with a criminal sanction in

consideration of the interest protected and the offensive magnitude of present and future damages. In the case of civil remedies to enforce the environmental law, the costs burden the citizens, with limited expenses for the public sector of the economy (state administration and its ramifications). The criminal sanctions are applied because it is assumed that they raise the level of environmental protection, with high costs to the public sector of economy in terms of investigation and imposition of penalties. The enforcement costs are relevant under binding budget constraint.

Shavell (1985, 1987) has identified five conditions that should be fulfilled to justify the application of nonmonetary sanctions for efficient criminal deterrence. Such conditions are (i) the probability of insolvency of the violator, (ii) the possibility that the criminal is able to avoid paying the monetary sanction, (iii) the illicit earnings for the private agent in infringing the environmental criminal law, (iv) the potential future damage to the environment, and (v) the magnitude of the damage. His conclusion is that the criminal sanctions are worth applying only when these five conditions are strong enough to outweigh the social costs to deter, punish, and hold in jail the authors of environmental crimes (Billiet and Rousseau 2014).

In a general theoretical framework, Emons (2003, 2007) found evidence that criminal sanctions should be applied in the case of repeat offenders, punishing the less harmful violation of the law with monetary penalties; escalating penalties are used not to make the criminal career less attractive but to make being honest more attractive. Despite these theoretical findings, the empirical question is whether the use of escalating sanctions based on offense history is valid when the benefit of crime and the probability of apprehension are high.

The recent guidelines of US sentences suggest inflicting a short period of imprisonment when the environmental crime leads to severe damage (e.g., an illegal dump of waste, river pollution, etc.). In accordance with previous theoretical findings, O'Hear (2004) reports that in Australia, the prison sentences for environmental crime

have been limited to the most serious offenses to the environment.

The sanction for a criminal defendant varies according to the crime and includes such measures as the sanction of death, not applicable in cases of environmental law infringement. Incarceration is a period of time that the violator spends in jail, as a consequence of a criminal sentence. Probation is the period during which a person, "the probationer," is subject to critical examination and evaluation. It is a trial period that must be completed before a person receives greater benefits or freedom. (The probationer criminal sanction could be considered quite similar to the suspended sentences that have a lower immediate effect but tend to deter future criminal behavior (Billiet and Rousseau 2014).) Community service is the period of time that the convicted spends doing civil services instead of jail or at the end of the period of imprisonment. Finally, monetary fines are monetary sanctions applied by the courts as a conclusion of a criminal proceeding (for a more complete explanation of the characteristics of criminal sanctions, see Öberg 2013).

Deterrence Effects of Criminal Sanctions

In the introduction we assume that the criminal sanctions are applied in cases of extremely offensive behavior toward the environment. The underground message is that criminal sanctions possess the greatest deterrence capacity to dissuade further environmental damage. The axiom that criminalization raises the level of environmental quality has been questioned several times (Stigler 1970; Polinsky 1980; Andreoni 1991; Heyes 1996), in consideration of greater costs for the use of criminal sanctions and the negative substitute effects on civil and administrative fines.

Recently, Goeschl and Jürgens (2014), moving from the initiatives of the European Commission to prevent environmental crimes, have shown that an increase in the criminalization level does not necessarily lead to higher environmental quality. The greater the area of criminal responsibility for environmental law violation, the lower will be the enforcement by using alternative measures.

Based on these considerations, the debate is shifted into the arena of applied studies.

Cherry's (2001) empirical analysis found strong empirical evidence that financial penalties have a deterrence effect stronger than prison sentences and that monetary sanctions imply a lower social cost to punish those responsible for crimes and to prevent future damages. The limitation of Cherry's study is that crimes against the environment are not considered but illegal behavior in general.

Faure and Heine (2005), in their research on the enforcement of environmental crimes in Europe, find that monetary sanctions are more widespread than criminal sanctions in the Old Continent. Almer and Goeschl (2010) found a positive relationship between environmental crime prosecution and the deterrence effect. This result is reinforced by the outcomes of another empirical study conducted, using microeconomic data, in Germany regarding in particular waste crimes, where persecutors faced inelasticity of 4% of resources devoted to environmental offices in response to a unitary increase in environmental crimes (Almer and Goeschl 2011).

Faure and Svatikova (2012) performed an empirical analysis on the effects of criminal sanction deterrence of illegal behavior against the environment when the penalty is used together with the administrative process to enforce environmental law. Their analysis is based on the theoretical assumption that to enforce environmental law, having a budget constraint, it is more efficient to use administrative fines, instead of criminal sanctions, because they are more flexible and cheaper than criminal proceedings.

Their evidence, based on a dataset available for four Western Europe nations (the Flemish Region, Germany, the Netherlands, and the United Kingdom), shows that administrative fines are a cheaper complement of criminal sanctions.

Billiet and Rousseau (2014) suggest that the relationship between monetary and nonmonetary sanctions is in complementary and not alternative terms. (For an empirical analysis performed on microdata regarding penalties applied by courts in Europe, see Billiet et al. 2014.) In

their empirical research, they found that the fines are usually applied together with other criminal sanctions, thus confirming the complementary nature of the different kinds of criminal sanctions.

Billiet and Rousseau (2014) investigated how prison sentences for environmental crime effects are used in fact and their deterrence effects to violators in Flanders, using statistics of seven courts of first instance and the Court of Appeal of Ghent, regarding the complete case law from 2003 to 2007. This empirical analysis confirms that in the EU, countries like Belgium, in 87.49% of criminal proceedings, only a fine sanction was applied to individual violators. These scholars in their interesting study found that the kind of environmental problem does not matter (Billiet and Rousseau 2014) and that prison sentences are not always executed. (Rousseau (2009) performed an interesting empirical analysis of the court sentences in proceedings regarding environmental crime, finding that judges protect the private assets and calculate the sanctions in consideration of the magnitude of the harm).

Almer and Goeschl (2015) performed an empirical analysis on a special kind of environmental crime regarding waste, considering a panel dataset of 44 counties from the German state of Baden-Württemberg between 1995 and 2005. Their outcomes support the theoretical assumption that criminal sanctions have a greater deterrence effect, although the magnitude of this effect depends on the intensity of enforcement. Moreover, the economic conditions and policy measures adopted at jurisdiction level play a significant role to explain why similar law infringements seem to be evaluated and punished in different ways.

The deterrence effect of criminal sanctions to punish the violators of environmental law seems to be a promising field of future research to draw guidelines for further legislative measures.

Cross-References

- [Commons, Anticommons, and Semicommons](#)
- [Externalities](#)

References

- Almer C, Goeschl T (2010) Environmental crime and punishment: empirical evidence from the German penal code. *Land Econ* 86(4):707–726
- Almer C, Goeschl T (2011) The political economy of the environmental criminal justice system: a production function approach. *Public Choice* 148:611–630
- Almer T, Goeschl T (2015) The Sopranos redux: the empirical economics of waste crime. *Reg Stud* (forthcoming) 49(11):1908–1921
- Andreoni J (1991) Reasonable doubt and the optimal magnitude of fines: should the penalty fit the crime? *RAND J Econ* 22(3):385–395
- Anisimov AP, Aleksseva AP, Melihov AI (2013) Current issues of ecological crime counteraction in Russia. *Criminol J Baikal Natl Univ Econ Law* 3:80–89
- Ayres I, Braithwaite J (1995) *Responsive regulation: transcending the deregulation debate*. Oxford University Press, Oxford
- Bator FM (1958) The anatomy of market failure. *Q J Econ* 72(3):351–379
- Billiet CM, Rousseau S (2014) How real is the threat of imprisonment for environmental crime? *Eur J Law Econ* 37(2):183–198
- Billiet CM, Blondiau T, Rousseau S (2014) Punishing environmental crimes: an empirical study from lower courts to the court of appeal. *Regul Gov* 8(4):472–496
- Cherry TD (2001) Financial penalties as an alternative criminal sanction: evidence from panel data. *Atl Econ J* 29(4):450–458
- Clifford M, Edwards TD (1998) Defining ‘environmental crime’. In: Clifford M (ed) *Environmental crime: enforcement, policy, and social responsibility*. Aspen Publishers, Gaithersburg, pp 121–145
- Di Vita G (2008) Differences in pollution levels among civil law countries: a possible interpretation. *Ener Pol* 36(10):3774–3786
- Earnhart D (2000) Environmental “citizen suits” in the Czech Republic. *Eur J Law Econ* 10(1):43–68
- Eman K, Meško G, Dobovšek B, Sotlar A (2013) Environmental crime and green criminology in South Eastern Europe – practice and research. *Crime Law Soc Change* 59:341–358
- Emons W (2003) A note on the optimal punishment for repeat offenders. *Int Rev Law Econ* 23:253–259
- Emons W (2007) Escalating penalties for repeat offenders. *Int Rev Law Econ* 27:170–178
- Farmer A (2007) *Handbook of environmental protection and enforcement – principles and practice*. Earthscan, London/Sterling City
- Faure MG (2012) Effectiveness of environmental law. What does the evidence tell us? *William Mary Environ Law Policy Rev* 36(2):293–338
- Faure M, Heine G (2005) *Criminal enforcement of environmental law in the European Union*. Kluwer Law International, London
- Faure MG, Svatikova K (2012) Criminal or administrative law to protect the environment? Evidence from Western Europe. *J Environ Law* 24(2):253–286
- Glicksman R, Earnhart D (2007) Depiction of the regulator-regulated entity relationship in the chemical industry: deterrence-based vs. cooperative enforcement. *William Mary Environ Law Policy Rev* 31(3):603–660
- Goeschl T, Jürgens O (2014) Criminalizing environmental offences: when the prosecutor’s helping hand hurts. *Eur J Law Econ* 37(2):199–219
- Gray WB, Shimshack JP (2011) The effectiveness of environmental monitoring and enforcement: a review of the empirical evidence. *Rev Environ Econ Policy* 5(1):3–24
- Heyes AG (1996) Cutting environmental penalties to protect the environment. *J Public Econ* 60(2):251–265
- Huisman W, Van Erp J (2013) Opportunities for environmental crime: a test of situational crime prevention theory. *Br J Criminol* 53(6):1178–1200
- Lynch MJ, Stretesky PB (2014) *Exploring green criminology: toward a green criminological revolution*. Ashgate, Farnham
- Mossop D (1993) Citizen suits – tools for improving compliance with environmental law. In: Cunningham N, Norberry J, McKillop S (eds) *Environmental crime: proceedings of a conference held 1–3 September 1993*. Australian Institute of Criminology, Canberra
- Naysnerski W, Tietenberg T (1992) Private enforcement of federal environmental law. *Land Econ* 68(1):24–48
- O’Hear MM (2004) Sentencing the green-collar offender: punishment, culpability, and environmental crime. *J Crim Law Criminol* 95(1):133–276
- Öberg J (2013) The definition of criminal sanctions in the EU. *Eur Crim Law Rev* 3:273
- Polinsky AM (1980) Private versus public enforcement of fines. *J Leg Stud* 9(1):105–127
- Rousseau S (2009) Empirical analysis of sanctions for environmental offenses. *Int Rev Environ Resour Econ* 3(3):161–194
- Rousseau S, Kjetil T (2010) On the existence of optimal fine for environmental fines. *Int Rev Law Econ* 30(4):329–337
- Shavell S (1985) Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Rev* 85:1232–1262
- Shavell S (1987) The optimal use of nonmonetary sanctions as a deterrent. *Am Econ Rev* 77(4):584–592
- Siebert H (1992) Environmental quality as a public good. In: *Economics of the environment: theory and policy*. Springer, Berlin/Heidelberg
- Sigman HA (2010) Environmental liability and redevelopment of old industrial land. *J Law Econ* 53:289–306
- Stigler GJ (1970) The optimum enforcement of laws. *J Polit Econ* 78(3):526–536
- Tietenberg T, Lewis L (2012) *Environmental and natural resource economics*, 9th edn. Pearson Prentice Hall, Upper Saddle River
- Walters R, Westerhuis DS (2013) Green crime and the role of environmental courts. *Crime Law Soc Change* 59(3):279–290
- Zong X, Liao G (2012) Sustainable development and the investigation of criminal liability for Chinese environmental pollution. In: *World Automation Congress (WAC) proceedings 2012*, Puerto Vallarta

Environmental Policy: Choice

Donatella Porrini

Department of Management, Economics,
Mathematics and Statistics, University of Salento,
Lecce, Italy

Abstract

The choice between environmental policies is traditionally considered in terms of *ex ante* versus *ex post* interventions on the behavior of (potential) injurers that (can) cause an environmental accident with a consequent environmental damage. Moreover, market-based policies can be implemented as an indirect form of incentive for correct behavior. The two market-based policies, taxes and tradable permits system, are compared on the basis of the difference between price and quantity instruments. And finally, the real situation of the presence of an environmental policy mix is considered as a research challenging topic.

Definition

An environmental policy, on an economic analysis of law point of view, is an instrument to correct malfunctions and subsequent inefficiencies that originate from the presence of market failures: the environment appears as a “public good” that may not be appropriated by anyone and has no market price; the damage to the environment is a case of “externality,” in that it is, fully or partly, a social cost that is not internalized into the accounts of the parties causing it (Cropper and Oates 1992).

Introduction

Different policies can be implemented to reach a given set of environmental protection objectives. But nearly all environmental policies consist of two components: the identification of an overall goal (such as a certain level of air quality or an

upper limit on emission rates) and the choice of instruments to achieve that goal. In practice, these two components are often linked within the political process, because both the selection of a goal and the mechanism for achieving that goal have important political implications.

“But looking at this problem from a law and economics perspective, we can move from the theoretical definition of the efficiency of different instruments to their practical, and so direct, potential to achieve concrete objectives. In particular, three objectives emerge as relevant in judging the practical efficiency of environmental policies: the first is paying accident compensation to the victims; the second is prevention, in the sense of providing incentives for firms to improve safety standards; and the third is connected with technological change in the sense of encouraging firms to adopt lower-risk technologies” (Porrini 2005, pp. 350, 351).

Hereafter, the choice between environmental policies is considered in terms of *ex ante* versus *ex post* interventions on the behavior of (potential) injurers that (can) cause an environmental accident with a consequent environmental damage. Moreover, market-based policies can be implemented as an indirect form of incentive for correct behavior. The two market-based policies, taxes and tradable permits system, are compared on the basis of the difference between price and quantity instruments. And finally, the real situation of the presence of an environmental policy mix is considered as a research challenging topic.

Ex Ante Command-and-Control Policy

Generally, an *ex ante* policy focuses on the potential injurer’s activity from which the damage originates, with effects that precede the occurrence of the same. In the environmental sector, in particular, the standard-setting instrument is most common and consists in the enforcement, by an agency, of a given prevention level in any way that may be quantitatively defined. Practically, this *ex ante* policy is founded on a centralized structure in charge of setting standards and then ensuring their compliance, according to the

so-called classical “command-and-control” process.

As to the US experience with *ex ante* policy, the activity of the EPA (Environmental Protection Agency) provides a clear example of policy of this kind implemented by an independent environmental authority. This agency performs its tasks through the enforcement of polluting emission thresholds and the performance of inspections and, possibly, of actions brought to the federal courts.

The choice to develop this *ex ante* policy provides the advantage of centralized agencies to assure a cost-effective calculation on the basis of the expected damage and of the marginal cost of the different technical preventive instruments. So, following the traditional economic analysis of law approach, well-defined standards generate the correct incentive for the firm to act with caution and take the best production and prevention decisions (Calabresi 1970).

Moreover, in its application, the centralized search facilities, the continual oversight of problems, and a range of regulatory tools make this kind of policy capable of systematically assessing environmental risks and implementing a comprehensive set of instruments. But, on the other hand, as disadvantages, the agencies may not be very flexible in adapting to changing conditions, and a centralized command structure, relying on expert advice, may be subject to political pressure as well as to collusion and capture by the regulated firms.

Summarizing, the choice to implement an *ex ante* command-and-control kind of policy responds to the need of “uniformity versus flexibility.” “To be efficient, such regulations, require information on alternative techniques and a balancing between the profit to the industry and the environmental impacts of various production techniques and processes. Specific production standards may, therefore, rapidly become obsolete” (Faure and Skogh 2003, p. 198).

Ex Post Liability Policy

The most common *ex post* policy consists in a liability system that provides for a legal authority

to identify a party responsible for the damage caused by an environmental accident.

Again, the experience in USA can be considered as an example given that the problem of environmental liability in that country has emerged more than 30 years ago. In fact, in 1980 the Congress issued the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) with the main purpose to bring quick relief and remedy action after an accident, to cope with the “decontamination” of polluted sites, and to recover the cleanup and compensation costs from the liable parties.

The *ex post* policy is based on the system of liability assignment to the polluting company according to the so-called polluter pays principle. Such principle has been defined as “economic” in that it provides for the economic option of charging the cost of the environmental damage to a specific party that is liable for the accident, rather than generally to the society.

A system of liability satisfies the need to compensate affected parties for physical and economic damage originated from a real event loss and to stimulate preventive measures against future accident.

On an economic analysis of law point of view, in a world of perfect (complete) information, this *ex post* policy is an efficient method to solve the problem of internalizing the economic effects of environmental accident. In fact, the firms face the proper incentive to take the optimal level of precaution, and the individuals, harmed by pollution, receive the proper compensation, possibly through an insurance provider.

But in practice, the assignment of individual responsibility shows relevant problems that we can summarize in:

- (i) a specific polluter could in many cases be difficult to identify because some consequences (i.e., a disease or a reduction in health) may be attributed to a number of different factors besides the pollution and, even if a link between a pollutant and the consequences may be established, it could turn out to be difficult to determine the firm responsible for the damage;

- (ii) compulsory insurance contracts that the firms are induced or forced to buy could be incomplete or insufficient because of the difficulty to determine the probability of accident and the distribution of the loss caused by environmental damage, hence making the pricing of the contracts complex;
- (iii) the polluter can in some cases be insolvent and unable to pay for cleanup or compensation costs because of the (small) size of the firms operating in dangerous activities in comparison with the (high) costs and penalties of environmental damage.

A liability system can be applied using either a negligence or a strict liability regime. The law and economic literature compares the two regimes considering how they provide a potential polluter with incentives to take adequate preventive measures (Cooter and Ulen 1988). And, generally, strict liability is preferred in the presence of informational issues (Epstein 1973). This could be the case of environmental accidents, where it is particularly difficult to determine the standard to assign liability on the basis of negligence: in reality, pollution has many sources and (potentially) many victims, and it is a hard task to prescribe efficient pollution standards based on a calculus of the abatement cost and the external harm of every source of pollution.

Ex Ante Versus Ex Post Policies

The traditional economic analysis of law approach compares ex ante and ex post policies on the basis of their role in achieving the efficiency goal to stimulate the socially optimal level of preventive care and in so doing controlling the environmental risk.

The authorities responsible for meeting these objectives are the regulatory agencies, which fix standards and check their compliance, and the courts, which assign liability. "Statutory regulation, unlike tort law, uses agency officials to decide individual cases instead of judges and juries; resolves some generic issues in rulemakings not linked to individual cases, uses nonjudicialized procedures to evaluate technocratic information, affects behavior ex ante

without waiting for harm to occur, and minimizes the inconsistent and unequal coverage arising from individual adjudication. In short, the differences involve who decides, at what time, with what information, under what procedures, and with what scope" (Rose-Ackerman 1991).

A seminal contribution by Shavell (1984) presents four determinants on the basis of which comparing the two different policy systems.

The first determinant is the difference in information between private parties and the regulatory authority. It clearly could happen that the nature of the activities carried out by the firms is such that the private parties have better knowledge about the cost of preventive care and of the risks involved. In such a case, a liability system is more efficient because it makes the private parties the residual claimants of the control of risks. But it may also happen that the regulator has better knowledge because of the possibility of centralizing information and decisions, in particular when knowledge of risks requires special replicable and reusable expertise. In such a case, direct regulation is likely to be more efficient.

A second determinant is the limited capacity of private parties to pay the full costs of an accident, either because of limited liability or of insufficient assets; this is the case called "judgment proof" (Shavell 1986). An ex post liability policy allows to reach a social optimal solution, only if the private parties can finally cover the damage in the dimension decided by the court. In all the cases in which the damage is superior to the resources of the private parties, a liability system does not provide private parties with proper incentives for preventive care. So the greater the probability or the severity of an accident are and the smaller the assets of the firm are (relative to the potential damages), then the greater the efficiency of ex ante regulation.

The third determinant is the probability with which the responsible parties would face a legal suit for harm caused. This problem is particularly present in environmental risks: in many cases the victims are widely dispersed with none of them motivated to initiate a legal action, harm may appear only after a long delay, and specifically responsible polluters may be difficult to identify.

In the comparison with *ex ante* policy, *ex post* liability is more uncertain and so less efficient.

The fourth determinant is connected with the general level of administrative expenses incurred by the private parties and the public. The cost of a liability system includes the cost of legal procedure and the public spending for maintaining legal institutions. The cost of an *ex ante* policy includes the private costs of compliance and the public spending for maintaining the regulatory agencies. In this case the advantage of the liability system is that legal costs are sustained only if a suit occurs, and, if the system works well, stimulating the efficient level of prevention, the number of suits will decrease, and therefore the costs are becoming lower in time. On the other hand, under an *ex ante* policy, the administrative costs are sustained whether or not the accident occurs because the process of regulation is costly by itself, and the regulator needs in any case to collect continually information about the parties, their activities, and the risks.

Summarizing, *ex ante* policy is better when harm is large, is spread among many victims or, takes a long time to show up, when accidents are not very rare events, and when standards or requirements are easy to find and control.

In another contribution, Kolstad et al. (1990) support the hypothesis of the complementarity between *ex ante* and *ex post* policies because, even if the economic literature has mainly studied separately the two systems, characterizing each of them by different inefficiencies, the phenomenon of joint use of *ex ante* and *ex post* systems is widespread.

Among the determinants presented by Shavell (1984) and above analyzed on the basis of which comparing the two different regulatory systems, the authors concentrate on the fact that potential injurer is in many circumstances uncertain about whether a court will hold him liable in the event of accident and subsequent suit. But an *ex ante* regulatory system can correct this inefficiency, at least in part.

The development in the contributions within the economic analysis of law literature shows an increasing attention to the relationship between *ex ante* and *ex post* policies characterized by

imperfections in their implementation, as complements or substitutes, both in providing the incentives to the optimal level of preventive care. So the debate about policy choice mainly focuses on the achievement of a given target in terms of efficiency in a framework where imperfections are considered as reasons to prefer one policy to the other.

Market-Based Policies: Taxes and Tradable Permits

Within the category of market-based instruments, there are policies that encourage behavior through market signals rather than through explicit directives to firms. So, in practice, rather than imposing uniform emission standards, market-based instruments introduce a cost for the firm in the form of a tax or of the price for a permit, leaving then the firms to deal with the problem to control and limit the pollution level on the base of their marginal costs.

So stressing an efficiency kind of argument, in the case of the implementation of market-based instruments, each firm determines until which level it is more convenient to reduce pollution given the possibility to pay a tax or to buy a polluting permit.

The two most important features of market-based instruments with regard to traditional command-and-control approach are cost-effectiveness and dynamic incentives for technology innovation and diffusion. On one hand, command-and-control policies, to set standards, require that policy makers obtain detailed information about the compliance costs, each firm faces the problem that such information may be not available to government; by contrast, market-based instruments provide for a cost-effective allocation of the environment control burden without the need for this information.

As market-based policies, the most common are tradable permits system and environmental taxes.

A system of tradable permits is based on the allocation of a number of permits to the firms, each of them allowing the emission of a given amount of a pollutant; if the facility is able to

reduce its emissions (preferably through the use of less polluting technologies), it can sell its remaining emission permits to other firms that are unable to meet their quotas.

It is clear that the advantage of this policy is the possibility to fix the level of pollution control and in the case of technological change, without additional government intervention, to freeze this level.

On the other hand, a policy based on taxation attributes a price on polluting activities that will be incorporated by the firm in the price of the products. In this case the incentive for the adoption of abatement techniques relies on the market mechanism because if a firm does not apply the optimal techniques, it will pay more taxes and sell its products at a higher price than other firms with negative effects in terms of competition.

Market-based instruments, as regulatory devices that shape behavior through market signals rather than explicit instructions on pollution control levels or methods, are often described as “harnessing market forces” because they can encourage firms and individuals to undertake actions that serve both their own financial interest and public policy goals (Stavins 1998).

Environmental taxes and tradable permits system are both market-based policy instruments, but their implementation is different: taxes fix the marginal cost for carbon emissions and allow quantities emitted to adjust, whereas tradable permits fix the total amount of carbon emissions and allow price level to change according to market forces. Because of these differences, the former are defined as “price” instruments for the correlated effect to increase the price of certain goods and services, thereby decreasing the quantity demanded, while tradable permits are defined as “quantity” instruments for the feature to directly fix the quantity through the number of permits.

The literature on environmental policy choice describes alternative instrument taxes as price control instruments and tradable permits as quantity control ones, and many contributions compare their relative performance in terms of efficiency under uncertainty.

The starting seminal article of Weitzman (1974) analyzed the optimal instrument choice under a static partial equilibrium framework,

consisting of a reduced form specification of costs and benefits from abatement. In the setup, an agency issues either a single price order (fixed price) or a single quantity order (fixed quantity), and these fixed policies result in different expected social welfare outcomes under uncertainty. Specifically, Weitzman shows that, with imperfect information about the abatement costs, the relative slopes of the marginal benefit (damage) function and the marginal cost function determine whether one instrument is preferred to another. If the expected marginal benefit function from reducing emissions is flatter than the marginal cost of abatement, then a price control is preferred. If, however, the marginal benefit function is steeper, then a quantity control is preferred.

In the law and economic literature, Kaplow and Shavell (2002) deal with the standard context of a single firm producing externality; moreover, they consider the case of nonlinear corrective tax, and multiple firms jointly create an externality, demonstrating the superiority of taxes to permits.

Despite the results of the majority of contributions that a taxation system is preferable to tradable permits system in terms of economic efficiency, this policy obviously faces political opposition. On the supply side of the market, companies oppose taxes, as a cost that implies a revenue transfer to the government and also as a factor that can imply negative effect on competition in an international context; on the demand side, consumers are typically not happy and pay at the end a higher price on the products, and environmental groups oppose taxes because, unlike tradable permits system, these fail to guarantee a particular reduction in the emission level.

Conclusive Remarks About the Mix of Policy Instruments

We have analyzed the different policies as alternative instruments that can be implemented to reach given environmental objectives considering, on a law and economic perspective, their different degree of efficiency.

But environmental policy instruments usually operate as part of a “mix” of instruments, and in

practice several different policies are applied to address a given environmental problem as broadly as possible with the target to cover all sources of pollution in every relevant sector of the economy.

The efficiency of these mixes can be enhanced by adhering to many of the same principles that guide the use of individual instruments and by explicitly considering the way in which different instruments interact.

For example, one possible mix could be tradable permits together with tax system.

On an efficiency point of view, while a policy based on “quantity,” such as a tradable permits system, can provide a degree of certainty as to the environmental outcome, the compliance costs that will eventually be faced by polluters are likely to be quite uncertain under these systems. But this uncertainty can be reduced by introducing a tax system as a “safety valve” in the permit price. In effect, this allows polluters to emit whatever amount they like, in return for paying a fixed price, the “tax,” for any emissions for which they do not hold an allowance, should the permit price exceed a predefined level.

This mix between environmental policies presents some economic advantages that are key motivating forces to try to develop research on this topic.

First of all, policy mix allows for exchanges across different systems and thereby facilitates cost-effectiveness, that is, achievement of the lowest-cost emission reductions across the set of linked systems, minimizing the overall cost of meeting the collective cap.

Mixed systems may also provide regulatory stability as an advantage for affected firms, in the sense that it may be more difficult to introduce changes in an emission-reduction scheme when those changes require some sort of coordination with other policies (Johnstone 2003).

There are also administrative benefits from the mix that come from sharing knowledge about the design and operation of different policies to find the best practice, but also from the reduction of administrative costs through the sharing of such costs and the avoidance of duplicative services.

Despite the just mentioned economic advantages, we cannot find in the law and economic literature until now so many researches that

develop theoretical models based on the mixed use of different environmental policy instruments that are still considered mainly as alternative.

References

- Calabresi G (1970) The cost of accident. Yale University Press, New Haven
- Cooter RD, Ulen TS (1988) Law and economics. Collins, Harper
- Cropper ML, Oates WE (1992) Environmental economics: a survey. *J Econ Lit* 30:675–740
- Epstein RA (1973) A theory of strict liability. *J Leg Stud* 2:151–204
- Faure M, Skogh G (2003) The economic analysis of environmental policy and law – an introduction. Edward Elgar, Cheltenham
- Johnstone N (2003) The use of tradable permits in combination with other environmental policy instruments. Report ENV/EPOC/WPNEP(2002)28/Final, OECD, Paris
- Kaplow L, Shavell S (2002) On the superiority of corrective taxes to quantity. *Am Law Econ Rev* 4:1–17
- Kolstad CD, Ulen TS, Johnson GV (1990) Ex post liability for harm vs. ex ante safety regulation: substitutes or complements. *Am Econ Rev* 80:888–901
- Porrini D (2005) Environmental policies choice as an issue of informational efficiency. In: Backhaus JG (ed) The Elgar companion to law and economics, 2nd edn. Edward Elgar, Cheltenham, pp 350–363
- Rose-Ackerman S (1991) Regulation and the law of torts. *Am Econ RevPap Proc* 81: 54–58
- Shavell S (1984) Liability for harms versus regulation of safety. *J Leg Stud* 13:357–374
- Shavell S (1986) The judgement proof problem. *Int Rev Law Econ* 6(June):45–58
- Stavins R (1998) Market-based environmental policies. Resources for the Future, Discussion Paper 98–26
- Weitzman ML (1974) Prices vs. quantities. *Rev Econ Stud* 41(4):477–491

Equilibrium Theory

Pasquale L. Scandizzo

CEIS (Centre for International Economic Studies), University of Rome “Tor Vergata”, Rome, Italy

Abstract

Equilibrium is a key concept of modern science, from classical mechanics to biology, so

that its importance for economics should not be a surprise. More surprising is perhaps its central role in the tentative transition of economics from an essentially institutional set of disciplines to a unified branch of modern science, based on a rigorously axiomatic system. The main attempt for such a transition, which has as protagonists several Nobel prizes, was consumed in the 1950s, in an atmosphere first of elation and then of disillusion that appears very similar to that experienced 20 years earlier by mathematicians because of the failure of the Hilbert unification program. In spite of the somewhat spectacular nature of both successes and failures of its mathematical theory, however, general equilibrium has a central role in the history of modern economics that goes much beyond its formal treatment. This role is inextricably related to the notion of the market and is perhaps the primary constituents of the eclectic nature and vitality of contemporary economics.

General Equilibrium as a Property of the Market

Perhaps because of its increasing importance to the modern economy, the concept of market equilibrium has been evolving over the course of the story, so as to embrace different categories, in some cases more inclusive and in others more specific, which extend over the entire arc of economic phenomena. This extension of a concept originally defined in a very simple fashion has led to considerable confusion, especially since the theory and practice of modern economic systems are, in fact, completely dependent on market equilibrium as a category of the spirit, even more than as a theoretical model.

What then is a market equilibrium and in what sense can it be defined as a “general equilibrium”? The word market was born as the past participle of the Latin verb *mercari*, which means trading, and designates the beginning of its use of a physical place where the goods available for exchange were exposed and where, therefore, negotiations and exchanges were carried out. The designation

of the place, still widely used in everyday language, came to denote at the same time the space, the exchange, and the agents engaged in trade of specific goods. As a result, the place evoked by the term became increasingly virtual, the transition between the material and the virtual occurring through the discovery of the properties of the markets that extend well beyond their apparent spatial limits and culminate in the concept of market equilibrium. Equilibrium, on the other hand, was a concept that only gradually came to characterize the market, as its assertion went hand in hand with the increasing virtuality of the exchange and the agents involved in making it happen.

The process by which the market became dematerialized, in fact, can be seen as an increasing perception of its role as a mechanism to balance the exchange, by ensuring that agents meet and act appropriately to achieve a balance of actions and desires. For example, one of the precursors of economic liberalism Richard Cantillon (1755) offers a description of the market and its reasons to demonstrate how the activity originating from the establishment of a periodical fair creates broader consequences:

There are some villages where markets have been established by the interest of some proprietor or gentleman at court. These markets, held once or twice a week, encourage several little undertakers and merchants to set themselves up there. They buy in the market the products brought from the surrounding villages in order to carry them to the large towns for sale. In the large towns they exchange them for iron, salt, sugar and other merchandise which they sell on market days to the villagers. Many small artisans also, like locksmiths, cabinet makers and others, settle down for the service of the villagers who have none in their villages, and at length these villages become market towns. A market town being placed in the centre of the villages, and at length these villages become market towns. A market town being placed in the centre of the villages whose people come to market, it is more natural and easy that the villagers should bring their products thither for sale on market days and buy the articles they need, than that the merchants and factors should transport them to the villages in exchange for their products. (1) For the merchants to go round the villages would unnecessarily increase the cost of carriage. (2) The merchants would perhaps be obliged to go to several villages

before finding the quality and quantity of produce which they wished to buy. (3) The villagers would generally be in their fields when the merchants arrived and not knowing what produce these needed would have nothing prepared and fit for sale. (4) It would be almost impossible to fix the price of the produce and the merchandise in the villages, between the merchants and the villagers. In one village the merchant would refuse the price asked for produce, hoping to find it cheaper in another village, and the villager would refuse the price offered for his merchandise in the hope that another merchant would come along and take it on better terms. (Cantillon 1755)

For Adam Smith, one of the founders of the notion of equilibrium, the market was also originally identified in a physical space, such as a place dedicated to trade in the city or even the entire city (as a market town) with respect to the surrounding countryside. For him, however, the essential feature of the market was its ability to feed and be fed by the division of labor. In this sense, the market becomes a place dedicated to the determination of value by a threefold change: (a) from primitive societies to modern societies, (b) from the family to the firm, and (c) from the countryside to the cities. Smith's conception of the market is therefore first and foremost ideological, as pointed out by Joan Robinson (1962). It identifies the transition between two ideal states "the economy of barter" and "the market economy," as a form of substantial progress that is the basis of value creation and prosperity of modern societies. In this sense, Smith sees the market as an anthropological construct, which explains the redemption of man from the "rough state" where the assets and exchanges are both limited by the absence of the division of labor.

Ricardo, who starts from the paradigm of the Smithian division of labor to develop his fundamental theory of comparative advantage, seems to show a total nonchalance towards the concept of the market as compared to that of equilibrium. The market as such is evoked only from time to time, as a place dedicated to trade, in which the goods produced are carried. However, in the great debate on international trade, it remains a kind of constituency space, which distinguishes the "foreign" from the "home" market, both defined almost as a side effect of trade between countries. In fact,

both Smith and Ricardo did not seem to attach much importance to the market, in the sense that they take its presence for granted within a system that they characterize more than as a "market economy," as an economy of capital. Yet the fundamental problem, in this system, is not to explain the nature and behavior of economic entities, but rather to solve the problem of achieving equilibrium through the formation of value and the distribution of wealth.

[...] On the contrary, as the rise in the real value of silver, in consequence of lowering the money price of corn, lowers somewhat the money price of all other commodities, it gives the industry of the country where it takes place, some advantage in all foreign markets, and thereby tends to encourage and increase that industry. But the extent of the home market for corn must be in proportion to the general industry of the country where it grows, or to the number of those who produce something else, to give in exchange for corn. But in every country the home market, as it is the nearest and most convenient, so is it likewise the greatest and most important market for corn. That rise in the real value of silver, therefore, which is the effect of lowering the average money price of corn, tends to enlarge the greatest and most important market for corn, and thereby to encourage, instead of discouraging, its growth. (Ricardo 1821, Chapter 24, p. 42)

Walras was to argue, with the *esprit de finesse* of his national tradition, that the market was not a single place, but a system of interdependent markets, and therefore, interdependence was its constitutive principle which also linked the participants. From the subjective point of view, the market consisted then of entrepreneurs, brokers, and owners of factors of production (workers, capitalists, and rentiers) connected with each other and with the goods exchanged by a mechanism (the *tâtonnement*) search of equilibrium prices. Walras does propose a unique identification of entrepreneurs, not so much because, according to his theory, they maximize their utility through the pursuit of maximum profit, but because they are characterized as intermediaries between the markets for factors of production and that of the goods. Under equilibrium conditions of production, the entrepreneur does not get either profits or losses (Walras 1877b, p. 232, 1954, p. 225). But equilibrium, for Walras, is actually

a theoretical notion that characterizes the market: it constitutes the normal state to which all variables tend perpetually and automatically in a competitive free economy (Walras 1954, p. 224). Since it contains implicitly the equilibrium of exchange, the market owns the further property of equality between the supply and demand for final and intermediate goods, as well as factors of production.

After Walras, it was Marshall who addressed more explicitly the problem of equilibrium, by characterizing the market as a meeting place of supply and demand. Although based on the comparison of “needs” or “desires” (wants), rather than the incipient neoclassical paradigms (demand, supply, perfect competition), this concept was somewhat overshadowed by the theories of trade in Smith and Ricardo. It needed, however, to be made explicit and descriptive, to accommodate the concept of price formation, bargaining power, and competition. In Book V of his *Principles of Economics*, Marshall begins by tracing a continuum of markets, wider or narrower depending on the position between the two extremes of easily transferable products from open markets and relatively immobile products captured by closed markets:

Thus at the one extreme are world markets in which competition acts directly from all parts of the globe; and at the other those secluded markets in which all direct competition from afar is shut out, though indirect and transmitted competition may make itself felt even in these; and about midway between these extremes lie the great majority of the markets which the economist and the business man have to study. (Marshall, *Principia*, Libro V, Cap. I)

In this continuum, the unique characteristic of markets is that they allow economic agents to find a balance between desires and efforts. The easiest way to find this balance, says Marshall, tracing in this way the classical phylogeny, is to a person who gets what she wants directly from her work. A second tool is the barter, but both of these mean being primitive and limited in their practical consequences; in the end, the emergence of the market appears to be a sheer necessity.

In the market two quantities are compared, subject to variation: the supply price, given by

the minimum price that the agent who proposes the sale is willing to accept for the goods offered, and the bid price, which is the maximum price that the agent who contemplates the purchase is willing to pay. Generally, if the goods are divisible, there are no particular problems in finding the equilibrium point, i.e., the quantity of goods exchanged for which the bid price and the supply price match. There are, however, particular markets where the achievement of this point is not trivial: the labor market, for example, is characterized by the fact that for anyone who offers the sale, only one unit has to be put on the market and may be in need, so that the worker may be willing to accept a very low salary. Her supply price, determined on the basis of the cost of her effort, may not coincide with the price for which she is forced to sell her work to survive.

Marshall's treatment is striking for its clarity but also, as in the case of the classics, for the essentially passive nature attributed to the market. The latter, in addition to be configured as the virtual space of the exchanges, i.e., as a locus operandi with desirable characters of full information, competitiveness, and so on, does not appear to have in any way a coordinating role. The coordination of trade is in fact an automatic consequence of the comparison between supply and demand (or, through the supply price and the bid price, the comparison between effort and desire) and does not even need to resort to the Walrasian auctioneer. In fact, the very concepts of demand and supply prices appear as abstractions, raising the question of balance in an essential way, and can be considered the disembodied protagonists that allow to bypass a theory of the functioning of the market. This is perhaps also due to the fact that, as noted by Colander (1995), Marshall was aware of a higher complexity of the general equilibrium problem and that his conception was more subtle and pervasive than Walras, combining the ideas of multimarket coordination and equilibrium “generality” with those of dynamic adjustment and coordination from a plurality of institutions. While the Walras model is addressed to the solution of a multiple matching problem between goods supplied and goods demanded in an interdependent perspective, Marshall

conceives general equilibrium as a process with “chaotic” characteristics. This process arises from the fact that preferences are constantly updated and economic agents strive to find the equilibrium again and again by moving along “corridors” of coordination which are contingent on the existence and operability of institutions.

A little less than a century later, in a text characterized by the almost total absence of the word “market,” Thorstein Veblen constructs a theory of the entrepreneur, based on the idea that general equilibrium is a dynamic process, depending on the figure of the “captain of industry,” motivated by the opportunities provided by the market to acquire a monopoly position in a economy dominated by machines and processes of product standardization.

Veblen’s model of the “captain of industry,” dear to the frontier capitalism, has the reference specimen of big businessmen, such as the railway investors, whose fortunes, however, consisted not only in the vision and passion for the gains but also in the ability to fight with success against other entrepreneurs and, on the side of consumers, seize the best opportunities for emerging markets. To indicate the wisdom of this opportunistic attitude, Veblen cites, in particular, the saying “charging what the traffic can bear” (Veblen 1904, Chapter 3), which refers to the nascent industry of American Railroads – a service rather than an industry – but characterized by the use of powerful machines.

The market and with it the very notion of equilibrium thus tend to disappear in a heroic perspective of entrepreneurship. Veblen anticipates what will be the central observation of Coase and the starting point of the neo-institutional economics, which identifies the company as an alternative to the market and its command and control structure as an alternative to the impersonal balance between demand and supply. Physiocrats, classical and neo-classical economists, and, in their specificity, even the members of the Austrian school, from Mengel to Böhm-Bawerk, in fact, consider the entrepreneur and the enterprise an integral part of the market, as opposed to the family economy and the model of consumption. In Veblen, however, the idea of a self-governing equilibrium reappears in the role

of the captains of industry as of those who, fighting with other entrepreneurs, to eliminate them from the market, achieve equilibrium by freeing the economy from an excess of management, because “[...] it appears that the greater the amount of pecuniary management, the smaller are the services provided to the community.”

The heroic function of the entrepreneur for Veblen coincides with a form of competition that seeks the disequilibrium and is therefore completely different from that of Walras and Marshall. The experience of American capitalism will inspire 20 years later Max Weber, in attributing a role equally dynamic and heroically ascetic to the moral entrepreneur created by the Protestant Reformation. It is interesting to contrast the points of view of these two authors, Veblen, an economist with sociological propensities, and Weber, a sociologist with inclinations as an economist, because they derive a heroic role for the entrepreneur from completely different views of his moral and creative features. For Veblen, the entrepreneur is spurred by the profit motive and the desire to amass a fortune. The drive towards these goals has its social utility, but to pursue them effectively, the entrepreneur must be substantially amoral and essentially disruptive. For Weber, on the contrary, the success of the capitalist entrepreneur embeds, through the Protestant ethic, moral character and a social quality dramatically higher for the market before birth and consolidation of the reformed religion.

More recently, Kirzner, the most influential living economist of the Austrian school, takes up the theme of the Socratic role of the entrepreneur, making him, however, the principal agent of the search for market equilibrium. According to Kirzner (1973), the function of the entrepreneur as an innovator gives impetus, through her tireless spirit of initiative, to the entire economy. In doing so, the entrepreneur plays a crucial role in correcting market failures, through the activities of arbitrage and speculation. In the process of finding and reaching, the entrepreneur achieves general equilibrium in a far broader sense, as balance, dynamics, and completeness of the markets. In addition, speculative activities, often held to blame, cause substantial benefits for

consumers, because they eliminate the rents due to disequilibrium and push the market towards more efficient conditions. For the entrepreneur to be able to fully release her energy in a beneficial way, it is necessary that the institutional environment is suitable, and entrepreneurial initiative is not mortified by the “traps and snares” of bureaucracy and regulation.

The entrepreneur is therefore the keystone of the economic system because she keeps alive the work, as in the heroic vision of Veblen and Weber, and at the same time, through her constant pursuit of profit, ensures that market activity goes to fruition. As an innovator, Kirzner (1989) argues that the entrepreneur also operates ethically even when she captures high incomes and excess profits. These are in fact the result of the fact that she has discovered the use of resources, thus making to come to life (economic) economic goods that did not exist before. As a beneficiary of distributive justice that rewards merit, the entrepreneur is therefore a fully consistent moral subject.

In the face of these attempts to explain the historical and institutional market, the long wave of the neoclassical school, which is rooted in the principle of specialization and cooperation, reemerges with great force. Wicksteed is perhaps the economist who argues more persuasively, without the use of mathematics, the paradigm and the marginalist theory of equilibrium which will become the core of the so-called neoclassical synthesis. His Presidential Address (1914) contains a concise summary of the marginalist theory of value and distribution and, at the same time, offers an incisive way in the implications of this theory for the nature of the institutions. First of all, the economy is based on a principle of cooperation: “The economic body [...] of an industrial society is an instrumentality for which each man, doing what he can for some of his peers, he gets what he wants from others”. Secondly, the principle of maximization does not imply global complex calculations, but only a comparison algorithm of local differences:

[...] which means, in common parlance, that what an individual will be willing to give something in return, but be rather lacking, is determined by comparison of the difference that he thinks that its

possession would do for him except that he would give anything in return would cause [...].

Moreover, it is this principle of equality of differences that explains both the balance of the individual and that of the market, i.e., between individuals, in the sense that as long as someone believes that an act of exchange may reduce the gap between two opposing differences, he will attempt to make the exchange [...].

What Is General Equilibrium?

Because of its intrinsic complexity and its very diverse history and in spite of its centrality in economic theory, the notion of general equilibrium that has gradually achieved consensus is both subtle and controversial and still subject to alternative interpretations by different scholars. The point of departure concerns the more general notion of “equilibrium,” where, since the physiocrats, the main governing idea is the objective property of a material “balance” between demand and supply of both goods and services. It is this balance, which is surprisingly attained by exchange through unknown and somewhat mysterious means (e.g., Smith’s “invisible hand”), that ensures that the economy is not clogged with unwanted stocks of commodities or plagued by involuntary unemployment or lack of goods or services needed or desired. Material balance generates the further idea that a second, essential property of equilibrium has a subjective nature: in equilibrium agents are content in the sense they feel themselves in a desirable state of balance, because they sell what they want to sell and buy what they want to buy at mutually agreeable rates of exchange. This subjective condition in turn leads to stipulate that in equilibrium agents realize their plans and are satisfied because they behave according to what they believe is right to pursue their objectives and have no need or desire to change.

So far the notion of equilibrium described does not have any special connotation as partial or general, but simply characterizes a condition where things are at rest, in the sense that there is no incentive to change the position of any agent,

either to reduce objective costs or to increase subjective satisfaction, unless underlying exogenous conditions change. In this context, the “partial” equilibrium model emerges from a desire to simplify the analysis of price formation, limiting the interaction between demand and supply and the role of prices to one market or to a subset of interdependent markets. Partial equilibrium, therefore, does not mean that the multiplicity of markets is necessarily ignored nor that any of the properties of equilibrium (material balance, subjective self-fulfilling plans) are discarded but only that one part of the economy is considered exogenous. More specifically, partial equilibrium generally ignores the linkage between price determination and income formation, as well as the wealth effect caused by the fact that the increase or decrease in prices changes the value of agents’ endowment of economic resources.

A further, often tacit, characterization of market equilibrium is that it is “competitive,” i.e., it arises in a market where there is a plurality of agents, each too small to determine, through her individual behavior, the outcomes of the exchange. Of course, monopolistic equilibrium has been cultivated as a theoretical field by several scholars, but it is typically the solution of a noncooperative game and does not possess the combination of objective (material balance) and subjective characteristics of competitive equilibria. As John Nash (1950, 1951) demonstrated, in fact, the notion of subjective equilibrium does not go hand in hand with the accomplishment of material balance. In the theory of noncooperative games, a Nash equilibrium is reached if each player maximizes her objective function under the full knowledge of the strategies of the other players. In such an equilibrium, a higher level of mutual rest is reached in that no player may benefit by changing her strategy if no other player does so. This does not mean, however, that players would not like to change the allocation of goods or services characterizing the equilibrium. In most cases they would, but they cannot, because such an allocation is determined by a higher-level equilibrium among the strategies leading to the allocation. A Nash equilibrium is thus a meta-equilibrium, and while a competitive

market equilibrium qualifies as a Nash equilibrium, a Nash equilibrium is not necessarily a competitive equilibrium and does not necessarily require material balances between demand and supply or agents’ contentment about the allocations realized by the enactment of their strategies.

What is then “general equilibrium”? One is tempted to reply that general equilibrium describes a condition where all markets are in equilibrium, both in the sense that all markets are cleared (demand equals supply) and all agents fulfill their plans. However, while this definition certainly appears simple and direct, it is not complete, since general equilibrium, unlike partial equilibrium, in addition to material balances and subjective fulfillment requires that the distribution of wealth is consistent with resource allocation. General equilibrium theory, in fact, concerns three different circles of causation: (i) between demand and supply of goods and services on one hand and prices and incomes on the other, (ii) between the formation of incomes from demand and supply of factors of production and their prices, and (iii) between the initial resource endowment and the redistribution caused by productive choices and institutional transfers. The precise way in which these three circles interact was not clear until in relatively late times, R. Stone and A. Brown (1962) formalized it in the so-called social accounting matrix (SAM). The focus on the proof of the existence by the celebrated Nobel Prize-winning economists Kenneth Arrow and Gerald Debreu, furthermore, while in itself a benchmark of economic theory, for a long time diverted the attention of the scientific community from more practical features of general equilibrium, such as its computability under alternative scenarios of competition and information and its use for planning and project evaluation. Computable general equilibrium (CGE) models, mainly developed as a result of research efforts at the World Bank in the 1970s, were thus mainly a spin-off of the application of SAM and even in their advanced, present-day form, they tend to evoke the initial dualism between a core set of social accounts and a complementary and highly variable set of behavioral and technical equations.

While at first it may not even be seen as the same subject, since the beginning of its being theorized, the outcome of general equilibrium has been identified with efficiency, in the sense that a fully competitive price system brings about more efficient outcomes than any other practical arrangement or planning mechanism. The linkage between competitive equilibrium and efficiency was formalized by Pareto (1909) and Bergson (1938) and, in its most advanced form, by Arrow (1951) and Debreu (1951) in the form of a full equivalence theorem. Pareto's definition – exemplar in its efficacy – characterizes efficiency as a feasible allocation, i.e., the choice of feasible sets of goods to consume and to produce on the part of the economic agents, that cannot be changed without making at least one agent worse off. Efficiency thus simply implies that it is not possible to modify an efficient allocation without damaging someone. It is the expression of the social trade-off that arises once all possible improvements with no impact on distribution have been made.

Two basic theorems of welfare economics link Pareto efficiency to the notion of general (Walrasian) equilibrium. The first says simply that any general equilibrium is Pareto efficient (Arrow 1951; Debreu 1951), while the second theorem states that under certain, rather general, conditions, given a Pareto-efficient allocation, it is possible to find a corresponding general equilibrium and, in particular, a supporting price vector. These two theorems have important policy implications, because they suggest that Pareto efficiency may be obtained either through a decentralized market system (as in the Walrasian model), by letting prices be freely determined by the exchange mechanism, or as a planned allocation, by imposing a system of prices that will support it.

The relationship between Pareto efficiency and general equilibrium foreshadows the question of its existence, which was taken up by Arrow and Debreu (1954), as well as McKenzie (1959) in a series of celebrated contributions. Their work, although focused on a rather narrow subproblem, essentially proved that under certain conditions aggregate demand and supply could be equated

by a set of nonnegative prices. This was a nontrivial result that was contingent on a series of rather restrictive assumptions but was obtained through a mathematical powerful and unifying instrument (the fixed-point theorem) that was in itself shining for originality and simplicity. The result had two drawbacks, however. First, it did not cover nor it proved to be a feasible base for finding circumstances under which the equilibrium was unique. Second, as proved in a series of important and somewhat astounding later contributions by Sonnenschein (1972, 1973), Debreu (1974) himself, and Mantel (1974), the base of the existence proof was a construct, named “the aggregate excess demand function,” which, although resulting from the aggregation of individual demand and supply, was not bound by the limitations deriving from the postulates of rationality. In what has been called “the everything goes” conclusion, in fact, it was proved that such a function, even though the result of individual rational behavior, is not characterized by any special mathematical property. Thus, the existence of general equilibrium seemed to be quite independent of its “micro-foundations,” as a consequence of an essential weakness of the microeconomic “rationality” assumptions, which were proved to be not sufficiently discriminating to impose anything resembling rationality on aggregate behavior.

General equilibrium theory, however, was not devoid of consequences for applied economics. In a series of important research attempts, in large part conducted at the World Bank, several generations of computable general equilibrium (CGE) models were developed and gradually became an important and useful tool for policy analysis. In these models, social accounting matrices (SAM) became the core of the representation of general equilibrium as a circular flow of production, consumption, and incomes, with prices in all markets as the equilibrating variables. Solving algorithm started with fixed-point (Scarf and Hansen 1973) and mathematical programming procedures (Norton and Scandizzo 1981; Walbroeck and Ginsburg, 1981) and gradually developed into nonlinear equation systems and local or global search solution methods (Devarajan et al. 1997). At present, while the macroeconomic models

prevailing in the 1970s have all but disappeared from the economic practice, CGE models are increasingly used around the world, in both their static and dynamic versions, as tools to analyze economic policy options.

Conclusions

General equilibrium has accompanied the development of economic theory both as a founding concept and as a scientific challenge. Its classical version interprets the market as being characterized by a search for coherence, fulfillment, and balance, from the point of view of both the agents and the goods and services traded. In this interpretation, the historical notion of the market as a place in space and time is progressively changed into that of a virtual space, whose prevalent feature is to harbor the dynamism rather than the locus of exchange. The neoclassical view continues this process of dematerialization of the market and pushes it further by focusing on the abstract nature of equilibrium as a mathematical construct and on its existence and uniqueness properties. The ambition of this second approach, which permeates the founding phase of mathematical economics, is to unify economic theory by providing a general and rigorous connection between microeconomic and aggregate behavior.

While successful in imposing a method of theoretical investigation that remains one of the assets of modern economic theory, the neoclassical program did not go beyond proving a theorem of existence of general equilibrium and of seemingly important but somewhat rarified properties of socially efficient allocations. Embarrassingly, the program also proved the micro-foundation illusory, as a consequence of an intrinsic incapacity of the rationality assumptions to characterize aggregate behavior. Surprisingly, general equilibrium theorizing has been instead more successful in applied economics, by providing, through the social accounting matrix (SAM) and computable general equilibrium (CGE) models, an apparatus of sufficient rigor and sophistication to systematize data collection and frame the analysis of complex economic policies.

References

- Arrow KJ (1951) An extension of the basic theorems of classical welfare economics. In: Neyman J (ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. University of California Press, Berkeley
- Arrow KJ, Debreu G (1954) Existence of an equilibrium for a competitive economy. *Econometrica* 22(3): 265–290. <https://doi.org/10.2307/1907353>. <https://doi.org/10.2307%2F1907353>
- Bergson A (1938) A reformulation of certain aspects of welfare economics. *Q J Econ* 310–334
- Cantillon R (1755) *Essai sur La Nature de Commerce en Général* (ed: Hibbs H). Macmillan, London, 1931
- Colander D (1995) Marshallian general equilibrium analysis. *East Econ J* 21(3)
- Debreu G (1951) The coefficient of resource utilization. *Econometrica* 19(3):273–292
- Debreu G (1974) Excess-demand functions. *J Math Econ* 1:15–21. [https://doi.org/10.1016/0304-4068\(74\)90032-9](https://doi.org/10.1016/0304-4068(74)90032-9). <https://doi.org/10.1016%2F0304-4068%2874%2990032-9>
- Devarajan S, Go D, Lewis J, Robinson S, Sinko P (1997) Simple general equilibrium modeling. In: Francois J, Reinert K (eds) *Applied methods for trade policy analysis*. Cambridge University Press, Cambridge, UK
- Kirzner IM (1973) *Discovery capitalism and distributive justice* (Oxford 1989)
- Mantel R (1974) On the characterization of aggregate excess-demand. *J Econ Theor* 7:348–353. [https://doi.org/10.1016/0022-0531\(74\)90100-8](https://doi.org/10.1016/0022-0531(74)90100-8). <https://doi.org/10.1016%2F0022-0531%2874%2990100-8>
- McKenzie LW (1959) On the existence of general equilibrium for a competitive economy. *Econometrica* 27(1):54–71. JSTOR 1907777
- Nash JF Jr (1950) Equilibrium points in n-person games. *Proc Natl Acad Sci U S A* 36:48–49
- Nash JF Jr (1951) Non-cooperative games. *Ann Math* 54:286–295
- Norton R, Scandizzo PL (1981) General equilibrium computations in activity analysis models. *Oper Res* 29(2):243–262
- Pareto V (1909) *Manual of political economy* (Trans: Kelley AM). New York 1971
- Ricardo D (1821) *On the principles of political economy on taxation*, 3rd edn (ed 1821). John Murray, London
- Robinson J (1962) *Economic philosophy*. Transaction Publisher
- Scarf HE, Hansen T (1973) *The computation of economic equilibria*, vol 24, Cowles Foundation for Research in economics at Yale University, Monograph. Yale University Press, New Haven/London
- Sonnenschein H (1972) Market excess-demand functions. *Econometrica* (The Econometric Society) 40(3): 549–563. <https://doi.org/10.2307/1913184>. <https://doi.org/10.2307%2F1913184>. JSTOR 1913184
- Sonnenschein H (1973) Do Walras' identity and continuity characterize the class of community excess-demand

- functions? *J Econ Theor* 6:345–354. [https://doi.org/10.1016/0022-0531\(73\)90066-5](https://doi.org/10.1016/0022-0531(73)90066-5). <https://doi.org/10.1016%2F0022-0531%2873%2990066-5>
- Stone R, Brown A (1962) *A computable model for economic growth*. Cambridge Growth Project, Cambridge, UK
- Veblen T (1904) *The theory of business enterprise*. Scribner New York
- Walbroeck JL, Ginsburg VA (1981) *Activity analysis and general equilibrium modelling*. North Holland Publishing, Amsterdam, Webley, O. 1986
- Walras L (1954) *Éléments d'économie politique pure, éléments of pure economics* (trans: William Jaffé from the 1926 ed). Richard D. Irwin, Homewood
- Varlas L (1877b) *Éléments d'économie politique pure, 2nd part 1st ed*. L.Cobraz, Lausanne
- Wicksteed PH (1914) *The scope and method of political economy in the light of the "marginal" theory of value and of distribution*. *Econ J* 1–23

Essential Facilities Doctrine

Frédéric Marty
CNRS - UMR 7321 Gredeg, Université Côte d'Azur, Université Nice Sophia Antipolis, Valbonne, France

Definition

The essential facility doctrine is a disputed concept in the field of competition law enforcement. According to this doctrine while a dominant operator controls an asset that its competitors cannot bypass to access the market because of its natural monopoly situation or because the unreasonableness of its replication in financial or in technical terms, it may be bound to provide them an access in fair, reasonable, and non-discriminatory terms. This approach may lead to far reaching remedies and is all the more challenged that it is also implemented to intangible assets.

Introduction

The essential facilities doctrine (hereafter the EFD) stems from the notion of market failure and, even more specifically, from the concept

of natural monopoly. The EFD may be used in a situation in which an economic operator access to the market exclusively depends on the decision of a facility owner without any available alternative. The facility owner may be one of its competitors in the downstream market or a market player who controls an infrastructure which plays as a bottleneck to enter the market. The facility at stake can be an infrastructure, an upstream product or service, or an intangible asset. In other words, the facility owner can be a network operator, an upstream monopolist, or a patent holder. The EFD may be activated within the scope of competition law if this upstream monopolist's refusal to grant an access can be qualified as an exclusionary or an exploitative abuse. It can also be used within the scope of antitrust law, at least in principle, if this refusal participates to a strategy aiming at monopolizing a downstream market.

The denial of access may be an absolute one – it is a clear cut refusal to deal – or a “relative” one. In such a case, the access is still possible but the facility owner imposes unfair or discriminatory conditions in terms of access charges or provides to its downstream competitors an access characterized by deteriorated technical conditions that impair the quality of the service provided to the final user. Such an altered access distorts competition.

The first situation may be commonly observed in deregulated industries. A former State-owned vertically integrated firm, controlling a natural monopoly segment (as the local loop in a telecommunication network, the transportation infrastructure in gas or electricity sector. . .) may impose excessive access charges to its new competitors in the downstream market. Margin squeeze strategies may be at stake in such situations. They may consist in imposing an excessive price for the upstream service and in charging an exclusionary price for the downstream one. Cases as *Deutsche Telekom* (case 37.451, European Commission, 21 May 2003) illustrate this type of strategy, implemented in this instance to impair the access to the market of new competitors (ADSL Internet access providers). The incumbent leverages its dominant position to downstream

competitive markets thought cross subsidizations or by raising its rival costs. Imposing excessive access prices may also be a way to accumulate profit margins to finance exclusionary prices in an associated market in a diversification strategy.

The second situation, characterized by voluntary and artificially degraded technical conditions of access, may be embodied by the *Trinko v. Verizon* decision of the US Supreme Court (540 U.S. 398, 2004). In this case, the incumbent, who controls the essential facility, artificially deteriorated the quality of the service provided to final users by its downstream competitor (here AT&T). In other words, its competitor cannot guarantee to its users the same quality service than the one the incumbent provides to its own customers on the downstream market. The access was not denied but the competition on the merits was structurally distorted.

We first consider the implementation of the EFD to network infrastructures in the USA, before analyzing in a second part its use in the EU context, both for tangible and intangible assets. Our third part opens a discussion.

No Man Is a Prophet in His Own Country: The EFD Under Intense Criticisms from US Scholars

The EFD was initially implemented while a “natural monopoly” infrastructure is at stake. Firstly, we assess to what extent this doctrine may be activated in network industries to impose a mandatory access in fair, reasonable, and not discriminatory conditions to the benefit of the downstream competitors of a vertically integrated former monopoly. Secondly, we consider the distinction between the cases of competition laws and of sector-specific regulations. Its implementation as a regulatory tool in the United States contrasts with the situation in the field of antitrust laws. While the concept was crafted within the scope of the Sherman Act enforcement, its activation within the scope of antitrust laws is increasingly challenged.

Reluctances to Implement the Concept in the Legal Sphere

The EFD was crafted in the field of network industries in the USA in the early twentieth

century. Its first use was related to the natural monopoly issue. As a competition through the infrastructures is technically impossible (or collectively suboptimal in terms of welfare), it is necessary to make possible a competition through the services. It supposes to guarantee a fair, undistorted, and equal access to the natural monopoly to any competitor wanting proposing its services to the final consumers, in the downstream market, in which the competition is possible. The most archetypal implementation of this approach was performed in the 1980s in the telecommunications sector with the AT&T decision.

Promoting a downstream services-based competition despite the existence of an upstream monopolistic bottleneck both implies network open access architecture and a regulation that ensures the effectiveness of a *level playing field*. The 1982 US Department of Justice consent decree in the AT&T case embodied such a logic (United States District Court of the District of Columbia, *US v AT&T Co.*, 552 F. Supp. 131, August 1982). If this case led to the incumbent break up, the key stone of the remedies consisted in a mandatory and undistorted access to the network guaranteed for all the *baby bells*, e.g., for all the downstream competitors in the telecommunication services markets. In the case at hand, the EFD was implemented in the framework of a market building strategy in the context of the sector liberalization. Such a strategy is also at stake through the EU directives promoting a structural unbundling of vertically integrated State owned monopolies in field of utilities. However, the EFD was not crafted in the field of network industries deregulation but as a competition tool 70 years before.

Indeed, the first occurrence of this concept can be found in the US Supreme Court case law in *Terminal Railroad Association* (224 U.S. 383, 1912). The Supreme Court admitted that while the unification of railroad facilities in Saint Louis, Mississippi, was permissible, considering the efficiency gains it produces, it remains that a refusal to grant access to a competitor to these ones may be analyzed as an antitrust law infringement as it discriminatorily impairs its capacities to access to the market.

Thereupon the agreement among competitors can be analyzed as a combination in the purpose to restraint interstate trade. Throughout the twentieth century several US courts decisions implemented this concept to unilateral and coordinated practices. According to this case law, an access to a bottleneck may be compelled as soon as a competitor has no alternative to enter the market. It was for instance the case in *Associated Press* (326 U.S. 1, 1945), *Lorain Journal* (42 U.S. 143, 1951), *Otter Tail* (410 U.S. 366, 1973), *Hecht Football* (570 F.2d 982, D.C. Cir., 1977), or in *Aspen Skiing* (427 U.S. 585, 1985).

However, the concept of essential facility, and more precisely its implementation within the scope of Antitrust laws, was formally rejected in harsh terms by the US Supreme Court in 2004 in its *Trinko* ruling. Quite surprisingly, according to the Supreme Court, the essential facilities doctrine was never endorsed by its own case law. Insofar as it is not an antitrust concept, it may be used, at best, in the field of sector specific regulation. . . If the US Supreme Court overturned its own long-standing jurisprudence (from 1912 to 1985), it may be related both to the specific conditions of this case (a class-action brought by a law-firm searching for damages) and to the evolution of the academic literature regarding this issue since the 1960s. Indeed, legal scholars have increasingly seen in the implementation of EFD a violation of fundamental rights as property rights or freedom to contract (Areeda 1990).

Such a mistrust is not so surprising considering other courts decisions both from the *Lochner* era (the conservative US Supreme Court case law before the progressive turn of the late 1930s' and the post war *Warren* era) and from the *Chicagoan* turn of the late 1970s. The Supreme Court stated in *Colgate* (250 U.S. 300, 1919) that a company has the power to choose with whom to contract and to set its conditions. Since the market power does not result from a *monopolization* strategy, antitrust laws have not to impose to contract with any other undertaking. In 1980, the *Berkey Photo v. Eastman Kodak* decision (444 U.S. 1093, 1980), confirms that a monopoly, since its

position results from its own merits, can charge the price it decides.

These two judgments have to be placed into their respective proper contexts. The first one is representative of the *classical legal thought*. Neither antitrust laws nor public regulations have to impair fundamental rights. The second one is typical from the *Chicagoan* approach of antitrust laws enforcement. The *GTE Sylvania* decision (*Continental Television v. GTE Sylvania*, 433 U.S. 36, 1977) widened the scope of the rule of reason at the detriment to *per se* prohibitions. The *Reiter Sonotone* decision (*Reiter v. Sonotone Corp.*, 442 U.S. 330, 1979) endorsed the consumer welfare approach. The *Eastman Kodak* decision confirmed the *Chicagoan* view according to which a monopoly can lawfully extract the competitive rents it had created. In this framework, Antitrust laws only have to sanction the extension of a market power on other basis than the merits. Extracting the surplus created is considered both as legitimate and as desirable in terms of market dynamic (Carlton and Heyer 2008).

Indeed, the Sherman Act, promulgated in 1890, does not prohibit the mere possession of a monopoly power. A monopoly position is not an antitrust issue in and of itself while its acquisition and its maintenance stem from the undertaking's own merits. As the Supreme Court stated in *Grinnell Corp* (384 US 563, 1966) "The offense of monopoly under § 2 of the Sherman Act has two elements: (1) the possession of monopoly power in the relevant market and (2) the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident." As a consequence, the notion of merits goes far beyond the scope of considerations narrowly related to past investments or to the risks initially taken by the incumbent.

Such a view contrasts with the underlying logic of the Judge Learned Hand's ruling in *Alcoa* (148 F.2d 416, 2d Cir. 1945). According to the current dominant conception of antitrust laws enforcement, a monopolist is not bound to adjust its market behavior in order to guarantee a *living*

profit to its competitors. On the contrary, Learned Hand considered that a vertically integrated undertaking that enjoys a monopoly position on the upstream market must set its prices in a manner that allows its downstream competitors to remain on the market, whatever their efficiency. This position is at odds with US current case law according to which a *legitimate* monopolist has no duty towards its downstream competitors. However, the situation may be different in regulated industries for which the market positions are not only related to the merits. This is precisely the point stressed by the Supreme Court in *Trinko*: the EFD is no longer seen as an antitrust-related concept but only as remedy available to a sector specific regulator to organize the third party access to a *natural monopoly* network.

A Risk in Terms of Incentives According to the Economic Literature

Economic scholars have also harshly criticized the implementation of the EFD in the competition law field. Its broad implementation may lead to bind a successful firm to make its competitors benefit from its past investments, e.g., to free-ride them. Such a mandatory access may pertain to an asymmetric regulation of competition and it may have a deterring effect on its incentives to invest. *Ex post*, imposing an access to the incumbent's assets can be seen as an expropriation. The undertaking controlling the essential facility is deprived of the right to refuse the access and of its capacity to set freely its price. The economic gains resulting from its investment will be partially lost. *Ex ante*, anticipating such a risk might deter to invest in this type of asset. Depending on the level of the access charge, it may also impair its incentives to spend money in its maintenance or in its development, as the dominant undertaking cannot expect to extract all the economic gains resulting from its investment. In addition, it may even impair the incentives to invest for nondominant undertakings. As the new entrants may prefer to use the incumbent's assets, it may reduce the incentives to develop a new (essential) facility. Indeed, this case law may lead to expect the possibility to be bound to provide an access benefiting to competitors and in doing so to reduce the

potential pay-offs associated to the initial investment.

As a consequence, imposing a mandatory access may be welfare-enhancing only in a static sense: it effectively increases the competition on the downstream market. However, in a dynamic sense, it may be a counterproductive remedy as it hinders the incentives to invest in the infrastructure and in strategies aiming at by-passing it. The consumer welfare gain resulting from the competitive remedy may be only a short-term one. Investments aiming at producing disruptive innovations may be discouraged and the asset owner may be daunted to invest to increase the quality of the service provided by its infrastructure or its capacities because it will also benefit to its competitors. In a worst-case scenario, the incumbent might be incentivized to under-invest to generate socially counter-productive bottlenecks in order to have an objective reason to deny the access to its competitors.

In addition, the very cautious approach of the economic literature considering the EFD implementation is even exacerbated for intangible assets as intellectual property rights (hereafter IP). For instance, how to characterize an intangible asset as absolutely necessary to a competitor to access the market and reasonably impossible to replicate in technical or in financial terms? Such preventions echo with the Stigler (1968) views according to which barriers to entry cannot really be technical or financial but are more commonly the product of public regulations. Moreover, public intervention through antitrust remedies are viewed with suspicion. It is particularly the case since these remedies are implemented within market building objectives (or in order to counterbalance individual market power) or since they aim at correcting the excess of IP rights protection. The risk of regulatory capture cannot be excluded.

Indeed, we may put into relief the long-standing mistrust of economic scholars towards antitrust suits initiated by competitors. The risk is to protect unduly competitors at the expense of consumers (Bork 1978). In addition, the incentives produced by competitive threats might be still possible despite a monopoly situation. The competitive pressure persists for as long as the

market remains a constable one. Compelling a monopoly to renounce to its contractual freedom and to limit its capacity to extract all the surplus created can eventually be prejudicial in terms of consumer welfare. In other words, compulsory licensing or mandatory access may lead to an asymmetric regulation of competition. If the EFD one may be acceptable as a regulatory tool in the framework of a sector specific liberalization, it is by far more debatable since the dominant undertaking's position is due to its merits and not to former exclusive rights.

A Broad Acclimatization in the EU Competition Law Enforcement

Even through the essential facility doctrine is now rejected as an antitrust tool in the USA, its implementation by the EU Commission and by the Member States' competition authorities is significant. The reason of this adherence may be explained by the EFD theoretical consistency with some of the EU competition policy underlying principles. Firstly, the EFD tackles the issues related to deny to access to the market (exclusionary abuse) or to excessive access prices (exploitative abuses). Secondly, as soon as the purpose of the competition law is conceived as encompassing the building competitive markets or the guarantee a level playing field, the EFD appears as an efficient tool to craft remedies to achieve such ends.

Implementation to Network Industries

The first uses of the EFD in the EU context naturally took place in the field of the liberalization of network industries with the 1992 EU Commission's decision in the case *B&I – Sealink* (Commission, case IV/34.174 – Sealink/B&I – Holyhead, 11 June 1992). It was also the case, 4 years later in the French case the 1996 competition authority's decision related to the Narbonne's heliport (French Competition Council, decision n°96-D-51, 3 September 1996). Through the EU Commission led liberalization process, the EFD was successively applied to telecommunications, electricity, gas, and railways sectors. We will see below that one additional special feature of the EU use of the EFD compared

to the US case has to be taken into account: its large implementation to intangible assets. For instance, incumbents are commonly bound to share their consumer's data bases with new entrants in order ensure a level playing field (see for instance the French Competition Authority decision, n°17-D-06 related to practices in the markets of gas natural, electricity, and energy related services, 21 March 2017).

The EFD was often implemented as a competitive remedy in the sectors concerned by a liberalization process. The first step of liberalization commonly consists in guaranteeing an open (and regulated) access to the incumbent's infrastructure for the new entrants in order to favour the development of a services-based competition. The second step may be an unbundling (incumbent's vertical de-integration) or the implementation of the so-called *ladder of investment approach* (Bourreau et al. 2010).

The vertical unbundling is a drastic way to guarantee an undistorted access. It consists in the vertical integrated incumbent break up, as it was the case in the 1982 *AT&T* US case. These remedies were implemented in the energy sector exclusively through negotiated procedures (see for instance the *ENI* Commission decision 39.315, 29 September 2010).

The ladder of investment approach is a less intrusive intervention (without mandatory divestitures as structural remedies) that can be implemented since the competition is not potentially limited to the service but can be expanded to a part of the facilities. This approach consists in providing a broad access to the incumbent's facilities in the first steps of the liberalization process in order to lower barriers to entry at the benefits of its new competitors. The expected effect consists in helping them to acquire a sufficient customer's portfolio to overcome investors' risk aversion to finance their own infrastructures. This broad access will be progressively reduced in order to promote incentives to invest both for them and also for the incumbent, since the first ones cannot durably free ride its own investments. This approach allows to incentivize the new entrants to shift from a services-based competition to a facilities-based one. It does not concern the

natural monopoly in itself but all the facilities necessary to the connection with this one. Naturally regulated access to the natural monopoly segment remains at the end of this period. One of the interests of such a facilities-based competition is to favour a service differentiation among the competitors, for instance in terms of quality. It can be seen as broad way to implement the EFD within the context of a sector specific liberalization.

Even so, the case of the ladder of investment approach illustrates some of the common risks and pitfalls of the EFD implementation in network industries. On one hand, if the access charge is too low, the new entrants may free ride incumbent past investments and can be discouraged to invest in their own assets. On the other hand, the incumbent's incentives to develop and to maintain its own infrastructures may be reduced by such a mandatory access. Setting the access charge raises several difficulties. A high access charge level may erect a barrier to entry and unduly preserve incumbent's rents. Even if the access charge has to be oriented to its costs, it remains that information asymmetries might lead to significant regulatory costs and to possible suboptimal settings. In the same time, the access to the incumbent's facilities may be counterproductive for the new entrants as it might perpetuate a situation of dependence toward their main competitor. This dependence on the facilities controlled by their main competitor exposes them to the consequences of noncooperative behaviors as raising rival costs strategies (through the pricing of nonregulated ancillary services) or strategies aiming at artificially reducing the quality of the service they provide to final users.

The EU Commissions have faced sharp criticisms about its broad implementation of the EFD in the field of network industries. The concept of ladder of investments, presented above, illustrates that the scope of the essential facility can go far beyond the natural monopoly segment narrowly defined. The risk is to excessively extend the notion of essential facilities to all the specific and expensive assets that a new entrant needs to dispose to enter the market. Investing in some specific assets may actually represent a significant risk as their amortization can be difficult for a new

entrant who only serves a small segment of the market. However, it is not so rather clear at the theoretical point of view that such investments can be considered as barriers to entry (see for instance Stigler 1968). The risk is to promote thought the EFD implementation an asymmetric regulation of the competition by mandating the incumbent to provide an access not only to its essential facilities but also to *convenient facilities* for its competitors (Ridyard 2004). The risk is to favor inefficient entries at the expense on final consumers.

A striking example of such concerns can be provided by the EU Commission *GVG* decision in 2003 related to an access of a German company to the Italian railways. The mandatory access covers all the components of the traction services encompassing the locomotives and the railway workers qualified for the Italian network themselves (European Commission, case COMP/37.685, *GVG/FS*, 27 August 2003). On the contrary, in the *Bronner* decision, the Court of Justice had decided that a mandatory access to an incumbent facility cannot be imposed since the new entrant can develop its own (Court of Justice, judgment C-7/97, *Oscar Bronner GmbH & Co. KG v Mediaprint Zeitungs, 26 November 1996*). An essential facility should be a non-replicable one and not only a less expensive one. The EU case law has been overturned at the benefit of a broader definition of the notion of the essential facility.

Another spectacular example of the broad implementation of the EFD in the EU competition policy applied to the liberalisation of network industries can be provided by a French competition authority decision related to the electricity market in December 2007, *Direct Energy* (French Competition Council, decision n°07-D-43 related to EDF market practices, 10 December 2007). These far reaching remedies can be explained by the very specific context of the French liberalization characterized by the coexistence of regulated tariffs at the retail level and market prices at the wholesale one. The new entrants were bound to set their retail prices at this threshold to attract customers. In the same time, as they did not control their own generation

capacities, they must buy the electricity on the wholesale market. Since, the regulated retail price is set according to the generation marginal cost of the plants, mainly nuclear, operated by the incumbent, they were potentially squeezed since the wholesale price might be higher than this last one. The competitive claim was grounded on this margin squeeze produced by the combined effect of the downstream (regulated) price and the upstream (market) price. The competition law-based remedy consisted in implementing the EFD to the energy generated by the nuclear power plants (de Hauteclocque et al. 2011). The new entrants benefit from a drawing right on the incumbent's nuclear plants originated generation. The EFD was implemented to energy itself through this quasi-structural remedy. Such types of remedies tend to ensure that the new entrants may be as efficient as the incumbent. The use of the EFD is not limited to guaranteeing an access to the market for the competitors of a vertically integrated incumbent but participates to a regulation of the competition. In other words, the access does not concern a network infrastructure but the upstream good. The purpose is to ensure a level playing field in the downstream market. . . even if it supposes to "subsidize" potentially less efficient competitors. The EFD in this context leads to horizontal wealth transfer among market players. In this instance, it was used to address a regulatory imperfection.

Implementation to Intellectual Property Rights

Beyond this broad implementation, the EU implementation of the EFD is also distinguishable from the US one, as it also broadly concerns intangible assets as intellectual property rights. The EFD was implemented in the EU case law on TV programs, data bases, and to interoperability devices between softwares. While the EFD as competition remedies may be admitted as the essential facility at stake resulted from exclusive rights, the case of intangible assets raises concerns as soon as the assets are the products of incumbent's merits. In the same time, these far reaching remedies can be analyzed as a mean to counterbalance the excessive protection of the incumbents produced by intellectual property rights. They can also be

analyzed as a mean to prevent an unchallengeable monopolization of a given market or to counteract leveraging strategies. The leverage from the then near-monopoly Windows operating system was one of the main concerns 11 years ago (EU Court of First Instance, case T-201/04, *Microsoft Corp. v. European Commission*, 17 September 2007); (EU Commission, 18 July 2018, case 40099) and some online platforms raise nowadays the same types of competition issues.

In a historical perspective, the first implementation to the EFD to intangible assets in the case of the EU competition law took place in the 1990s with the Irish TV programs case (EU Court of Justice, cases C-241/91P and C-242/91P, *RTE v European Commission*, 6 April 1995). The dominant broadcasting operator edited its own TV programs magazine and refused to provide its TV listings to a new entrant who aimed at proposing a multi-channels TV programs magazine. The 1995 EU Court of Justice rulings imposed to communicate to this entrant the TV listings by considering that the refusal impair it to propose a new product to consumers. A second emblematic case was the IMS Health one (EU Court of Justice, case C-418/01, *IMS Health GmbH v NDC GmbH*, 29 April 2004). An undertaking proposing software-based reporting tools for pharmaceutical sales data refused that a new entrant adopts the same database modular structure than its one, based on German Zip codes. The EU Court of Justice dismissed its claims in 2004 and considered that it must provide a licence for allowing its competitor to use the same database structure despite its protection through intellectual property rights. What is specific in this second case is that there is no new product at stake as it was the case in the 1996 ruling.

This point is all the more relevant that the initial EU case law imposed to fulfil several criteria to decide of a compulsory licensing remedy under the EFD. The 1995 Magill decision was for instance grounded on the future availability of a new product for consumers. In the IMS judgment, this condition was not required. The access may be required even if the new competitor will provide the same service in the same relevant

market. In the same way, in the Microsoft case, the Commission imposed as a remedy the sharing of the interface protocols with its OS, considering that the interoperability will favor the future development of new products. Such pragmatism was also at stake concerning another condition to implement the EFD for intangible assets as for instance the requirement to implement a balance of incentives before imposing a licensing. Indeed, the EU case law initially imposed to perform such a balance in order to assess the net effect on the incentives to invest and to innovate both for the incumbent and for the beneficiary of the compulsory licensing. Such an assessment, which echoes with the more economic approach of the competition law enforcement – e.g., its effects-based approach – is seldom, or almost never, implemented in the EU case law related to the EFD (Marty and Pillot 2012). In the EU Commission guidance related to its enforcement priority regarding the article 82 of the Treaty (now the article 102), issued in February 2009, only three criteria remain: the refusal must relate to a product or a service objectively necessary to effectively compete on the market; the refusal may lead to an elimination of the effective competition; and this refusal is likely to lead to consumer harm).

Discussion

Implementing the EFD for intangible assets raises specific concerns (Castaldo and Nicita 2007). Setting the licensing fee is one of the more delicate aspects. Defining the access charge for a physical network may already be a challengeable task. Even though the principles are clear (the access price should be oriented toward the cost), the influence of accounting choices and the imperfections of information may lead to significant difficulties. Moreover, if we consider the Direct Energy remedies how to set the real cost of an electricity generated by nuclear power plants? The solution applied in this specific case – an auction procedure – illustrates the difficulties encountered to implement such remedies. Things are even worse for intangible assets for which the cost are by far more difficult to assess. The access

and replication costs are negligible. The initial investments and the risks taken cannot be easy to evaluate and may be irrelevant, as testifies the IMS case, in which the essential facility at stake was an IP right on the structure of data base in accordance with Zip codes.

In addition, a broad implementation of the EFD to intangibles might lead to price regulation and to some extend to competition regulation. Defining the terms of a FRAND licence terms (*Fair, Reasonable, and Non Discriminatory*) or balancing market powers among competitors impose to define the concepts of fairness and reasonableness. Furthermore, using the EFD to promote the liberalization of a network industry, or even worse to make a dominated market contestable, may lead the competition authority to act as a sector specific regulator aiming at favoring the development of the competitors, possibly at the expense of final consumers.

Finally, the EFD implementation as a competition law remedy illustrates the transatlantic divergences on several topics. First, it stresses the debates on the respective scopes of competition law enforcement and sector specific regulation. Second, it illustrates the divergences about the possibility to use competition remedies to address issues as the excessive protection granted by intellectual property laws. Third, it underlines the disagreements on the use of competition law remedies as tools to prevent or to correct excessive concentrations of economic powers, even if these ones mainly result from the merits.

However, a striking point in the analysis of the implementation of the EFD in the EU competition law enforcement should be put into relief. It consists in the sharp contrast between the frequency of its use and the reluctance of the competition authorities to ground decisions and to justify remedies explicitly on this doctrine. Several reasons might be highlighted. A first one may be found in its lack of well-grounded economic theory foundations, especially in the framework of an effects-based approach (Geradin 2004). A second one may be related to the relative fuzziness of the concept at legal point of view and to the possible risks induced in terms of decision reversal in the judicial control. As a consequence, the doctrine is

in the same time significantly used to shape competition law remedies in the EU case and rarely qualified as a legal basis to ground them.

Cross-References

► [Negotiated Procedures in EU Competition Law](#)

References

- Areeda P (1990) essential facilities: an epithet in need of limiting principles. *Antitrust Law J* 58:841
- Bork R (1978) *The antitrust paradox*. Free Press, New York
- Bourreau M, Doan P, Manant M (2010) A critical review of the “ladder investment” approach. *Telecommun Policy* 34(11):683–696
- Carlton D, Heyer K (2008) Assessing single-firm conduct. *Comp Policy Int* 4(2):285–305
- Castaldo A, Nicita A (2007) Essential facility access in europe: building a test for antitrust policy. *Rev Law Econ* 3:83–110
- de Hauteclocque A, Marty F, Pillot J (2011) The essential facilities doctrine in European competition policy: the case of the energy sector. In: Glachant J-M, Finon D, de Hauteclocque A (eds) *Competition, contracts and electricity markets: a new perspective*. Edward Elgar, London, pp 259–292
- Geradin D (2004) Limiting the scope of Article 82 of the EC treaty: what can the EU learn from the US Supreme Court’s Judgment in *Trinko in the Wake of Microsoft, IMS, and Deutsche Telekom*. *Common Mark Law Rev* 41:1519–1533
- Marty F, Pillot J (2012) Intellectual property rights, interoperability and compulsory licensing: merits and limits of the European approach. *J Innov Econ* 9(1):35–61
- Ridyard D (2004) Compulsory access under EU competition law: a new doctrine of “convenient facilities” and the case for price regulation. *Eur Comp Law Rev* 11:669–674
- Stigler GJ (1968) *The organization of industry*. University of Chicago Press, Chicago

Ethical Review Board

► [Institutional Review Board](#)

EU Citizenship

► [European Nationality](#)

EU Microsoft Competition Case

Luca Rubini

University of Birmingham, Birmingham, UK

Abstract

The investigation and litigation that involved Microsoft before the EU authorities in the 2000–2012 period is arguably the first case that put the complex issues of the new economy to the test of EU competition laws. The case, which attracted a lot of attention, also beyond competition law circles, was a complex one, at various levels – technical, economic, and legal – and one that eventually interrogated the ultimate goals of competition policy in a given legal system. This entry takes the reader through the various issues and findings of the European Commission and the Court of First Instance of the European Union before offering few notes of analysis and a view on the future implications of this landmark case.

EL classification K4 Legal Procedure, the Legal System, and Illegal Behavior.

Background and Significance of the Case

“The Microsoft case was the monopolization case of the new economy just as *Standard Oil* had been the icon of the old” (Peritz 2010). This quote makes reference to the US case involving Microsoft and draws a momentous parallel with the famous *Standard Oil* case of 1911 where the US Supreme Court found Standard Oil guilty of monopolization of the petroleum industry and famously divided the company into several different and competing companies. This quote gives a very good idea of the importance of the competition law tangles that have involved Microsoft, the American new economy giant, in the USA. Equally important was the investigation and litigation that involved Microsoft in the EU in the 2000s and which produced some of the most

controversial decisions of the EU Commission and of the General Court.

Microsoft, founded by Bill Gates in the 1970s in California, has fast become one of the most successful companies of our time. It belongs to that inner circle of businesses strongly characterized by a unique blend of innovation and creativity and vision and aggressiveness – which can often be attributed to the leadership of unique founder individuals. Microsoft operates in some of the most crucial and innovative markets. Its products, and notably its operating systems and software applications, are ubiquitous. Their importance cannot be understated. They have influenced, indeed shaped, everyday life of billions of people around the world. Commercial success and especially market power attract the attention of antitrust control. Microsoft has thus been subject to various investigations and actions, based on antitrust laws, at both sides of the Atlantic. Since the beginning of the 1990s, Microsoft started to be the subject of the scrutiny of US authorities (on the US case, see Page and Lopatka 2007; Peritz 2010; Gavil and First 2014). This culminated in a much-publicized litigation at the turn of the millennium where it was charged with the allegedly illegal tying of the Internet Explorer web browser with the Windows operating system. Microsoft would have leveraged on the sheer dissemination of its PC operating system to extend its dominance in the web browser market. Following, among other things, a change in administration (from Clinton to Bush Jr.) and numerous settlements between the various parties involved, this litigation finished with no action. The focus shifted to the other side of the Atlantic. At the beginning of the 2000, the EU Commission officially started to investigate two conducts of Microsoft – the tying of its media player with Windows and the refusal of certain interoperability information to communicate with its server operating systems – which, in a momentous decision of 2004, were concluded to be breaches of EU antitrust law. In 2007, in a long-awaited decision, the General Court of the EU essentially confirmed the findings and rulings of the EU Commission. In 2009, the EU Commission started a new investigation focused (like the US

case) on the bundling of Windows with Internet Explorer, which was settled the next year.

After summarizing the arguments and the findings of the EU case (section “[The Decisions of the EU Commission and the General Court](#)”), this entry provides few notes of analysis of the technical, economic, and legal aspects of this case (section “[Analysis of the Case](#)”). This analysis paves the way to the consideration of the broader implications of the case (section “[Implications for the Future](#)”).

The Decisions of the EU Commission and the General Court

Article 102 of the Treaty on the Functioning of the EU (TFEU) (formerly, Article 82 of the EC Treaty) was the key provision at the center of both the EU Commission investigation and the litigation before the General Court. In the parts relevant to the Microsoft case, this provision, which dates back to the original Treaty of Rome of 1957, reads:

Any abuse by one or more undertakings of a dominant position within the internal market or in a substantial part of it shall be prohibited as incompatible with the internal market in so far as it may affect trade between Member States.

Such abuse may, in particular, consist in:

...

(b) limiting production, markets or technical development to the prejudice of consumers;

...

(d) making the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.

The reader can find in the literature (Rubini 2010) a uniquely interesting repetition of the EU investigation and litigation, performed by its very same actors: the Commission (Banasevic and Hellström 2010), Microsoft (Forrester 2010; Bellis and Kasten 2010), and the General Court (Vesterdorf 2010). The following section summarizes the key arguments and findings of the case. This overview will introduce the reader to the various levels of complexity of this momentous case.

The Decision of the EU Commission

After a lengthy investigation, the EU Commission issued its decision on 24 March 2004. It concluded that Microsoft had violated Article 82 of the EC Treaty by abusing its dominant position in client PC operating system (OS) in two ways. Firstly, it had illegally refused to supply interoperability information which was indispensable for rival vendors to compete in the work group server operating system market. Secondly, it had illegally made the availability of the Windows client PC operating system conditional on the simultaneous acquisition of the Windows Media Player software. The Commission ordered Microsoft to disclose interoperability information and appointed a trustee to monitor the compliance with this duty. It also ordered Microsoft to provide a version of Windows without Windows Media Player. The Commission finally imposed a fine of 497,196 million euros for what it found to be a very serious infringement of Article 102 TFEU.

With respect to the **licensing** part of the case, the Commission found that Microsoft had breached Article 102 TFEU by refusing to supply interoperability information to its competitors. As noted, letter (b) of the same provision enlists: “limiting production, markets or technical development to the prejudice of consumers.” Two Microsoft products, and markets, were under examination in the licensing claim: first, the Windows PC OS and, second, the Windows Server OS. The Commission found that, with market shares of, respectively, 93% and 60%, Microsoft was superdominant and dominant in the relevant product markets. The servers’ market was the one where there had been, and there still was, some competition. Some companies were competing with Microsoft and were concerned by the reduction in the interoperability information available for communication between their servers and Windows PCs. The main issue of contention concerned the degree of interoperability, and amount of information, which should be guaranteed. The Commission opined that Microsoft refused information that was indispensable for external servers to interoperate with Windows, which limited innovation and risked eliminating competition to the detriment of consumers. In

economic terms, the scenario was one of leveraging in the server OS market on the basis of the virtual monopoly held in the PC OS market. The significant network effects that are typical of these markets exacerbated the effects of this exclusionary conduct (Gil-Moltó 2010). The Commission therefore ordered Microsoft to license on reasonable and nondiscriminatory terms the relevant interoperability information to its competitors.

With respect to the **tying** claim, the Commission found that Microsoft had infringed Article 102 TFEU by bundling its Windows Media Player with its operating system. This had enabled Microsoft to expand its market power and foreclose competition in the media player market. The Commission based its decision on four steps: (i) the tying (i.e., Windows) and tied (i.e., Windows Media Player) products are separate products, (ii) Microsoft is dominant in the market for Windows, (iii) it does not give its customers a choice to obtain the tying product without the tied product, and (iv) this practice forecloses competition (in market of media players). The Commission thus ordered Microsoft to offer a version of Windows without the Windows Media Player.

The Judgment of the EU General Court

With a decision dated 17 September 2007, the General Court (at the time Court of First Instance) of the EU upheld the substance of the Commission’s decision with respect to both abuses as well as the fine. The General Court’s decision was not appealed before the Court of Justice of the EU.

The Duty to Disclose Interoperability Information
Before the General Court, Microsoft put forward various arguments to discredit the assessment of the EU Commission of the need for a duty to disclose interoperability information. Firstly, the concept of interoperability used by the Commission was too broad, and to follow it would have meant to enable Microsoft’s competitors to have access to highly technologically innovative protocols and to clone its products. Secondly, the Commission did not follow the strict conditions that, according to the case law, need to be present before antitrust law can impose a duty to license

intellectual property. In particular, the presence of five alternative routes for interoperability defeated the existence of indispensability. In addition, a high threshold for antitrust intervention should be followed. The presence of competing servers in the market proved that competition was not being eliminated. Moreover, the emergence of no “new product” was prevented by Microsoft’s conduct. Finally, Microsoft’s refusal was justified on the grounds of protecting valuable intellectual property, and the forced disclosure of information would encourage copying and reduce incentives to innovate.

The General Court in turn rejected each of these arguments.

Firstly, the General Court confirmed the degree of interoperability the Commission found to be necessary to remain viable in the market and in the specific context of a Windows work group network (paras. 229–230). In particular, the Court found that the Commission was right in concluding that “the common ability to be part of [the Windows domain architecture] is a feature of compatibility between Windows client PCs and Windows work group servers” (para. 189). Consequently, the Commission was held to be correct in considering that the required interoperability information should cover “the complete and accurate specifications for all the protocols [that are] implemented in Windows work group server operating systems and that are used by Windows work group servers to deliver file and print services and group and user administration services, including the Windows domain controller services, Active Directory services and Group Policy services, to Windows work group networks” (para. 195). The Court also noted that this information does not extend “to the internal structure or to the source code of its products” (para. 206). The Court also rejected Microsoft’s argument that this degree of interoperability would have allowed its competitors to clone its products or certain features of those products (paras. 234–242).

Secondly, the General Court also found the Commission did not err when it found that the information concerning the interoperability with the Windows domain architecture was **indispensable**. The availability of the required information was

necessary in order to be able to compete viably with Windows work group server OSs on an equal footing.

On the one hand, the Court rejected that the five alternatives to disclosure could ensure the necessary degree of interoperability. Microsoft had put forward five methods which, though not ensuring the perfect substitutability that the Commission considers essential, made nonetheless it possible to achieve the “minimum level of interoperability required for effective competition,” “to work well together” (paras. 345–346). It is interesting to note here that, in its Decision, the Commission had already rejected “reverse engineering” as a viable alternative to disclosure of interoperability information (see para 562 of the Decision).

On the other hand, on the issue of what level of **elimination of competition** is necessary to trigger antitrust intervention, Microsoft contended that the Commission had been satisfied with establishing a mere “risk” of the elimination of competition in the work group server OS market. What should have been required was the “likelihood” or, in other words, the “high probability” of distorting competition. The General Court responded that Microsoft’s complaint was “purely one of terminology” and “wholly irrelevant” (para. 561). The Court went on noting,

The expressions ‘risk of elimination of competition’ and ‘likely to eliminate competition’ are used without distinction by the Community judicature to reflect the same idea, namely that Article [102 TFEU] does not apply only from the time when there is no more, or practically no more, competition on the market. If the Commission were required to wait until competitors were eliminated from the market, or until their elimination was sufficiently imminent, before being able to take action under Article [102 TFEU], that would clearly run counter to the objective of that provision, which is to maintain undistorted competition in the common market and, in particular, to safeguard the competition that still exists on the relevant market.

In this case, the Commission had all the more reason to apply Article [102 TFEU] before the elimination of competition on the work group server operating systems market had become a reality because that market is characterized by significant network effects and because the elimination of competition would therefore be difficult to reverse. . . .

Nor is it necessary to demonstrate that all competition on the market would be eliminated. What matters, for the purposes of establishing an infringement of Article [102 TFEU], is that the refusal at issue is liable to, or is likely to, eliminate all effective competition on the market. It must be made clear that the fact that the competitors of the dominant undertaking retain a marginal presence in certain niches on the market cannot suffice to substantiate the existence of such competition (paras. 561–563).

Applying this reasoning to the actual market data, the General Court concluded that “Microsoft’s refusal has the consequence that its competitors’ products are confined to marginal positions or even made unprofitable. The fact that there may be marginal competition between operators on the market cannot therefore invalidate the Commission’s argument that all effective competition was at risk of being eliminated on that market” (para. 593).

Citing the previous case law, Microsoft argued that it has not been established that its refusal prevented the appearance of a “new product” for which there is unsatisfied consumer demand. The General Court noted that the appearance of a new product “cannot be the only parameter which determines whether a refusal to license an intellectual property right is capable of causing prejudice to consumers” (para. 647). Such prejudice may arise, as the Commission had found, also when there is a limitation of technical development, as provided in Article 102 TFEU, letter (b) (*ibid.*).

The General Court then endorsed the Commission’s finding that “the information at issue does not extend to implementation details or to other features of Microsoft’s source code,” representing “only a minimum part of the entire set of protocols implemented in Windows work group server operating systems” (paras. 657–658). Nor – the Court noted – “would Microsoft’s competitors have any interest in merely reproducing Windows work group server operating systems.” Once the necessary interoperability information is made available, “they will have no other choice, if they wish to take advantage of a competitive advantage over Microsoft and maintain a profitable presence on the market, than to differentiate

their products from Microsoft’s products” (para. 658).

Finally, the General Court noted that the mere fact that the relevant technology is covered by intellectual property protection is not per se a sufficient objective justification (para. 690). Microsoft had not sufficiently established that, if it were required to disclose the information, disclosure would have a significant negative impact on its incentives to innovate (para. 701).

The Bundling of the Windows Media Player

As regards the tying claim, the analysis immediately focused on whether the Commission introduced new law in the area by examining whether the dominant undertaking “does not give customers a choice to obtain the tying product without the tied product” (para. 845). The General Court dismissed Microsoft claim that this was any different from what Article 102 EC (d) requires, i.e., that bundling assumes that consumers are compelled, directly or indirectly, to accept supplementary obligations, such as those referred to in Article 102(d) EC (para. 864). The Court points out that “coercion is mainly applied first of all to OEMs [Original Equipment Manufacturers], who then pass it on to the end user” (para. 865). The analysis then shifted to the question whether the Commission introduced a new condition, analyzing whether Microsoft’s conduct foreclosed competition. Microsoft accused the Commission of having added a new test, nor provided for in the law (and in particular Article 102 TFEU, letter d), notably that the tying conduct is foreclosing competition (para. 846). The Court acknowledged that, in light of the specific circumstances of the case, the Commission did not “merely assume, as it normally does in cases of abusive tying, that the tying of a specific dominant product has by its nature a foreclosure effect” (para. 868). By contrast, it “examined more closely the actual effects which the bundling had already had on the streaming media player market and also the way in which that market was likely to evolve” (*ibid.*). The Commission was right in doing this, since the “list of abusive practices set out in the second paragraph of Article 82 EC is not exhaustive and that the practices

mentioned there are merely examples of abuse of a dominant position” (para. 860). More specifically, bundling by an undertaking in a dominant position may also infringe Article 82 EC where it does not correspond to the example given in Article 82(d) EC. Accordingly, in order to establish the existence of abusive bundling, the Commission was correct to rely in the contested decision on Article 82 EC in its entirety and not exclusively on Article 82(d) EC (para. 861). The Court, in any event, concluded that constituent elements of abusive tying identified by the Commission coincide effectively with the conditions laid down in Article 82(d) EC (para. 862).

The second point of contention focused on whether media functionality is a separate product from the PC operating system, which is a necessary condition to have tying. Microsoft argued this was not the case (para. 912). Media functionality forms an integral part of the operating system with the consequence that what is at issue is a single product. This would have been confirmed by customers’ expectation that any PC OS have essential audio and video functionalities. While recognizing the rapid evolution of the industry, where products that are initially separate are then subsequently regarded as forming a single product, the Court concluded that the Commission was correct to find two separate products in the period of the investigation (paras. 913–914). The Court noted that there exists a separate consumer demand for streaming media players (para. 917). The simple fact that two separate products are complementary does not exclude their difference (paras. 921–922). The Court confirmed the Commission’s finding that media functionality is not linked, by nature or by commercial usage, to PC OSs (para. 938). Firstly, it does not seem that the two constitute by nature indissociable products (para. 939). In any event, the Court importantly underlines that “it is settled case-law that even when the tying of two products is consistent with commercial usage or when there is a natural link between the two products in question, it may none the less constitute abuse within the meaning of Article 82 EC, unless it is objectively justified” (para. 942). Secondly, “it is difficult to speak of commercial usage in an industry that is 95%

controlled by Microsoft” (para. 940). Furthermore, the fact that vendors of competing client PC OSs also bundle those systems with a media player was not conclusive (para. 941). The Court crucially noted that “it is settled case-law that even when the tying of two products is consistent with commercial usage or when there is a natural link between the two products in question, it may none the less constitute abuse within the meaning of Article [102 TFEU], unless it is objectively justified” (para. 942). Microsoft’s argument that the integration of Windows Media Player in the OS was dictated by technical reasons was found not to be substantiated (para. 937).

There was no real contention with respect to Microsoft’s dominance in the PC OS market (para. 870) or with respect to the inability of consumers to obtain Windows without Windows Media Player (para. 961).

The Court then moved to the actual assessment of how Microsoft’s conduct would have foreclosed competition; the General Court upheld the Commission’s analysis.

Microsoft contended that the Commission, recognizing that it was not dealing with a classical tying case, had to apply a new and highly speculative theory, relying on a prospective analysis of the possible reactions of third parties, in order to reach the conclusion that the tying at issue was likely to foreclose competition. The General Court noted that “it is clear that, owing to the bundling, Windows Media Player enjoyed an unparalleled presence on clients PCs throughout the world, because it automatically achieved a level of market penetration corresponding to that of the Windows client PC operating system” (para. 1038). The pre-installation of Windows Media Player made users less likely to use alternative players (para. 1041). The said bundling created disincentives for OEMs to ship third-party media players on their client PCs for technical (e.g., higher usage of hard-disk space, risk of confusion on the part of users, increase of customer support and testing costs) and economic (e.g., higher price of the PC) reasons (paras. 1043–1045). The Court also found that the Commission was also correct to find that methods of distributing media players other than

pre-installation by OEMs could not offset Windows Media Player's ubiquity (paras. 1049–1057). The Court confirmed the Commission's finding that the market for streaming media player was characterized by significant indirect network effects (paras. 1061–1076). That expression describes the phenomenon where the greater the number of users of a given software platform, the more there will be invested in developing products compatible with that platform, which, in turn, reinforces the popularity of that platform with users (para. 1061).

Finally, Microsoft's arguments that the integration of media functionality in Windows was "indispensable in order for software developers and internet site creators to be able to continue to benefit from the significant advantages offered by the "stable and well-defined" Windows platform" were rejected (para. 1146). The Court further noted that, although standardization may be positive (but not necessarily wanted by third parties), it cannot be allowed to be imposed unilaterally by a dominant firm (paras. 1152–1153). Similarly, the claim that the Commission was interfering with Microsoft's business model was not accepted (paras. 1149–1150). The Commission did not deny that Microsoft could provide a version of Windows with Windows Media Player. It objected to the fact that this is the only available version and, more specifically, that no version of the ubiquitous OS is offered without Windows Media Player.

Analysis of the Case

The summary exposition of section II gives a good flavor of the complexity of this case. Indeed, it is various factors of technical, economic, and legal nature, each adding to the other, that make it complex. In turn, this complexity explains how the assessment of both the EU Commission and the General Court could be so controversial and generate much debate. Quite probably, the most interesting remarks in this discussion come from Bo Vesterdorf, the President of the General Court in the very same case, who, focusing on the duty to disclose interoperability information, expressed

concerns about the possible negative impact of the General Court's decision on investment and innovation (Vesterdorf 2008).

The first layer of complexity comes from the highly **technical** nature of the subject matter. What clearly comes out from an overview of the literature is that, without a proper understanding of the "basic technology" issues, any economic or legal assessment is doomed to fail (see the excellent primer of Jackson 2010). The "technical basis" of the EC case "was significantly more complex than its US counterpart," but, unfortunately, "many of the articles written about the Microsoft proceedings fail to convey the technical complexity of the issues at stake" (Jackson 2010). It is sometimes (wrongly) assumed that the computing sector is just like any other industry and that a stylized account of the technical facts would suffice. It should also be highlighted that it is in particular the "interoperability" part of the case that requires an unusually detailed level of technical understanding, which probably explains why this claim attracted particular attention. To name just few of the technical issues of the case and indeed the most general ones: What is interoperability? In particular, what degree of interoperability is necessary for computer programs to be meaningfully interfaced? Consequently, what type of information is necessary? Apart from disclosure, what paths are available to software developers to obtain this information? What difficulties do they each involve? How effective are they to ensure interoperability? What does define a set of instructions as a computer program? Is it possible to distinguish the latter from a mere functionality? When do separate computer programs stop being so and become one single program? Can functionalities be removed from a computer program without impairing the integrity of the program? (for a discussion of these questions, see Walsh 2010; Andreangeli 2010).

The second level of complexity, which is directly based on the first one, concerns the **economics** of this sector and of the behavior of consumers and competitors. Some unique features characterize the computer industry, such as its very fast pace of development and innovation and the high relevance of network effects (see

Gil-Moltó 2010; Walsh 2010; Liebowitz and Margolis 2001; Evans et al. 2000). Thus, on the one hand, the tendency toward network effects makes foreclosure more likely. On the other hand, the fast development of the industry where today's winner is tomorrow's loser (and vice versa) represents a crucial concern for legal and regulatory processes, in two respects. First, the speed and uncertainty of change may make the assessment of harm and efficiencies difficult. Second, the time required for investigations may render legal processes continually and, inherently, obsolete. Equally, in such a dynamic environment, remedies are difficult to be tailored and effective (Economides and Lianos 2010).

The third, and final, level of complexity lies in the law. After almost 60 years from its entry into force, the provision at issue, Article 102 TFEU, is still vague. "Despite a significant volume of case law expanding, refining, clarifying and apparently applying the law, there is still considerable uncertainty about the exact conditions amounting to abuse" (Walsh 2010). The fact is that Article 102 TFEU, with its broad and general language, is a "standard" (rather than a "rule"), which means that the regulator or judge has to define its content on a case-by-case basis (see Kaplow 1992). This process may lead to more accurate results but is certainly more costly and uncertain. This uncertainty is particularly acute in technological sectors like those of the new economy and does significantly depend on the ambiguous signals that come from the first two layers of complexity set out above.

More deeply, the root of the uncertainty of Article 102 TFEU depends on the fact that its meaning changes depending on the different conceptions of what competition and antitrust law should be about, conceptions that succeed and prevail one after the other. In 60 years (and quite probably in many years to come), the language is (and will be) the same, but the understanding of competition policy and of the regulation of unilateral behavior by dominant firms has changed (and will change). What is competition law about? Is it (only) about efficiency? What do we mean by efficiency? Is competition law more about safeguarding a competitive process, characterized

by market access and contestability, or, more simply, an efficient outcome of the market contest? What other competition policy objectives play within the frame of competition law? (For commentary around these questions, see Kerber 2008 and Fox 2008.)

That being said, one has to ask whether the Commission and the General Court have really been revolutionary in their legal analysis. We focus here on the interoperability issue. In particular, did they really depart so dramatically from the previous *Magill* and *IMS Health* case law and their formulation of the "exceptional circumstances" under which a duty to deal and supply can be enforced? Fox (2010) rightly observes that "[t]he Microsoft facts did not fit the factors very snugly; but they fit the concept of essentiality much better than the facts of either *IMS Health* or *Magill*". While the General Court (and the Commission) may have been essentially right, Fox goes on noting that it would have been "much more satisfying" if the General Court had more openly and directly asked the following questions:

- (1) Are consumers and the market seriously disadvantaged by denial of full access to interoperability information? If the answer is "yes":
- (2) Would the respondent and the market be seriously disadvantaged by a duty to grant access?

According to the facts on file, it seems that first answer would indeed have been positive (in the Court's decision, it is repeatedly mentioned that users preferred certain rivals' products on all qualities except interoperability) and the second one negative (it should again be remembered that, according to the facts, before achieving a significant presence in the market of server OSs, Microsoft provided complete interoperability information).

A similar exercise – but with probably a more uncertain outcome – could be carried out for bundling (see Andreangeli 2010). In brief, one could ask whether, through the bundling of the Windows Media Player with Windows, consumers

were really “coerced” to use Windows Media Player and, as a result, consumers (and competitors) were harmed by this conduct.

The relativity of competition conceptions, and the ensuing flexibility of antitrust laws, and in particular the regulation of unilateral behavior (the most flexible of them all), becomes especially apparent if a parallel is made between how the EU dealt with the conduct of Microsoft and how the US treated it – or, better, would treat it. This is particularly apparent with respect to the interoperability issue and to the imposition, via antitrust laws, of the duty to supply. As Fox (2010) again shows:

Any analyst applying the law and spirit of *Trinko* [landmark decision where the US Supreme Court emphatically underlined that there is no duty to deal in American antitrust law] would not start the analysis with the question posed above: Are consumers seriously disadvantaged by work group suppliers’ lack of seamless access to the standard operating-system network? Analysis would start with quite a different question: Why should Microsoft be ordered to share its property with anyone, let alone rivals? US courts generally presume that a duty to deal will seriously impair a monopoly firm’s incentives – to the harm of the market and innovation.

A similar conclusion on the EU-US divide can be taken for the claim of illegal bundling. As shown by the Court of Appeals of DC in the US Microsoft case, US Courts do require a rule or reason assessment and, in so doing, set the threshold pretty high before concluding that bundling is anticompetitive in a new economy setting (see Andreangeli 2010).

Implications for the Future

Many have highlighted the true exceptionality of the facts of the Microsoft case, which would suggest that it is difficult to draw any lessons for the future. Still, we believe, two main implications can be detected.

Microsoft was the first in a new string of cases dealing with the new economy and forcing antitrust laws, which in decades had developed with more traditional industries, to be confronted with extremely difficult questions. More than ever

economic and technical knowledge are essential. Current (at the time of writing) investigations like *Google* raise important questions on the definition of the relevant markets and of abuse in highly complex and interconnected market (see Pollock 2010; Lianos and Motchenkova 2012; Bork and Sidak 2012).

Secondly, it has been noted above that the EU has taken a “typically European” approach, which in the end is definitely more interventionist as compared to the US one. One may wonder whether this will continue in other new economy cases or whether a rapprochement with what happens across the Atlantic can be expected.

References

- Andreangeli A (2010) Tying, technological integration and Article 82 EC: where do we go after the Microsoft case? In: Rubini L (2010), pp 318–343
- Banasevic N, Hellström P (2010) Windows into the world of abuse of dominance: an analysis of the Commission’s 2004 Microsoft Decision and the CFI’s 2007 judgment’. In: Rubini L (2010), pp 47–75
- Bellis JF, Kasten T (2010) The Microsoft Windows Media Player tying case. In: Rubini L (2010), pp 127–165
- Bork RH, Sidak JG (2012) What does the Chicago School teach about Internet search and the antitrust treatment of Google. *J Comp Law Econ* 8(4):633–700
- Economides N, Lianos I (2010) The quest for appropriate remedies in the EC Microsoft cases: a comparative appraisal. In: Rubini L (2010), pp 393–462
- Evans DS, Fisher FM, Rubinfeld DL, Schmalensee RL (2000) Did microsoft harm consumers? Two opposing views. AEI-Brookings Joint Center for Regulatory Studies, Washington, DC
- Forrester IS (2010) *Victa placet mihi causa*: the compulsory licensing part of the Microsoft case In: Rubini L (2010), pp 76–126
- Fox EM (2008) The efficiency paradox. In: Pitofsky (ed) *How the Chicago School overshot the mark: the effect of conservative economic analysis on US antitrust*. Oxford University Press, New York, pp 77–100
- Fox EM (2010) The EC Microsoft case and the duty to deal: the transatlantic divide. In: Rubini L (2010), pp 274–281
- Gavil AI, First H (2014) *The Microsoft antitrust cases. Competition policy for the twenty-first century*. MIT Press, Cambridge, MA
- Gil-Moltó MJ (2010) Economic aspects of the Microsoft case: networks, interoperability and competition. In: Rubini L (2010), pp 344–368

- Jackson C (2010) The basic technology issues at stake. In: Rubini L (2010), pp 3–46
- Kaplow L (1992) Rules versus standards: an economic analysis. *Duke Law J* 42:557–629
- Kerber W (2008) Should competition law promote efficiency? Some reflections of an economist on the economic foundations of competition law. In: Drexel J, Idot L, Moneger J (eds) *Economic theory and competition law*. Edward Elgar, Cheltenham/Northampton, pp 93–120
- Lianos I, Motchenkova E (2012) Market dominance and quality of search results in the search engine market, TILEC discussion paper DP 2012/036, 31 October 2012
- Liebowitz SJ, Margolis SE (2001) *Winners, losers & Microsoft. Competition and antitrust in high technology*. The Independent Institute, Oakland
- Page WH, Lopatka J (2007) *The Microsoft case*. Chicago University Press, Chicago
- Pertitz JRR (2010) The Microsoft chronicles. In: Rubini L (2010), pp 205–257
- Pollock R (2010) Is Google the next Microsoft? Competition, welfare and regulation in online search. *Rev Netw Econ* 9(4):1–29
- Rubini L (2010) Microsoft on trial. Legal and economic analysis of a transatlantic antitrust case. Edward Elgar Publishing, Northampton
- Vesterdorf B (2008) Article 82 EC: where do we stand after the *Microsoft* judgment? *Glob Antitrust Rev* 1–41
- Vesterdorf B (2010) Epilogue. In: Rubini L (2010), pp 487–489
- Walsh A (2010) *Microsoft v Commission: interoperability, emerging standards and innovation in the software industry*. In: Rubini L (2010), pp 282–317

Eucken, Walter

Stefan Kolev

Wilhelm Röpke Institute, Erfurt and University of Applied Sciences Zwickau, Zwickau, Germany

Abstract

The purpose of this entry is to delineate the political economy of Walter Eucken. To reach this goal, a history of economics approach is harnessed. First, the entry concisely reconstructs Eucken's life, intellectual evolution, and heritage. Second, it presents the specificities of his "theory of orders" and his "order-based policy" gateway to a rule-based political economy.

Biography

Walter Eucken (1891–1950) was doubtlessly one of Germany's most important twentieth-century economists, especially in the field of political economy. The Freiburg School, which he co-initiated in the 1930s as an interdisciplinary group of economists and legal scholars, has had a crucial impact on the evolution of postwar German economics and legal scholarship, but also on the practical trajectory in the economic policies of the Federal Republic and of European integration. This introduction aims at embedding Eucken in his time and at depicting his role in several contexts in science as well as at the interface between science and society, before subsequently turning to his contributions to political economy.

Eucken was born in Jena into the family of the philosopher and later Nobel Prize laureate Rudolf Eucken. He studied a combination of history, economics, law, and administrative science at Kiel, Bonn, and Jena and was heavily influenced in his socialization in economics by the Younger Historical School, even though his teachers were not totally hostile to theorizing (Goldschmidt 2013, pp. 127–129). Having completed a dissertation (at Bonn) and a habilitation (at Berlin) on historicist grounds, he received a call to Tübingen in 1925 and subsequently to Freiburg in 1927, where he remained for the rest of his life. What would later become famous as the Freiburg School of Ordoliberalism came into existence in the early 1930s, when Eucken formed a scholarly community with the law professors Franz Böhm (1895–1977) and Hans Großmann-Doerth (1894–1944), also attracting talented younger scholars like Friedrich Lutz (1901–1975) and Leonhard Miksch (1901–1950) to Freiburg (Goldschmidt and Wohlgemuth 2008a). The cooperation of Eucken with Böhm and Großmann-Doerth steadily intensified, with the book series "Order of the Economy" initiated in 1936 as a milestone – its introduction under the title "Our Mission" became the programmatic manifesto of the incipient ordoliberal understanding of the role of law and economics in science and in society (Böhm et al. 2008; Goldschmidt

and Wohlgemuth 2008c). In the adverse intellectual climate of the time, Eucken and Böhm actively rejected National Socialism in theory and in practice, with Eucken openly opposing Martin Heidegger's rectorate policies in the university senate in 1933 (Nicholls 1994, pp. 60–67; Klinckowstroem 2000, pp. 85–88). Eucken was also among the very few who remained openly loyal to Edmund Husserl until his death in 1938, with his own epistemology heavily influenced by Husserl's phenomenology. Also, in the late 1930s and early 1940s, Eucken was an active participant in intellectual resistance groups, later to become known as the Freiburg Circles (Glossner 2010, pp. 31–38). Despite the isolation particularly during the war, Eucken remained connected to intellectual allies like F.A. Hayek and Wilhelm Röpke and in the immediate postwar years became a seminal figure in the international revitalization of liberalism, among others during the founding years of the Mont Pèlerin Society (Kolev et al. 2014). Simultaneously, Eucken was one of the principal policy advisors to the allies in Germany and also to Ludwig Erhard's increasingly influential policy strategy, later to become famous as the Social Market Economy (Goldschmidt and Wohlgemuth 2008b, pp. 262–264; White 2012, pp. 233–238). Eucken passed away in March 1950 while giving a lecture series at the London School of Economics upon Hayek's invitation.

While the relevance of ordoliberalism for post-war Germany and Europe is discussed in the entry dedicated to ordoliberalism, the impact of Eucken's heritage as a person deserves special attention. In 1954, the Walter Eucken Institute was founded by friends, colleagues, and students in the house of his family in Freiburg and is until today an influential research institute and think tank. Eucken's personality still attracts paramount attention: in 2014 Federal President Gauck delivered an address in Freiburg upon the 60th anniversary of the Walter Eucken Institut, in 2016 Chancellor Merkel delivered an address in Freiburg upon Eucken's 125th birthday, while today's economists argue about "Walter Eucken's long shadow" in the German policy responses to the Euro crisis (Feld et al. 2015; Bofinger 2016) and

his relevance for related rule-based research programs like Constitutional Political Economy (Vanberg 1988; Buchanan 2012; Köhler and Kolev 2013).

Order, Science, and Policy

As described above, Eucken lived in a time as hostile as possible to liberty – and in a time which was characterized by extreme degrees of arbitrariness and chaos. This is one of the reasons why Eucken and his associates focused their political economy and social philosophy on the concept of "order" – a term which is among the most complex and most multifaceted in intellectual history (Anter 2007, pp. 127–158). Eucken's specific research program took shape relatively late in his life – it "matured slowly" (Hayek 1951, p. 337) – and can be interpreted as concentrating upon two central goals: first to enable a higher degree of order for designing economic theory and second to enable a higher degree of order for conducting economy policy. These two goals are also the focal points of two key terms in Eucken's system and of his two major books: *Ordnungstheorie* as his contribution to economic theory, presented 1940 in *The Foundations of Economics* (1992), and *Ordnungspolitik* as his contribution to economic policy, presented 1952 in his posthumous *Principles of Economic Policy* (2004). Important to underscore and to show below, both Eucken's theory and his perspective on economic policy are very much in line with the "problem of constitutional choice, i.e., as a question of how desirable economic order can be generated by creating an appropriate economic constitution" (Vanberg 2001, p. 40). The two domains of theory and policy offer a helpful structure for continuing this exposition.

Theory of Orders: An Alternative to the Ruins of Historicism and Pure Abstraction

German economics has been famous, and at times notorious, for its inclination toward extensive methodological and epistemological debates.

The notoriety stems from periods like the Weimar Republic, when many economists continued fighting about methods and epistemology without being able (or willing) to tackle the pressing problems of economic life (Köster 2011, pp. 41–60). Eucken's project was targeted at the very opposite: with his "Foundations" he attempted to identify a solid methodological and epistemological basis for theorizing not as an aim in itself, but rather to enable such kind of theorizing which can alleviate the immense problems of the age after National Socialism. This section focuses on those theoretical concepts that have direct practical impact and are thus indispensable for understanding Eucken's political economy.

First and foremost, the separation between "order" and "process" is crucial: "economic order" for Eucken is to be understood as the sum of market forms and monetary systems which frame the interactions of the individuals, whereas "economic process" stands for the interactions themselves. Synonymous for the framework of the economic order are "the rules of the game," whereas the economic process can be translated as "the moves of the game" (Eucken 1992, pp. 223–232, 2004, p. 54). An additional cornerstone of Eucken's system is "the interdependence of orders" – a concept emphasizing the embeddedness of the economic order within the other social orders, notably the law and the political order of the state. This is of special relevance since it shows that even though the different social orders have their own individual logics, these orders have to be thought in their diverse interrelations. Also, the combinability of economic, legal, and political orders is not arbitrary, i.e., specific combinations like a centrally planned economy and rule of law are hardly stable in the long term. At the same time, he underscored that having a well-developed rule of law is not a sufficient condition for a well-ordered market economy – rather, it is through the cooperation of economists and legal scholars that specific principles of the rule of law will be identified as indispensable preconditions and prerequisites for the market economy (Eucken 1992, pp. 85–90). Eucken's theory of orders thus made

it possible to do away with the ruins of atheoretical or even antitheoretical historicism left by the Younger and the Youngest Historical Schools well into the 1930s, but at the same time he warned that abstract theorizing can be dangerous if its results are applied to any orders without carefully considering their specificities of time and space (Eucken 1992, pp. 41–44).

Order-Based Policy: An Alternative to Laissez-Faire, Central Planning, and Interventionism

The careful and systematic shaping of the economic order (synonymously: of the economic constitution) is at the core of Eucken's political economy. The above theoretical distinction of "order" and "process" enabled him to coin his famous *Ordnungspolitik*. A term difficult to translate precisely into any other language, it has been translated into English with the general term "rule-based policy" or with the more specific one "order-based policy." Its central target is a clear-cut conceptual alternative to laissez-faire, to the centrally planned economy, and to interventionism, using two criteria which Eucken distilled as essential for judging the merits of economic orders: the material criterion if an order is productive (i.e., overcoming scarcity as much as possible), and the ideal criterion if an order is humane (i.e., enabling a life in self-determination for the individuals) (Eucken 1992, pp. 239–241). Neither a laissez-faire regime nor a centrally planned economy fulfill these two criteria – a laissez-faire regime is deficient in terms of the framework of its economic order (with the implied belief in its automatic self-generation), while a centrally planned economy is deficient in terms of the properties of its economic process (and the impossibility of a self-determined life in it) (Eucken 1948). Eucken's "order-based policy" aims at optimally setting the rule-based framework of its economic order by a strong state envisioned as a referee impartial vis-à-vis vested interests, a state which (unlike the arbitrariness of interventionism) abstains from interventions into the economic

process except in well-defined exceptions (Giersch et al. 1992, pp. 28–32). Such a policy is aimed at fighting the first and foremost evil of social life: according to the Freiburg School, this evil is always the phenomenon of power in all forms of power relations – stemming from the state, from the market, and from other social orders (Eucken 2004, pp. 175–179; Foucault 2008, pp. 129–158). As first presented by Böhm and discussed in the entry covering his work, the ordoliberal instrument against power is competition. Eucken’s contribution here is to specify the kind of principles a “competitive order” – the order conforming both to the “productive” and to the “humane” criterion – has to follow (Oliver 1960, pp. 133–140). He devised two sets of principles, which are at the core of this “competitive order”: the “constitutive” and “regulating” principles. The former comprise a functioning price system, a sound currency, open markets, private property, freedom of contract, liability, and constancy of economic policy, while the latter aim at additionally curbing monopoly power, income inequalities, externalities, and problems of the labor market (Sally 1998, pp. 111–114; Kolev 2015, pp. 428–430).

Many of these problems may seem almost trivial today, but at the time of Eucken’s death, they were anything but trivial – not in Europe and not in the United States, as the parallels between Eucken’s endeavor and the project of the “Old Chicago” School clearly indicate (Van Horn 2009; Buchanan 2012; Köhler and Kolev 2013). Other problems, especially the issues of power in the digital age, the lack of liability in large sections of the political order and the financial system, or the search for maxims how to avoid the arbitrariness of interventionism and instead to base economic policy on rules, have lost nothing of their relevance.

Cross-References

- Austrian School of Economics
- Böhm, Franz
- Constitutional Political Economy

- Hayek, Friedrich August von
- Mises, Ludwig von
- Ordoliberalism
- Political Economy
- Röpke, Wilhelm
- Schmoller, Gustav von

References

- Anter A (2007) Die Macht der Ordnung. Aspekte einer Grundkategorie des Politischen, 2nd edn. Mohr Siebeck, Tübingen
- Bofinger P (2016) German macroeconomics: the long shadow of Walter Eucken. CEPR’s Policy Portal of June 7 2016. Available at: <http://voxeu.org/article/german-macroeconomics-long-shadow-walter-eucken>
- Böhm F, Eucken E, Großmann-Doerth H (2008) Unsere Aufgabe. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, p 27–37. English translation In: Peacock A, Willgerodt H (eds) (1989) Germany’s social market economy: origins and evolution. St. Martin’s Press, New York, p 15–26
- Buchanan JM (2012) Presentations 2010–2012 at the Jepson School of Leadership, Summer Institute for the History of Economic Thought, University of Richmond. Available at: <http://jepson.richmond.edu/conferences/summer-institute/index.html>
- Eucken W (1948) On the theory of the centrally administered economy. An analysis of the German experiment, parts I and II. *Economica* 15(58):79–100. 15(59): 173–193
- Eucken W (1992) The foundations of economics. History and theory in the analysis of economic reality, 2nd edn. Springer, Berlin
- Eucken W (2004) Grundsätze der Wirtschaftspolitik, 7th edn. Mohr Siebeck, Tübingen
- Feld LP, Köhler EA, Nientiedt D (2015) Ordoliberalism, pragmatism and the eurozone crisis: how the German tradition shaped economic policy in Europe. *Eur Rev Int Stud* 2(3):48–61
- Foucault M (2008) The birth of biopolitics. Lectures at the Collège de France 1978–1979. Palgrave Macmillan, New York
- Giersch H, Paqué K-H, Schmieding H (1992) The fading miracle. Four decades of market economy in Germany. Cambridge University Press, Cambridge
- Glossner CL (2010) The making of the German post-war economy. Political communication and public reception of the social market economy after world war II. Tauris Academic Studies, London
- Goldschmidt N (2013) Walter Eucken’s place in the history of ideas. *Rev Austrian Econ* 26(3): 127–147

- Goldschmidt N, Wohlgemuth M (2008a) Entstehung und Vermächtnis der Freiburger Tradition der Ordnungsökonomik. In: Goldschmidt N, Wohlgemuth M (eds) *Grundtexte zur Freiburger Tradition der Ordnungsökonomik*. Mohr Siebeck, Tübingen, pp 1–16
- Goldschmidt N, Wohlgemuth M (2008b) Social market economy: origins, meanings, interpretations. *Constit Polit Econ* 19(3):261–276
- Goldschmidt N, Wohlgemuth M (2008c) Zur Einführung: Unsere Aufgabe (1936). In: Goldschmidt N, Wohlgemuth M (eds) *Grundtexte zur Freiburger Tradition der Ordnungsökonomik*. Schweizer Monatshefte, Tübingen, pp 21–25
- Hayek FA (1951) Die Überlieferung der Ideale der Wirtschaftsfreiheit. *Schweizer Monatshefte* 31(6): 333–338
- Klinckowstroem W (2000) Walter Eucken: Eine biographische Skizze. In: Gerken L (ed) *Walter Eucken und sein Werk*. Mohr Siebeck, Tübingen, pp 53–115
- Köhler EA, Kolev S (2013) The conjoint quest for a liberal positive program: “Old Chicago”, Freiburg, and Hayek. In: Peart SJ, Levy DM (eds) F. A. Hayek and the modern economy: economic organization and activity. Palgrave Macmillan, New York, pp 211–228
- Kolev S (2015) Ordoliberalism and the Austrian school. In: Boettke PJ, Coyne CJ (eds) *The Oxford handbook of Austrian economics*. Oxford University Press, Oxford, pp 419–444
- Kolev S, Goldschmidt N, Hesse J-O (2014) Walter Eucken’s role in the early history of the Mont Pèlerin Society. Discussion paper 14/02. Walter Eucken Institut, Freiburg
- Köster R (2011) Die Wissenschaft der Außenseiter. Die Krise der Nationalökonomie in der Weimarer Republik. Vandenhoeck & Ruprecht, Göttingen
- Nicholls AJ (1994) Freedom with responsibility. The social market economy in Germany, 1918–1963. Clarendon, Oxford
- Oliver HM (1960) German neoliberalism. *Q J Econ* 74(1):117–149
- Sally R (1998) Classical liberalism and international economic order: studies in theory and intellectual history. Routledge, London
- Van Horn R (2009) Reinventing monopoly and the role of corporations: the roots of Chicago law and economics. In: Mirowski P, Plehwe D (eds) *The road from Mont Pèlerin: the making of the neoliberal thought collective*. Harvard University Press, Cambridge, MA, pp 204–237
- Vanberg VJ (1988) “Ordnungstheorie” as constitutional economics. The German conception of a “social market economy”. *ORDO Jahrb Ordn Wirtsch Ges* 39: 17–31
- Vanberg VJ (2001) The Freiburg school of law and economics: predecessor of constitutional economics. In: Vanberg VJ *The constitution of markets. Essays in political economy*. Routledge, London, pp 37–51
- White LH (2012) The clash of economic ideas. The great policy debates and experiments of the last hundred years. Cambridge University Press, Cambridge

European Community Law

Klaus Heine¹ and Anna Sting²

¹Rotterdam Institute of Law and Economics, Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

²International and European Union Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Abstract

European Community law has evolved over the last 60 years as a process of European economic integration. European law has developed alongside the stages of economic integration in a reciprocal process, with euro-area Member States being close to total economic integration. However, economic integration has not been unhalted – political and economic realities have forced the EU to adapt to current developments and negotiate treaty amendments but sometimes also to act outside the Treaty framework. Other influences, such as interest groups, have also shaped the dynamics of interest groups.

Synonyms

[European law](#); [European Union law](#)

Definition

European community law consists of the law that is found in the Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU). In addition European community law comprises regulations, directives, or decisions, as well as the soft-law form of recommendations and opinions.

The Legal Framework of the European Union

Introduction: The History of Integration and Main Legal Texts

The European Union (EU) as we know it today is a result of a process of European integration that

has lasted for the most part of the last seven decades. While this might seem like a long time, it is only a blink of an eye in the overall history of the European continent.

The European Union has evolved from what was earlier called the European Community, which in turn was based on three communities founded in 1951 (the European Coal and Steel Community (ECSC)) and more importantly in 1957 (the European Economic Community (EEC) and the European Atomic Energy Community (EURATOM)) by means of the treaties of Paris and Rome, respectively (Lenaerts and van Nuffel 2011). While the Treaty on the Coal and Steel Community was concluded for 50 years only, the Communities as of 1957 were entered into for an unlimited period (art. 356 TFEU). Six Member States, Belgium, France, Germany, Italy, Luxembourg, and the Netherlands, committed themselves to the process of economic integration and gave away sovereign rights to a supranational order. They agreed to have some policy areas governed by four principal institutions: The Council as a representation of Member States, the Commission (or at that time the High Authority), the Assembly (later to be the European Parliament), and the Court. The Council and the Commission were triplicated for all three communities but joint for the purposes of efficiency and convenience by the Merger Treaty in 1965 (Craig and De Burca 2011; Hartley 2014; Lenaerts and van Nuffel 2011).

While the ECSC Treaty was very specific as regards the policy to be pursued by the institutions, the EEC treaty left much more leeway to the institutions to fill in their objectives. In addition, the mainly economic aims of the Communities were complemented by ideas of a political, social, and cultural union. Probably the greatest step for the Communities was the establishment of the European Union by the Maastricht Treaty in 1992, which also paved the way for the Economic and Monetary Union (Hartley 2014). This treaty was founded on the communities and introduced two more policy domains to be dealt with by cooperation: a common foreign and security policy and cooperation in the field of justice and home affairs (Craig and de Burca 2011).

However, the process of European integration did not only concern the deepening of integration but also the widening. Membership became open for other States in Europe: Even though the UK was initially reluctant to join the EU, it ended up applying for membership, first blocked by Charles de Gaulle but later realized in a group accession including Ireland, Denmark, and Norway in 1972. In the years 1979 and 1981, Greece, Spain, and Portugal joined, before some of the EFTA countries, Sweden, Finland, and Austria, acceded to the Union in 1994. The biggest accession process, however, concerned a group of former communist countries in the East of Europe (Lithuania, Latvia, Estonia, Poland, the Czech Republic, Slovakia, Hungary, Slovenia) and two Mediterranean islands, Cyprus and Malta (Nugent 2010). More recently, Bulgaria and Romania have joined and the currently last Member to the EU is Croatia, which acceded in July 2013 (Hartley 2014).

The Treaty of Lisbon, a result of the failed constitutional treaty for the European Union, replaced the European Community by the European Union as the sole entity and led to a deepening of the EU competences and codified institutional developments in the Union (Craig and de Burca 2011).

European Union Law as a Body of Law

Status of European Union Law and Main Legal Texts

The law of the European Union can roughly be divided into primary and secondary legislation. Primary law is the law that is found in the treaties, the Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU). The TFEU superseded the Treaty on the European Community as a result of the Treaty of Lisbon (Fairhurst 2012).

Secondary law is EU law that is adopted on the basis of primary law, the so-called legal bases, and can take the shape of regulations, directives, or decisions, as well as the soft-law form of recommendations and opinions (art.288 TFEU). Regulations have general application and are binding in their entirety, while directives are binding as to the result to be achieved (and need to be implemented by national authorities); however, Member States

are free to choose form and methods. Decisions are binding in their entirety on those to whom they are addressed. If there is no legal basis in the treaty, the EU is not allowed to adopt any legislation; this is called the principle of conferral as laid down in art.4 (1) and art.5 (1) TEU (Craig and de Burca 2011). Those areas in which the Union is allowed to legislate are divided in three categories:

Exclusive competence areas are those on which only the Union is allowed to legislate, the customs union, competition rules, monetary policy for euro-area Member States, conservation of marine resources under the common fisheries policy, and the common commercial policy (art.3(1) TFEU). This competence also applies in external context for the negotiation of international treaties, which the Commission is allowed to sign on behalf of the EU for these policy areas (art.3(2) TFEU).

Shared competence areas are those in which both the Union and its Member States are allowed to take legislative action and include the social policy; economic, social, and territorial cohesion; and transport (art.4(2) TFEU includes a full list). However, the level on which decisions are actually taken is governed by two important principles of the EU: subsidiarity and proportionality (Lenaerts and van Nuffel 2011). Both principles were put in the Treaty of Maastricht (TEU) and describe that:

Under the principle of subsidiarity, in areas which do not fall under the exclusive competence of the Union, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level (art.5(3) TEU)

and

under the principle of proportionality, the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties. (art.5(4) TEU)

There is one competence that is the odd one out – economic coordination. While monetary policy is an exclusive competence of the EU and

is conducted by the independent European Central Bank, Member States are to “coordinate their economic policies within the Union” (art.5(1) TFEU). This discrepancy in competence division is arguably one of the reasons for the euro crisis (de Grauwe 2012).

Not only can one categorize European Law on the basis of competences, another divide that is often made is the distinction between Internal Market law and institutional/constitutional law (see below). The notion Internal Market describes a Common Market between the members of the EU. While economic integration was initially the main purpose of the Union, the institutional setting has only evolved alongside the process of integration.

The Institutions

After the Treaty of Lisbon, the European Union is officially governed by seven institutions: the European Parliament, the European Council, the Council, the Commission, the Court of Justice of the European Union, the European Central Bank, and the Court of Auditors (art.13 TEU). These institutions have developed over time, although most of them were part of the Union framework even before they were officially mentioned as institution in the treaty. However, some institutional developments cannot go unmentioned (Hartley 2014).

The European Parliament used to be called the Assembly until 1962 (EP Resolution of 30 March 1962, JO1962) and hardly had any powers at all in decision-making – it was not even recognized as an institution. That changed in the case *European Parliament v. Council* Case 302/87 (1988) ECR 5615 concerning the Parliament’s legal standing in front of the Court, which, before the judgment, was not “privileged,” i.e. directly admissible with any action before the Court, a prerogative reserved for institutions (Hartley 2014). The European Parliament then started to play an ever greater role in the EU, culminating in the first direct elections for the European Parliament in 1979, and together with the national parliaments, it is now seen as the democratic underpinning of the European Union. It is comprised of currently 751 members representing the European people

by degressively proportionate representation, with a slight overrepresentation of small Member States.

More recently, the introduction of the so-called ordinary legislative procedure by the Treaty of Lisbon basically made the European Parliament co-legislator with the Council in most areas of European law (art. 294 TFEU). The Council, as the main legislator, is comprised of ministerial delegates who are authorized to commit their governments (art. 16(2) TEU); the Council does not have a fixed constellation, but the ministerial delegates will change depending on the topic to be discussed – finance ministers will legislate on financial matters, ministers of agriculture will legislate on agriculture, etc. The presidency of the Council rotates on a half-year basis among Member States; however the situation is different when it comes to a, one could call, special configuration of the Council, the European Council. The European Council is comprised of the Member States' heads of state or government and has, since the Treaty of Lisbon, a permanent president who is elected for two and half years. The European Council is the highest political organ of the European Union and provides it with the necessary impetus for its development and defines general political directions and priorities (art. 15 TEU; Craig and de Burca 2011). Especially in the euro crisis, the European Council has played a major role with regard to the measures taken to tackle the crisis, since the situation required a consensus on highest political level.

The Commission is an independent body representing the Union interest and is led by 28 Commissioners, one from each Member State. Commissioners are proposed by the European Council, while the European Parliament has to consent to the appointment of the president of the Commission and his or her Commissioners by the European Council (art. 17(5) TEU). The Commission is the only institution to have legislative initiative; it formulates proposals for new EU policies, mediates between Member States, and oversees the execution of Union policies. It could be regarded as the EU executive (Craig and de Burca 2011).

The European Central Bank (Hartley 2014, p. 31f) only became an official institution once the Treaty of Lisbon entered into force, even though it existed in the same shape before – since the inception of the Economic and Monetary Union (EMU). Its independence in conducting monetary policy for the Member States whose currency is the euro is enshrined in the treaty (art. 283(3) TFEU) and is only limited by its mandate being that of securing price stability (art. 127(1) TFEU).

The Court of Auditors consists of one national per Member State from within a pool of national auditors. Their task is to examine the accounts of revenue and expenditure of the European Union and to provide the Union legislator with reports on the regularity and legality of such (Hartley 2014, p. 31). The Court of Auditors is strictly a monitoring institution and is not to be confused with the Court of Justice of the European Union (CJEU). The CJEU has, in the context of the reforms of the Lisbon Treaty, undergone a major restructuring and is now divided in the General Court (a court of first instance and for the less difficult cases), the European Court of Justice (for seminal cases, it always sits in full court), and the Civil Service Tribunal (for cases involving the staff of the EU). The Court of Justice has played a major role in the process of European integration (Craig and de Burca 2011), which will be discussed below.

In general, however, the division of tasks among the institutions in the European Union is very well thought through and quite balanced. The Commission, together with national authorities represented in the European Council, acts as an executive; the Council and the Parliament now share the legislative competence, while the CJEU is the obvious judiciary. The European Central Bank is probably the only central bank whose independence is enshrined in a document of arguably constitutional value (the treaty). The democratic legitimacy of the European Union is based on the European Parliament and the indirect control that national parliaments have on their governments when they take decisions in the Council, as well as the principle of subsidiarity which is controlled by national parliaments (Lenaerts and van Nuffel 2011).

While the establishment of a democratic Union is one of the proclaimed aims of the EU, one has to be aware of a debate that has mainly started with the Treaty on European Union in Maastricht and that deals with the question of a democratic deficit in the European Union. Two strands of literature claim the existence of such a deficit. First, institutionally speaking, the European Parliament is argued to be unrepresentative of the European people since voter turnout is relatively low and the ordinary legislative procedure is not applicable in important policy areas such as the Common Foreign and Security Policy and the monetary policy. In addition, the involvement of national parliaments is often seen as too little (Follesdal and Hix 2005). Second, a sociopsychological democratic deficit is often attested for the Union. This refers to the lack of a common European people (demos), which in turn leads to a conceptual lack of a European democracy. Opponents on the other hand argue that there is no democratic deficit, because the Union is mainly a technocratic order and is also designed as such (Majone 1998). While this debate is worthwhile exploring, it would go beyond the scope of this entry to discuss it in detail.

European Union Law as a Legal Order

The treaties are essentially treaties that have been entered into under the rules of public international law. However, the European Union legal order has evolved as a legal order standing on its own. In this, the European Court of Justice has played a major role in the early years of integration. Two seminal cases defined the status of European Union law. First, in the case *Van Gend en Loos v Nederlandse Administratie der Belastingen* (1963) Case 26/62, which concerned a reclassification of a chemical in another customs category by the Benelux countries, the Court stated that the EEC Treaty was a legal order on its own, capable of creating legal rights which can be enforced directly by natural or legal persons before the courts of the Community's Member States, if the respective provision was "clear, precise, and unconditional." This principle is now called the principle of direct effect and is probably one of the most important principles of Union law

and has been further developed (Craig and de Burca 2011).

Another seminal case that thrived European integration is the case *Flaminio Costa v ENEL* [1964] ECR 585 (6/64) which regarded an alleged incompatibility of an Italian domestic law with the treaty. Here, the Court stated that "the law stemming from the treaty, an independent source of law, could not (...) be overridden by domestic legal provisions, (...), without being deprived of its character as community law and without the legal basis of the community itself being called into question." It thereby effectively established supremacy of the Union law over national law, which even applies in the case of national constitutions (*Internationale Handelsgesellschaft und Vorratsstelle für Getreide und Futtermittel* Case 11-70, ECR 1970 1125; Lenaerts and van Nuffel 2011).

The Dynamics of the European Union Law

The legal order of the EU has not only changed and developed because of the proactive Court of Justice. Rather, the EU has always reacted to social, economic, and political realities and has changed accordingly. Unfortunately, this has not always been to the benefit of a more coherent EU law. In order to change the treaties and provide the Union with more competences, all Member States have to be in agreement (art. 47 TEU) and ratification in all Member States is required. Increasingly, the Member States find it hard to reach consensus among, currently, 28. Some Member States are keen to drive the integration process forward towards a political and social union, while others are more reluctant. In some policy areas, therefore, some Member States have negotiated an opt out (such as the UK and Denmark regarding the monetary union) or those Member States that wanted to further integrate opted for different (intergovernmental) solutions (Craig and de Burca 2011).

The Schengen Treaty is one example of such further cooperation which started out as an international treaty outside the EU framework. Some Member States committed themselves to abolish border controls among themselves in order to facilitate the free movement of persons. Later,

however, the Schengen Treaty was incorporated into the EU treaty framework and is now part of the so-called *acquis communautaire*, the Union law that needs to be accepted and implemented by all new Member States. Only those Member States that did not ratify the Schengen Treaty before it was part of the Union framework (the UK and Ireland) are not obliged to comply with Schengen (Hartley 2014).

Similar developments have taken place with regard to the tackling of the economic and debt crisis. While EU measures enhancing the economic agreements under the so-called Stability and Growth Pact (SGP) have been adopted and implemented in the form of the six pack (six regulations and one directive) and the two pack (two regulations), agreement on the establishment of a permanent financial stability mechanism could not be reached. In addition, mechanisms for immediate financial assistance for Member States in case of an economic crisis were not foreseen in the treaty. The result was the establishment of somewhat hybrid institutions such as the European Financial Stability Facility, a *societe anonieme* under Luxembourg law backed up by euro-area Member States lending money to other euro-area Member States in financial trouble, the granting of bilateral loans to Greece, and the European Financial Stability Mechanism, all of which are now replaced by the permanent European Stability Mechanism, the ESM (de Grauwe 2012). The ESM as well is founded in a treaty outside the EU framework ratified by 25 Member States (apart from the Czech Republic, the UK, and Croatia), although a treaty change to art.136 TFEU has been made allowing the euro-area Member States to establish among themselves a European Stability Mechanism. Another inter-governmental treaty, the Treaty on Stability, Coordination, and Governance (TSCG), has been ratified by the same countries, in order to improve fiscal discipline and economic coordination among the signatories. Both treaties are designed to include the EU institutions such as the Commission and the Court in their working, and the EU aims to incorporate these treaties in the EU framework midterm. However, it is striking that intergovernmental decision-making in the context

of treaties or the European Council has gained momentum in recent years. While it used to be the aim to bring more and more policy areas under the supranational pillar of decision-making (the ordinary legislative procedure), particularly the economic and monetary union is increasingly governed intergovernmentally. The European Council might be evolving from an institution that gives political impetus to an institution with a more legislative role. While this might arguably sidestep the European Parliament, it is also an example for the dynamics of EU institutional law. The mode of decision-making seems to increasingly depend on the sensitivity of the policy area concerned: Internal Market law on the other hand is still a prime example for supranational decision-making in the EU.

The Co-development of European Community Law and Economic Integration

The Economic Rationale for European Community Law and Integration

The body of European Community law implies a great promise for the European Member States, companies, and citizens. Tearing down barriers to trade and closer institutional integration can reap considerable welfare gains.

Even though the beneficial effects of free trade are known since the works of Adam Smith (1776) and David Ricardo (1817), it was the Cecchini-report from 1988 (Cecchini Report 1988) that estimated an overall cost reduction of 200 billion ECU for the Member States, if all trade barriers (tariff and non-tariff) would fall. However, the report provided no detailed analysis about the growth effects and distributional consequences for and between regions (e.g., center and periphery) (Baldwin 1989).

Because some Member States win and some lose with regard to their absolute welfare position, when trade barriers are removed (while overall welfare increases), it did not come as a surprise that in the aftermath of the Cecchini-report, the potentially losing Member States resisted a simple

abolishment of trade barriers (for a theoretical analysis, see Pierson 1996).

The insight that the distributional consequences of abolishing trade barriers which must be taken into account leads to three basic propositions that drive the economic analysis with regard to European Community law.

1. Economic integration is a process in time in which the Member States go through *stages of integration*, thereby deepening integration.
2. European Community law is mirroring the stages of integration, thereby aiming towards a European *constitution*.
3. At the micro-level the integration process is driven by *interest groups* and stakeholders that impact on the content of European Community law.

Stages of Integration

It is common to differentiate five generic stages of economic integration (Balassa 1961; Baldwin and Wyplosz 2012). Economic integration usually starts at the first or second stage and progresses then over time towards stages of higher integration.

1. The first stage is mainly associated with a bilateral or multilateral free trade agreement (FTA) between countries. While the abandoning of tariffs between the signatories of the FTA reaps some welfare gains, running an FTA is not trivial and not easily administered. That is because of the FTA countries' freedom to determine the tariff with third countries on their own. For importers and FTA countries alike, this creates opportunities for arbitrage (Tarr 2009). As a consequence rules of origin have to be introduced and maintained. One may conceive these rules of origin as a very nascent step into the direction of a common understanding over the rules of the game between FTA countries. One may regard this as a first step towards a premature constitution.
2. The second stage is the customs union (CU). The signatories of a CU abandon tariffs against each other, but they also agree on a common tariff against third countries. This overcomes

the problem of maintaining and controlling complex rules of origins as it is necessary under an FTA. From an economic perspective, the Treaty of Rome (1957) constituted a CU.

3. The third stage is the establishment of a common market (CM). The Treaty of Rome implied already the antecedents of a CM, although it took time till the 1980s to establish and enforce the features of a CM. Central to a CM is that non-tariff barriers to trade (national regulations) which impede trade become removed and that the free movement of people/companies, services, and capital becomes feasible. Two features of a CM need special attention: First, non-tariff barriers for goods, people/companies, services, and capital (four freedoms) can either be overcome by harmonizing laws and regulations or by mutual recognition of laws and regulations. Both ways play out in the European economic integration process. Second, to safeguard the four freedoms, a CM has to establish a common competition policy, preventing the Member States to give unfair advantages to their industries (esp. state aid control). Both features have in common that they need policy coordination between Member States, implying that there is a basic agreement about the vertical delineation of the competencies between the central level (EU level) and the signatories (Member States). In parallel power has to be vested to institutions and administrations at the central level, in order to execute the common policies. This process of institutionalization becomes apparent, for example, in the increasing role that the Court of Justice of the European Union plays.
4. The fourth stage is the economic union (EUN), which is a CM where a number of key policy areas are harmonized (esp. coordination of monetary and fiscal policies, labor market, regional development, infrastructure, and industrial policy). While at the first three stages of economic integration the removal of obstacles to free trade are center stage (negative integration), at the fourth stage the particular institutional design of the economically integrated area becomes important. This process of

institutionalization goes hand in hand with the strengthening of administrative bodies managing for the particular design of integration (positive integration). However, the implementation and enforcement of policies remain with the associated countries. The establishment of the European Single Market (Maastricht Treaty) is a good example of a EUN.

5. The fifth stage is total economic integration (TEI). At this stage, a vast number of key policies become harmonized, *inter alia* those as social policy and monetary and fiscal policy. A TEI exploits all possible economic advantages from free trade and factor mobility. The momentum of positive integration becomes amplified by shifting even more power to the central administration. The policies from the central level become binding for the associated countries, which role is only to execute the prescribed policies from the central level. The EU has yet not reached the stage of TEI, but the members of the euro area are close to it.

Towards a European Constitution

European Community law has yet not been consolidated in a European constitution. The “Treaty establishing a Constitution for Europe” from 2004 was not ratified by all Member States; as a consequence in 2005, the project of a European constitution was given up. Instead in 2009 the Treaty of Lisbon implemented parts of the constitutional project (e.g., qualified majority voting) through the established European treaties.

Insofar, in a strict legal sense, European Community law has yet not reached the status of a constitution. However, European Community law has undoubtedly evolved over time towards a tighter set of rules and regulations (institutional integration). From a law and economics perspective, the crucial question is whether the institutional integration of the EU corresponds indeed with more transborder transactions and hence more gains from trade. In other words, does the process of European constitutionalization lead to welfare gains?

This question can only be answered empirically and it bears the question of how to measure institutional integration. However, Mongelli

et al. (2005) can show that a higher degree of institutional integration goes hand in hand with more economic integration (trade). Thereby it seems that causation runs both ways: More institutional integration triggers more economic integration, but more economic integration also leads to more institutional integration. From that observation follows that there is a coevolution between European Community law and economic integration. This implies that the European constitutionalization process is driven by policy choices for the background of a deepening economic integration, while at the same time the deepening of economic integration depends on the policy choices that were made.

The Role of Interest Groups

European Community law is the product of actors on the national and the EU level which try to shape the governing rules and the law-making process into a for them favorable direction. Usually those actors are labeled as interest groups. Even though the term “interest groups” must not have a negative connotation but can refer to quite a lot of meanings in the process of European integration (Eising 2008), from an economics perspective it refers mainly to well-organized groups that try to get an economic advantage that is not the result of their economic performance but of their successful attempts to get favors via the political process (rent seeking) (Mueller 2003; with special reference to federations, see Weingast 1995). For example, the persistent high levels of subsidies that go into the common agricultural policy (CAP) can only be explained by the influence that the interest group of farmers has on European politics (Baldwin and Wyplosz 2012).

The role of interest groups in the process of European integration in general and the impact of interest groups on European Community law in particular can be studied from a great number of perspectives (for an overview, see Eising 2008). From a law and economics perspective, it is important to understand the making of European Community law as the result of a coevolutionary process of diverse interest groups, which influence laws and regulations, and the body of European law, which creates the forum in which

interest groups and stakeholders can play out. Thereby the European Community law becomes over time more differentiated and adapted to the challenges it faces in an ever more economically integrated area, but at the same time the increasing complexity of European Community law creates gaps and loopholes that can be targeted by interest groups. This implies on the one hand that interest groups have to evolve to organizations that are able to target those loopholes and gaps. On the other hand, it implies that Community law reacts to the activities of interest groups with a further differentiation of law and regulations.

While the relevance of interest groups for the development of European institutions and Community law cannot be doubted, the operation of interest groups in the EU is yet not fully understood. For example, it is yet not clear whether there is a Europeanization of interest groups (making the level of Member States less important) or whether successful interest groups on the European level need a strong anchoring in the Member States. Moreover, one may ask whether the specific politico-legal system of a Member State predetermines the effectiveness of interest groups on the European level (Eising 2008).

References

- Balassa B (1961) The theory of economic integration. Irwin, Homewood
- Baldwin R (1989) The growth effects of 1992. *Econ Policy* 4:247–282
- Baldwin R, Wyplosz C (2012) The economics of European integration, 4th edn. McGraw-Hill, London
- Cecchini-report (1988) The European challenge. Gower, Aldershot
- Craig P, de Burca G (2011) EU law- text, cases and materials, 5th edn, Oxford University Press, Oxford
- De Grauwe P (2012) The governance of a fragile eurozone. *Aust Econ Rev* 45(3):255–268
- Eising R (2008) Interest groups in EU policy making. *Living Rev Eur Gov* 3:4–32
- European parliament resolution of 30 March 1962, JO1962
- Fairhurst J (2012) Law of the European union law, 8th edn. Pearson, Harlow
- Follesdal A, Hix S (2005) Why there is a democratic deficit in the EU. European Governance papers (EUROGOV) C-05-02
- Hartley T (2014) The foundations of European union law, 8th edn, Oxford University Press
- Lenaerts K, Van Nuffel K (2011) European union law, 3rd edn. Sweet & Maxwell, London
- Majone G (1998) Europe's 'democratic deficit': the question of standards. *Eur Law J* 4(1):5–28
- Mongelli FP, Dorrucchi E, Agur I (2005) What does European institutional integration tell us about trade integration? ECB occasional paper series 40
- Mueller DC (2003) Public choice III. Cambridge University Press, Cambridge
- Nugent N (2010) The government and politics of the European union, 7th edn. Palgrave MacMillan, Basingstoke
- Pierson P (1996) The path to European integration: a historical institutionalist analysis. *Comp Polit Stud* 29:123–163
- Ricardo D (1817) On the principles of political economy and taxation. John Murray, London
- Smith A (1776) An inquiry into the nature and causes of the wealth of Nations. Methuen, London
- Tarr RD (2009) Customs union. In: Reinert KA, Rajan RS (eds) The Princeton encyclopedia of the world economy, vol 1. Princeton University Press, Princeton, pp 256–259
- Weingast B (1995) The economic role of political institutions: market-preserving federalism and economic development. *J Law Econ Organ* 20:1–31

Further Reading

- Chalmers D, Davies G, Monti G (2010) European union law, 2nd edn. Cambridge University Press, Cambridge
- Moravcsik A (2002) In defence of the democratic deficit: reassessing legitimacy in the European Union. *J Common Mark Stud* 40(4):603–624

European Court of Human Rights

Federica Gerbaldo

IEL- Institutions, Economics and Law, Collegio Carlo Alberto, University of Torino, Turin, Italy

Abstract

The European Court of Human Rights is a Supra-national Court established in 1959 with the European Convention on Human Rights. It stands as a monitoring mechanism to ensure the observance of the commitments to ensure fundamental human rights undertaken signing the Convention.

This entry focuses on the functioning mechanism of the Court either from a prescriptive and an empirical point of view. The first aspect

concerns organization and procedural issues, with special regard to the sanctioning mechanism, whereas the second aspect involves issues of effectiveness of the Court.

Definition

The European Court of Human Rights (ECtHR) is a supranational court founded by the European Convention of Human Rights (ECHR) in 1959 and based in Strasbourg (France). It carries out a supervisory function monitoring that the 47 member states of the Council of Europe (COE) that have ratified the Convention comply with its substantive provisions. The court fulfills its task scrutinizing claims alleging that the defendant state breached its commitment violating one or more ECHR provisions.

History

Founded in 1949, the Council of Europe (COE) became the very first political organization in Europe, although its approach was different from the European institutions. In the era of reconstruction after World War II, the original member states chose the path of cooperation and signed the founding treaty of COE as a reaction to the atrocities of the war and the growth of East-West tension. States parties made efforts in strengthening cooperation, signing international treaties, publishing peer reviews, exercising pressure, and promoting training and good practice; reinforcing citizenship and democratic governance was the ultimate goal.

The ambitious project of *achieving a greater unity between its members for the purpose of safeguarding and realizing economic and social progress* (article 1 Statute of council of Europe, 1949. Add to documents references Council of Europe, Statute of the council of Europe, 05 May 1949) aimed to guarantee lasting peace and prosperity on a continental scale. In this sense, the most successful achievement is the ECHR signed in 1950 and entered into force in 1953 and the creation of a monitoring mechanism with the

establishment of ECtHR. As a response to the inertia of the United Nations and its inconclusive attempts to transmit the principles proclaimed in the 1948 Universal Declaration of Human Rights into an internationally binding bill of rights, some of the Western European countries egged on the COE to proceed on its own. Thanks to the obduracy of those countries that share *a common heritage of political tradition, ideals, freedom and the rule of law* (European Convention of Human Rights, preamble, 4th November 1950, 213 UNTS, 221), the COE took the first steps for the collective enforcement of some of the human rights stated in the Universal Declaration.

Even though ECHR scope is more specific, the Convention arranges one of the strongest regional mechanisms for protection of fundamental individual human rights and stands as a model for other regional systems (Buergetal 2006). Signing the Convention, high contracting parties undertake to grant to every individual within their jurisdiction civil liberties typical of an effective political democracy. Among others, the Convention grants life, fair hearings, private and family life, freedom of expression, freedom of thought, conscience and religion, and property rights and prohibits torture or inhuman treatment, slavery and forced labor, arbitrary and unlawful detention, and discrimination. Besides the declaration of a list of human rights that contracting states commit to recognize and grant (see Harris et al. 2014 for an analysis of the rights granted), the Convention made protection effective establishing a monitoring mechanism to *ensure the observance of the engagement undertaking by the high contracting parties* and providing remedies for victims of violations.

After the entry into force, member states implemented the Convention with Protocols oriented to adjust structure and functioning mechanism to the constant increasing workload of the court (Egli 2007, p. 7). According to the original design, the court envisaged a double-step mechanism: a commission performed monitoring tasks analyzing claims' conformity to admissibility criteria and attempting friendly solutions, whereas a court performed adjudicatory tasks. The accession to ECHR of new member states after the fall of the Berlin Wall dramatically increased the

number of claims: Protocol 11, entered into force on 1 November 1998, was imagined to deal with the huge number of pending applications. The articulated mechanism moved from a mix of monitoring and adjudicatory tasks to an entire judiciary system, based on a single court working on a permanent basis: the elimination of the commission screening function simplified and made the procedure more concise. Furthermore, it allowed individual applicants to present claims directly to the court.

Despite its purpose, Protocol 11 failed to deal with the ECtHR workload, and a number of applications incessantly increased during years. In 1999, right after the introduction of Protocol 11, the number of pending applications was 12,600; 10 years later, pending applications amounted to 119,300 (ECtHR statistics 2013, p. 7). Member states issued Protocol 14, which entered into force on 1 June 2010 and provided further reforms aiming to grant long-term efficiency and strengthening applications' process. The protocol reduced the decision body formation for the simplest claims, conditioned the admissibility of claims to further admissibility criteria requiring "a significant disadvantage" for the applicants, and prolonged the judges' term of office from 6 years with option of reelection to nonrenewable 9 years term providing for expiration of the office when they reach 70 years of age.

Currently, members have adopted Protocol 15 introducing references to the principle of subsidiarity and the doctrine of the margin of appreciation and reducing from 6 to 4 months after the date of the final domestic decision the time limit within which the claims should be filed and Protocol 16 allowing the court to give advisory opinions on questions of principle related to interpretation and application of rights, upon request of the highest domestic courts. Both protocols will enter into force only once contracting states will sign and ratify them.

Organization

Issuing durable legal rules per se does not make a regional human rights system effective, if there is

no assignment to third parties of the task of interpreting and applying rules. ECtHR carries out exactly this role and makes the international regional system established by ECHR capable of enforcement. The effects of the incorporation of ECHR together with the functioning of the court as a constitutional court and the effects of ECtHR judgements allow denial of those theories that support a qualitative difference between international regimes and states; thus doubting some form of constitutionalism beyond states is feasible (Rosenfeld 2010). Recent theories argue that the ECHR framework evolves into a transnational constitutional regime (Stone Sweet 2012).

The ECtHR has its headquarters in Strasbourg and is composed by elected judges, chosen either among judges eligible for high judicial office or jurists of esteemed competencies. The number of judges is the same as the number of the high contracting parties (47 at present). The Parliamentary Assembly elects judges with majority of votes, among a shortlist of three candidates presented by each contracting party.

Judges must have high morality: although representing one state, they hear the claims in their individual capacity. Thus, they must avoid any activity which endangers their impartiality and independence.

Depending on the complexity of the case, the ECtHR scrutinizes the claim sitting in one of its four judicial formations. When inadmissibility of the claim clearly appears without any further investigation, the single judge has the power to strike the complaint out of the list of the cases or declares its inadmissibility with final decision. When this is not the case, the single judge forwards the claim either to a Committee or to a Chamber. The Committee unanimously decides either to declare inadmissible the claim or strike it out when no further investigation is required or to declare it admissible and simultaneously issue the judgement on the merit when the question is the object of well-established jurisprudence. Alternatively, the Chamber rules on the merit of the claim, together with a decision on admissibility or separately. In the exceptional situation in which the case concerns a question seriously affecting ECHR interpretation or which might

lead to a decision contrasting with a previous judgements, the Chamber relinquishes jurisdiction in favor of the Grand Chamber of 17 judges. Moreover, one or both parties might require referral to the Grand Chamber within a period of 3 months from Chamber judgement's delivery.

Procedural Issue

ECtHR has jurisdiction to hear claims filed by individuals or state and alleging that one of the high contracting parties violated the ECHR. Jurisdiction includes all the matters related to interpretation and application of the ECHR; however, the court shall not take case on its own motion. Since the introductory phase, the procedure stands out for its accessibility and effectiveness. Both states and individuals, either person, group of people, or non-governmental organizations, have the right to present claims, thus granting widest access to justice.

A complaint undergoes two steps, with some faint adjustments concerning the analysis depending on the judiciary formation scrutinizing it. The first is the admissibility stage during which judges verify the observance of some formal requirements.

Claims must concern one of the rights covered by ECHR, not to be manifestly ill-founded and be direct against one or more states parties of the Convention. Following a general rule of International Customary Law, applicants shall exhaust all the available domestic remedies, up to the highest level of jurisdiction, and resort the ECtHR within 6 months following the date of the last judicial decision. Last, the applicants shall have suffered a significant disadvantage: the violation, although real from a mere legal point, shall overtake a minimum threshold to deserve international judges' attention. These requirements reflect different purposes: the former grants national authorities the opportunity to prevent, or at least to fix, violations, whereas the latter facilitates the activity of the court making it easier when dealing with unmeritorious claims and allowing devotion of more time to serious claims. Both rest on the same assumption: the subsidiarity of ECtHR

with respect to national mechanism of safeguarding human rights. Thus, on one side, national courts have at first instance the chance to deal with questions regarding the compatibility of domestic law with the Convention. On the other side, the court stands as a supranational court of final instance deserving the analysis of substantial claims and denies the role of an alternative channel to which require further monetary damages.

Past the admissibility screening, judges scrutinize the merit of the claim and undertake further investigation, if needed, inviting parties to submit further evidence and written observations or fixing public hearings with the representatives of the parties. At the end of the investigation, judges deliberate on the question and issue a judgement, providing reasons for the decision, which becomes final if parties declare to renounce or do not request referral to the Grand Chamber within 3 months.

In principle, the final judgement is declaratory in nature (*Marckx v. Belgium*, appl. number 6833/74 1979 par. 58). Only in 2011, the court introduced an original mechanism aiming at optimizing time and resources allocation when dealing with cascade of repetitive claims, which does not necessarily end with a declaration that a violation has been committed. When applications underline the existence of a structural or systemic problem, the court might issue a pilot judgement after processing them as a matter of priority. The pilot judgement contains both the identification of the nature of the structural problem or the dysfunction and the suitable remedies the defendant state is required to adopt.

Final judgements are binding upon the parties of the cases. The declaratory character implies that the court renounces to impose on the breaching state any obligation to adopt specific measures necessary to ensure compliance, a part for some exceptional cases. The execution of the final judgement consists into the performance of two groups of obligations, depending on the measures required. The first one demands adoption of general measures and includes the obligation to execute the violated provisions and to prevent the occurrence of further violations. The second one consists into the obligation to put an end to the

violation and fix the negative effect, the execution of which required individual measures.

General measures aim to prevent future violations similar to the ones found by the court and most of the time require changes in legislation. However, when national authorities recognize direct effect to judgements, publication and dissemination of the decision is sometimes enough to induce national effective remedies.

Individual measures aim to put an end to the violation and to *restore as far as possible the situation existing before the breach* (Brumarescu v. Romania 1999, par. 19). Thus, following customary rules of international law, *restitutio in integrum*, aiming to take the injured back as far as possible in the same situation as the one he enjoyed before the violation, represents the preferred relief. When the restoration of the injured victim is otherwise impossible or only partial, the court might award monetary compensation.

The Committee of Ministers is the organ charged of supervising the execution of the judgements. The Committee organizes two meetings a year and gathers one representative for each contracting party (in principle the Minister of Foreign Affairs), each of whom is entitled to one vote. The monitoring function of the organ includes the supervision of the execution of friendly settlement and judgements (White and Ovey 2010, p. 53). The Committee invites the respondent state to explain the measure taken to avoid the consequence of violation and provide evidence of compensation payment: the Committee supervises regularly each case until it is satisfied with the adoption of general measures necessary to ensure compliance.

Just Satisfaction

Besides the general freedom in the execution of the judgements, the ECHR provides the court with the power to award compensation for injuries suffered, as a residual remedy for those damages that cannot otherwise be repaired. The core of the system lies in article 41 that provides for just

satisfaction. The provision envisages that if a violation is found and *if the internal law of the contracting party concerned allows only partial reparation to be made, the court shall, if necessary, afford just satisfaction to the injured party*. The court might exercise the power to afford just satisfaction to victims in the same judgement if the question is ready for decision and claimants specifically require monetary compensation. Just satisfaction is a compensation for an actual harm: finding of evidence of violation does not automatically grant a positive award, being a matter of court discretion. The letter of the provision is clear when specifying that the award is not a direct and automatic consequence of the finding of violation. Judges determine to award just satisfaction upon a specific applicants' request supported by appropriate documentary evidence. The approval of the request is conditional on the inadequacy of national remedies to provide full reparation to victims. In addition to subsidiarity, the award shall be necessary and just.

Economic analysis of law acknowledges damages as an instrument having twofold function: restoring the injured for the harm suffered (compensation) and providing injurers with behavioral incentives fostering internalization of the externalities (deterrence). Due to their economic function, in general, damages shall provide full compensation for the victims' losses because only in this case the injurer internalizes the negative externalities produced (Posner 2003, p. 192). Nevertheless, compensation of damages within the international public field assumes a peculiar meaning considering that the court exercises only a weak form of review (Bernhardt 1994, p. 297), but aims to provide "constitutional justice," rather than "individual justice." On one side, when finding a violation of the Convention, judges' declaration of incompatibility with ECHR does not replace the questioned law with one of its own making, nor directly affect the validity of that law in the national legal system. On the other side, the actual function of the court is to ensure that administrative and judicial processes in member states effectively conform to pan-European convention standards, rather than seeking to provide every deserving applicant

with a remedy for the Convention violation (Greer 2003, p. 405).

Traditional economic theories argue that damages do not have to fully compensate harm, as long as they make taking due care the most attractive strategy (Visser 2009, p. 155). However, having regard to the international human rights protection framework, under-compensation of the harm caused might not be sufficient in view of obtaining the ultimate goal of the court, namely, the adoption of a standard of protection appropriate to the ECHR. On the other side, the rules specifically clarify that the court does not aim to punish the breaching state for its responsibility in violating the Convention and invite to charge as inadequate any claims for *punitive, aggravated or exemplary* damages. Just satisfaction turns out to be the result of a balance between the interest of injured applicants to receive some sort of satisfaction and the public interest of defendant states, with an eye to the local economic circumstances of the state.

The award, if any, is exclusively in the form of a sum of money to be paid by the respondent state within a time limit that usually the court sets in 3 months. Judges might either award a global sum, either breakdown just satisfaction into three components that correspond to material losses, moral losses, and costs and expenses. Practical indications on the computation of material damages refer to the principles of International Customary Law codified by the United Nations Law Commission (chapter IV.E.1, article 28 and following). The basic principle rests upon the assumption that the injured applicant *should be placed, as far as possible, in the position in which he or she would have been had the violation found not taken place*. Material damages awards concern economic losses and would in principle reflect *the full calculated amount of damages* or a reasonable estimation based on observable facts.

Moral damages include a mixture of highly intertwined elements, many of which involved social and psychological aspects: it is a matter of civil law jurisdiction, rather than an instrument of public international law (Parish et al. 2011, p. 225). This concept recurs in context of tort

that causes intangible, or otherwise difficult to quantify, injuries to a person and/or his/her rights, including pain and suffering, anguish and distress, and loss of opportunities. Assessment of moral damages is difficult because moral losses are not directly visible (Shavell 2004, p. 242; Aarlen 2013, p. 439). The court itself admits that *it is in the nature of nonpecuniary damages that it does not lend itself to precise calculation* (Practice Directions, section III, par. 13). As a matter of fact, assessment of damages grounds on equitable arguments and standards emerged by case law.

Empirical Studies Focusing on ECtHR

ECtHR counts the biggest caseload among international courts, with more than 17.000 judgments delivered since its creation, and makes most of the information about the cases publicly available. This makes the court as a suitable object of attention when investigating the more relevant issues related to effectiveness of international adjudication.

Voeten (2008) investigates international judicial behavior, focuses on impartiality of ECtHR judges, and considers the court as a committee of individual judges, rather than a single actor. Focusing on the minority opinions expressed by judges, he evaluates what determinants other than law, if any, systematically affect judges. Findings show that ECtHR judges' behavior is more similar to that of national review courts, rather than political institutions: there is no evidence that judges' domestic legal culture is a matter of concern, and the overall effects of career motivation are moderate (despite some evidence that judges show national bias in evaluating politically sensitive case). The study is overall positive toward the possibility of impartial review and suggests ECtHR judges are politically motivated actors having political preferences on to the best way to apply abstract human rights, but they do not use their power to settle geopolitical scores.

Helfer and Voeten (2014) take the ECtHR rulings on a specific issue, namely, lesbian, gay,

bisexual, and transgender rights, to investigate the effectiveness of international tribunals as agent of policy changes. The study aims to investigate whether the court affects the behavior of actors other than parties to a dispute or, contrary, the court simply reflects evolving social trends and adapts its jurisprudence to the evolution of national policies. The study contributes with an empirical assessment to the path of literature trying to provide analytical explanation of the effectiveness of international adjudication activities (Helfer 2014). In particular, it focuses on what has been classified as the *erga omnes* effectiveness and *norm development effectiveness*. Findings show that international law created by ECtHR jurisprudence systematically affects domestic judicial review and legal changes. ECtHR judgements increase the likelihood that all COE members, even those that are not found to violate the Convention, adopt friendly LGBT rights reforms, and the effect is much more evident where support for LGBT rights is relatively low. Based on statistical findings, the authors argue that the ECtHR court contributes to legitimize and justify political changes especially with respect to lazy countries. The court engages in majoritarian activism, rather than aggressive policy, and reverses previous decision only when at least a majority of COE members have already done so.

References

- Aarlen J (2013) Research handbook on the economics of Torts. Edward Elgar, Cheltenham
- Bernhardt R (1994) Human rights and judicial review: the European Court of Human Rights. In: Beatty DM (ed) Human rights and judicial review: a comparative perspective. Martinus Nijhoff Publishers, Dordrecht, pp 297–319
- Buergenthal T (2006) The evolving international human rights system. *Am J Int Law* 100(4):783–807
- Egli P (2007) Protocol no. 14 to the European convention on human rights and fundamental freedoms: towards a more effective control mechanism? *J Transl Law Policy* 17(1):1–34
- Greer S (2003) Constitutionalizing adjudication under the European Convention on Human Rights. *Oxf J Leg Stud* 23(3):405–433
- Harris D, O’Boyle M, Bates E, Buckley C (2014) Law of the European convention on human rights, 3rd edn. Oxford University Press, Oxford
- Helfer LR (2014) The effectiveness of international adjudicators. In: Romano C, Alter K, Shany Y (eds) The Oxford handbook of international adjudication. Oxford University Press, Oxford, pp 464–482
- Helfer LR, Voeten E (2014) International courts as agents of legal change: evidence from LGBT rights in Europe. *Int Organ* 68(01):77–110
- Parish MT, Nelson AK, Rosenberg CB (2011) Awarding moral damages to respondent states in investment arbitration. *Berkeley J Int Law* 29(1):225–245
- Posner RA (2003) Economic analysis of law, 6th edn. Aspen Publisher, New York
- Rosenfeld M (2010) The identity of the constitutional subject: selfhood, citizenship, culture and community. Routledge, London, pp 61–65
- Shavell S (2004) Foundations of economic analysis of law. Harvard University Press, Harvard, p 242
- Stone Sweet A (2012) The European convention on human rights and national constitutional reordering. *Cardozo Law Rev* 33(12):1859–1868
- Visser LT (2009) Tort damages. In: Faure MG (ed) Tort law and economics, encyclopedia of law and economics, 2nd edn. Edward Elgar, Cheltenham, pp 153–200
- Voeten E (2008) The impartiality of international judges: evidence from the European Court of Human Rights. *Am Polit Sci Rev* 102(4):417–433
- White RCA, Ovey C (2010) White and Ovey. The European convention on human rights, 5th edn. Oxford University Press, Oxford

Documents

- Brumarescu v. Romania, appl. number 28342/95 (28 Oct 1999) 1999-VII; 35 EHRR 887. Available at <http://hudoc.echr.coe.int/sites/eng/pages/search.aspx?i=001-58337>
- Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 Nov 1950, ETS 5. Available at http://www.echr.coe.int/Documents/Convention_ENG.pdf
- Council of Europe, Rules of the court, 14 Apr 2014. Available at http://www.echr.coe.int/Documents/Rules_Court_ENG.pdf
- European court of Human Rights, Analysis of statistics 2013, Jan 2014. Available at http://www.echr.coe.int/Documents/Stats_analysis_2012_ENG.pdf
- Marckx v. Belgium, appl. number 6833/74 (13 June 1979) series A n. 31. Available at <http://hudoc.echr.coe.int/sites/eng/pages/search.aspx?i=001-57534>
- Rules of the court, practice directions, just satisfaction claims. Available at http://www.echr.coe.int/Documents/PD_satisfaction_claims_ENG.pdf
- United Nations Law Commission, Draft articles on responsibility of states for internationally wrongful acts, Nov 2001, Supplement Nr. 10 (A/56/10). Available at http://legal.un.org/ilc/texts/instruments/english/draft%20articles/9_6_2001.pdf

Further Reading

- Bates E (2010) *The evolution of the European convention on human rights. From its inception to the creation of a permanent court of human rights*. Oxford University Press, Oxford
- Brems E, Janneke G (2014) *Shaping rights in the ECHR: the role of the European court of human rights in determining the scope of human rights*. Cambridge University Press, Cambridge
- Fenyves A, Karner E, Koziol H, Steiner E (eds) (2011) *Tort law in the jurisprudence of the European Court of Human Rights*. Walter DE Gruyter, Berlin
- Greer S (2006) *The European convention on human rights achievements, problems, and prospects*. Cambridge University Press, Cambridge
- Harmsen R (2007) *The European Court of Human Rights as a "Constitutional Court": definitional debates and the dynamics of reform*. In: Morison J, McEvoy K, Anthony G (eds) *Judges, transition and human rights cultures*. Oxford University Press, Oxford, pp 33–53
- Keller H, Sweet AS (2008) *A Europe of rights: the impact of ECHR on national legal system*. Oxford University Press, Oxford
- Senden H (2011) *Interpretation of fundamental rights in a multilevel legal system: an analysis of the European Court of Human Rights and the Court of Justice of the European Union*. Intersentia, Cambridge
- Xenos D (2012) *The positive obligations of the State under the European Convention of Human Rights*. Routledge, London

European Identity

► European Nationality

European Integration

Aurelien Portuese

University of Westminster, London, UK

Abstract

European integration is a process. It is a process of institutionalization by which the old continent is continuously transformed either through incremental changes or through across-the-board reforms. European integration designates the changing current institutional framework of the European Union (hereafter EU). The European continent

experiences other institutional integration than the EU's – mostly the European Convention on Human Rights – but the focus shall be put on the EU as this regional organization is the most significant and integrated one.

The history of the European integration has paved the way for an entire new legal order with a strong economic rationale to come to the fore (I). The emphasis of institution-building and market-building has nevertheless not been without difficulties in terms of construing a common political structure, democratically legitimate for the European national democracies (II).

Definition

European integration is an expression describing a process, by opposition to a state of affairs that would be mainly unchanged. Hence, the European integration defines an institutional process taking place within a regional organization by opposition to an institutional state of affairs created suddenly through different forms of state such as a federation, a regional state, a nation-state, etc.

The EU is not a state; it is an organization so much of its own that it is commonly described as being a *sui generis* organization. The EU should not be looked at or investigated as photography but rather as a movie in the making – hence the expression “European *integration*.”

The history of the European integration has paved the way for an entire new legal order with a strong economic rationale to come to the fore (I). The emphasis of institution-building and market-building has nevertheless not been without difficulties in terms of construing a common political structure, democratically legitimate for the European national democracies (II).

The Law and Economics of European Integration

The history of the European integration portrays a strong economic underpinning of the common

destiny the European states have decided to embrace (a). Out of the economic system set up in the aftermath of the Second World War, an ongoing process of institutionalization has continuously transformed the European integration to become nowadays the most institutionalized region of the world. This process has spawned a new legal order – the European Union (EU) law – which is imbued with economic considerations (b). The European institutional framework and practice that result from this evolution is, from a law and economics perspective, driven by an economic efficiency rationale (c).

Economics of European Integration: Peace and Prosperity Through Reciprocal Stakes

On a continent repeatedly ravaged by wars and erstwhile enmities, and on a continent continuously disheartened by peace treaties violated, new peace-making solutions have appeared to become imperative during the postwar period. Diplomatic relations had not been so far satisfactory: only economic interdependence would make peace durable.

Because peace leads to prosperity, the argument of this Kantian perspective is circular: an economic integration paves the way for an ever-greater economic interdependence among European states allowing for peace and prosperity on the continent (Kant 1992).

The interlinkage between commercial knots and peace-building is well known since Montesquieu wrote: “The natural effect of commerce is to bring peace. Two nations that negotiate between themselves become reciprocally dependent, if one has an interest in buying and the other in selling. And all unions are based on mutual needs” (Montesquieu 1758, p. XX).

Therefore, the architects of the European integration were driven by fostering the self-interest of each state to make peace: the challenge was not so much about having a possible peace treaty to be signed but more about having war made impossible. Indeed, the economic interdependence created by free trade would render war impossible among rational and self-interested states.

The birth of European integration has consequently been triggered not by setting up

a “superstate,” a federation, or any other grand projects against which nation-states were reluctant to build. It has been driven by rational choice perspective and by the progressive deconstruction of the protectionist powers of the state.

Without coincidence, the first sectors of the economy to be integrated into a common pool were the coal and steel of European nations since coal and steel were the raw material for wars. The Schuman Declaration of 1950 in which Robert Schuman, French Minister of Foreign Affairs, called for a “step-by-step” process of integration leading to a European federation. This declaration heralded the creation in 1951 of the European Coal and Steel Community (ECSC), with a High Authority managing the production and the distribution of coal and steel in Europe. Consequently, with the ECSC, war has been rendered materially impossible (Pelkmans 2006).

In light of the success of the ECSC, the founding members – France, Germany, Italy, Belgium, Luxembourg, and the Netherlands – generalized this success to all sectors of the economy by creating a free trade area through the abolition of customs duties and the proclamation of the free movement of workers, capital, goods, and services. This was the creation of the European Economic Community (EEC) in Rome in 1957 (Duchene 1994).

The common market laid down by the EEC was constituted of a customs union made of free trade area (abolition of tariffs within Member States) and a common external tariff with respect to non-Member States. Also, the four economic freedoms (free movement of goods, services, workers, and capital) were proclaimed. Finally, few common policies were envisaged, such as the Common Agricultural Policy.

There are four stages of economic integration: a free trade area, a customs union, a common market, and finally an economic and monetary union. Before embracing a monetary union with the creation of the Euro in 1992 with the Maastricht Treaty, the EU has therefore jumped from its inception in 1957 into a very integrated regional union – a common market with a common external tariff and with common policies.

A 12-year transition period was granted in order for Member States to abolish their custom tariffs. In 1969, the common market was accomplished. The removal of barriers to trade and the large-scale competition allowed for over the entire continent would stimulate modernization of the economies by the lowering transaction costs, the enjoyment of economies of scale, and the enhancement economic efficiency. The belief in the net positive effects of a customs union was widely shared among economists.

The most influential theory at that time was Jacob Viner's theory of trade creation and trade diversion. He argued that economic integration leads to two simultaneous welfare effects: trade creation among Member States and trade diversion among Member States. Whether trade creation (welfare-enhancing effects) is greater than trade diversion (welfare-decreasing effects) depends on the structure of the economies to be integrated. The relative similarity of the European economies in 1957 led economists to vouch for economic integration through a common market given its plausible net positive effects in terms of trade creation (Pelkmans 2006).

The common market created by the EEC was the means to achieve an everlasting peace and prosperity on the continent. The common market has been a success in its realization. Indeed, from 1958 to 1972, while trade between the six founding Member States and the rest of the world had tripled, intra-community trade had been multiplied by 9 (Pelkmans 2006).

If the European integration had strong economic rationale due to the creation of a common market to the six founding Member States, the European common market could not last without a sufficient degree of institutionalization and a European law that encapsulate this economic rationale.

Law of European Integration: Institutions as Engines of Economic Integration

The effective working of the European integration requires a delegation of authority, even sovereignty, to supranational institutions. This requirement is derived both from the high reluctance of Member States to rely exclusively on other

Member States for the implementation of the rules creating the common market and from the economic rationale of such institutions since decision-making costs are reduced, thanks to the specialization of supranational institutions.

The European Institutions

The institutions of the EEC are representatives of different interests. While intergovernmental institutions within which Member States play a key role represent national interests, supranational institutions within which European institutional actors are predominant represent the "community" interest.

The intergovernmental institutions of the EU are the Council of the European Union and the European Council. The supranational institutions of the EU are the European Parliament and the European Court of Justice.

The European Commission is the similar to a cabinet government, composed of 28 members of the Commission ("commissioners") – one per Member State – who execute the European Treaties. It represents the general interest of the EU. The President of the European Commission is proposed by the European Council and elected by the European Parliament.

The European Parliament is the directly elected legislative assembly of the EU since 1989 where European citizens are represented. Composed of 766 members, the European Parliament has incrementally gained legislative powers to achieve equal footing with the Council of the European Union under the ordinary legislative procedure generalized since the Maastricht Treaty of 1992. Under the ordinary legislative procedure provided at Article 294 of the Treaty on the Functioning of the European Union ("Lisbon Treaty" of 2009), the European Commission has the monopoly of legislative proposal. The European Parliament adopts a first position; if this position is approved by the Council, the piece of legislation is adopted. If the Council disapproves the Parliament's position, it adopts its own position and then communicates to the Parliament who either approves or disapproves the Council's position. In the latter case, a so-called conciliation committee is convened, composed of an equal number of Council's

members and of Parliament's members. If the conciliation committee adopts a joint text later approved by both the Council and the Parliament, the legislative act is finally adopted. If the conciliation committee fails to adopt a joint text, or adopts a joint text but not later approved, the legislative act is not adopted.

The European Court of Justice is designed to be the supreme court of the European Union by ensuring full compliance with European law by the Member States and European citizens. Seated in Luxembourg, the European Court of Justice is composed of independent judges from different Member States representing the different legal traditions present in Europe. Helped by the General Court and the Tribunal for Civil Servants, the European Court of Justice can hear cases either directly or from national tribunals (Turk 2010).

The Council of the European Union composed of national ministers represents the Member States with each Member States being allocated different voting weights according to its political power.

The European Council, originally in 1974 an informal of the head of states and governments of the Member States, has been recognized as a European institution in 1992. This high-level meeting instills the political impetus to the European Commission and, since the Lisbon Treaty of 2009, is chaired by the President of the European Union. The President is elected by the member of the European Council and its term lasts 2 years and half, renewable once.

This sophisticated institutional framework progressively, through treaty amendments, resembles one of a federation of states despite an added complexity due to the reluctance of Member States to renege upon their sovereignties.

The European institutions have been granted powers independent from Member States only to the extent that the Member States benefited from these delegations of powers in favor of these institutions acting as their agents. Indeed, the European institutions have been tasked with the completion of an internal market within the European Union. In order reach such result, the necessary functions have been delegated to the European institutions – this is the “neo-functionalism” theory as we shall discuss below.

The European Law

The completion of the internal market required European law to materialize the economic objectives of the European integration – namely, a customs union, the four economic freedoms, and a competition policy. European economic integration could only emerge through European legal integration. “Integration through law” has been the major force toward the completion of the internal market, with the European Court of Justice working as the primary engine of integration.

First, the European Court of Justice has elaborated a thought-provoking case law ensuring the greatest effectiveness of European law over national laws (Weiler 1991; Turk 2010). The European judges have granted to European law direct effect, making most of European law directly invokable by litigants before any national and European court. Also, the European judges have trumped national laws by continuously proclaiming the primacy of European law over any national laws (Weiler 1991). Moreover, the responsibility of Member States has been engaged any time European law is not respected in the given Member State by whatever national institution, citizen, or organization of that state. Also, the European Court of Justice reaffirmed its role as the supreme court of Europe by interpreting largely its monopoly of interpretation of European law and of annulling European law in light of the European Treaties (Weiler 1991). Therefore, not only have European judges confirmed the essential role of European law in the national legal orders, but also have they placed themselves as powerful judges in the most influential court in Europe (Bernard 2010).

Second, after having ensured its maximum effectiveness, European law has been developed and interpreted in order to maximize the extent to which the process of building the internal market is conveyed. Ernst Haas, one of the fathers of the functionalism theory, said that “the most inviting index of integration – because it can be verified statistically – is the economic one.”

Market-building requires rules on the prohibition of discrimination since goods, services, capital, and labor must compete within a market on

a fair and open basis. The customs union is efficiency-enhancing on one part because it creates some trade but can also be efficiency-decreasing on the other part because trade can be diverted due to the discriminatory rule a customs union creates in favor of its members and in disfavor of the nonmembers. Hence, the customs duties (Article 30 TFEU) and “measures having an equivalent effect” (Article 34) are prohibited within the EU after the 12-year transitory period. The notion of “measures having an equivalent effect” has been given great scope by EU judges to encompass lots of regulatory measures. Indeed, “all trading rules enacted by Member States which are capable of hindering directly or indirectly, actually or potentially, intra-Community trade are to be considered as measures having equivalent effect to quantitative restrictions” said the EU judges in the famous *Dassonville* case of 1974. Despite the fact that some mandatory requirements can potentially be accepted as justification for some national regulations (see case *Cassis de Dijon* of 1979) above the justifications provided for in the treaties, the *Dassonville* formula has been continuously reaffirmed by the case law (Bernard 2010; Turk 2010).

Therefore, the EU legal order has been designed around the notion of competitive order – in other words, the internal market requires a law that ensures that competitive forces work openly, fairly, and on an equal basis for all market participants. Thus, in order to limit the reduction of economic efficiency and to increase the internal efficiency of the market, two legal dispositions had to be added at the top of the prohibition of customs duties and measures having an equivalent to quantitative restrictions. First, a nondiscrimination rule had to be an overarching principle of market-building. Second, a strong enforcement of a common competition law had to be ensured. The European successfully encapsulate both dispositions so that the internal market could be considered to have been completed by 1992 with the Maastricht Treaty.

First, the prohibition of discrimination on the basis of nationality enshrined in Article 6 of the Treaty of the European Union, in Article 21 of the Charter of Fundamental Rights of 2000 (a sort of

EU “Bill of Rights”), ensures equality in the internal market, hence its smooth working. It is no surprise that the EU judges have interpreted the nondiscrimination principle in a very ambitious manner. The nondiscrimination principle applied very broadly in the field of the free movement of goods is progressively applied in a comprehensive manner to other economic freedoms. Hence, the “market access” has, in the eyes of the EU judges, to be secured in the broadest manner in all fields of the internal market through the resort of the nondiscrimination principle writ large (Bernard 2010). Market access must be ensured due to the benefits of transactional efficiency (in terms of minimization of transaction costs created by national regulations) expected to be reaped out by internal market participants (Portuese 2012, pp. 361–402).

Second, European competition law had preserved the internal market from obstacles to its achievement (the so-called objective of “market access”). The basic EU competition rules are found in the TFEU at Articles 101 and 102, which lay down a prohibition against restrictive agreements (cartels and collusions) and abuse of dominance, respectively. Articles 101 and 102 TFEU apply to practices or behavior that “may affect trade between Member States” (Toth 2008). The concept of “trade between Member States,” defined in Commission Guidelines (No 2004/C 101/07 of 27 April 2004), implies that there must be cross-border economic activity involving at least two Member States – but, when trade between Member States might be affected, a restrictive agreement applied only in a single Member State will not be excluded from the scope of EU competition rules. Also, mergers are controlled according to Regulation 139/2004 of 2004. Finally, state aids are circumscribed in accordance to Article 107 of the TFEU. Entrusted with the task of investigating anticompetitive practices, the European Commission is invited to refer to the for potential cases to be tried. The high enforcement of European competition rules have participated in making the internal market a place where a fair and undistorted competition provides the economies efficiencies of a dynamic economy (Bernard 2010).

Overall, the removal of customs duties and charges having equivalent effect, the respect of the nondiscrimination and equality principle, and the strong enforcement of European competition rules have contributed to materialize European integration over decades. If European integration has been propelled by the European institutional actors such as the European Commission and the ECJ, it is also due to the particular appeal the pursuance of the notion of economic efficiency had on these supranational actors.

European Institutional Actors as Agents of Economic Efficiency: The Efficiency Hypothesis

The foundations of European law are substantiated in the general principles of EU Law. With the principle of equality and nondiscrimination discussed above, there are three main general principles that govern European law – the principle of subsidiarity, the principle of proportionality, and the principle of legal certainty.

The principle of subsidiarity is double-edged. Enshrined in the Treaty of the EU at Article 5 (3), it can be used to justify both increased centralization and increased decentralization. The principle of subsidiarity can justify further centralization because it requires that the most appropriate level of governance be chosen for exercising a particular power. A study of the principle in general, and in practice as interpreted by the EU judges, reveals that the principle of subsidiarity both is a principle of economic governance and is interpreted in an economically efficient way by the EU judges (Portuese 2011).

The principle of proportionality is a general principle of EU law explicitly stated in the EU Treaties nowadays at Article 5 (4). The EU proportionality principle can be divided in different sub-principles: (i) the review of the necessity of the measures to achieve the desired objective, (ii) the review of the suitability (or less-restrictive means test) of the measures for the achievement of the objective, and (iii) the review of the proportionality *stricto sensu*, whereby the burden imposed must be of proportion with the goal desired. If the first sub-principle is tantamount to an effectiveness criterion, the second

sub-principle ensures the efficiency (ratio means-ends) of the measure, while the third sub-principle involves a balancing exercise of interests similar to a cost-benefit analysis which itself fosters Kaldor-Hicks efficiency. Therefore, the overall principle is and is interpreted as an efficiency principle (Portuese 2013).

Finally, the EU principle of legal certainty protects the legitimate expectation of market actors. Therefore, it ensures that changes in the law occur at a predictable path and in a reasonable way. If this guarantee is not respected, then damages must be awarded in order to compensate the loss of market actors' proprietary interests. This principle has been jurisprudentially created in the EU, so the European integration can happen without loss of proprietary interests, thus without disincentivized internal market actors (Portuese 2014a, b).

Along the general rules of European integration such as free movement rules and competitions rules, the general principles of EU law are efficiency-enhancing rules and have been interpreted as such consistently by the EU judges.

There is no coincidence however.

Indeed, one must recall the high distrust Member States had one another in the aftermath of the Second World War. Therefore, a relatively strong supranational institutionalization has been erected in order to eschew reciprocal and the tit-for-tat ill-fated strategies of classical treaties. Consequently, any political answers favoring one specific Member State or a given group of Member States at the expense of other Member States had to be seriously sidestepped. Indeed, welfare distribution would have been seen at that time as a means for a Member State to get a hold on the overall European integration. Thus, the European construction was only left, politically speaking, with a solution of welfare creation: an economic solution secured by the law (Portuese 2012).

In that prospect, economic efficiency is a rather neutral notion, despite its criticisms. It does not favor a specific Member State over another; it promotes the general interest of Europe inasmuch as supranational institutions were tasked to promote. Hence, supranational institutions of the EU have been keen to develop the economic

efficiency within the internal market because it is both an engine of integration and, very notably and interestingly for them, a politically accepted objective.

In light of the cursory economic analysis of European law discussed above, it can be said that there is tendency of European law in general and of European case law in particular to develop in sense that promotes economic efficiency. Therefore, a hypothesis of the economic efficiency of the European case law can be made, by reference to the classical and contested economic efficiency hypothesis of the common law (Portuese 2012).

If European integration has been in great majority about economic integration, the creation of a supranational polity having manifest powers over the economic environments of Member States but without possessing the democratic legitimacy of national democracies raises the everlasting concerns of European integration.

The Law and Politics of European Integration

If European integration has been for long driven by economic grounds unfolded by the ECJ case law (Turk 2010), the political integration was more diffident since Member States were both reluctant to share the prestige of their national legitimacies and to set up, from the onset or later at the end of twentieth century, a European polity capable of being compared to any form of federation or superstate (Weiler 2000). This reluctance can be understandable but is nevertheless problematic for at least two reasons: the very essence of European law becomes under fire given the low legitimacy of this “undemocratic” law (a), and the very essence of the European institutions also evolves into an easy target for criticisms due to the lack of the commonly recognizable features of national institutions shared by the European institutions (b).

In fact, the European Union is based on an inverted construction where market-building has preceded and primed over institutions building. The European Union, a *sui generis* organization,

can be defined by the concept I introduce of *soft federation* (c).

Harmonizing Details and Enlarging High-Tail: Challenges to the Law

The European construction has developed European law to an extent that it becomes difficult to find sectors of the economy where regulations have not been harmonized at the European law. Indeed, the EU works predominantly as a regulator, privileging the normative power over military power.

The EU can only regulate, however, when it is directly related with the internal market. This is in accordance with the principle of conferral of powers.

When sectors are not harmonized, the so-called principle of mutual recognition applies: it means that the national rules have to be mutually recognized by the other Member States (Armstrong 2003). Hence, traders do not have to comply with extra rules that are purported to achieve objectives for which the rule of their home country has already been designed to achieve. Extra national regulatory costs are avoided unless mandatory requirements justify compliance with the host state rule. This process is often called “negative harmonization.”

Therefore, as an incentive, the European institutions have been keen to harmonize at great pace. The removal of all barriers to trade can sometimes lead to the hastened removal of national rules. Therefore, harmonization with one single European rule is needed in order to foster the integration of the internal market. This process is often called “positive harmonization”.

Negative harmonization is deregulatory; positive harmonization is re-regulatory.

The interesting point to raise in the following lines is not about how much, what, and when have we harmonized regulations in Europe. It is more about what consequences this process of continuous harmonization bears as challenge to the law as we understand it.

Coupled with the great pace of harmonization is the issue of enlargement (Weiler 2003). Not only have we integrated through harmonization with European rules throughout the European

construction, but also we have enlarged the EU from 6 founding Member States with 169 million inhabitants to 28 Member States with 500 million inhabitants, and it keeps enlarging!

These two major trends of European integration beg the question whether or not European law has enough democratic legitimacy in order to keep both harmonizing in its depth and enlarging in its ambit.

European regulations have recently been under greater scrutiny as for their costs and legitimacy, with the so-called Regulatory Impact Assessment ("RIA") that intends to balance the costs and benefits of each regulation. This tool has been developed under the framework of the "Better Regulation" program – the EU has to be a better regulator rather than a merely effective regulator. But the ambition of better regulating the EU leads to the very notion of legislative acceptability of EU law by citizens with respect to as EU law is perceived.

The law is traditionally understood as being a body of normative rules coherently forming a legal order and applied to a specific territory where subjects are legally binding themselves to this order with full consent. In other words, should the subjects change their mind in a democratically erected legal order, the institutions will be responsive and change the law. Here is the necessity of not binding one generation with another. That's why we have this expression "King in his Parliament": the rules in force are only those expressly or implicitly consented by the people to comply with.

But, as EU institutions can increasingly powers of economic and social regulations, the possibility for national democratically elected representatives to change the law applicable to their citizens. This argument does not work only for deregulating purposes: the mere changes in law are extremely difficult due to the decision-making process taking place at European level that leaves national representatives unpowered.

The notion of EU legal order faces multiple challenges not only for competing with national legal orders that bear high democratic legitimacy but also for ascertaining a minimum prestige owed by European citizens to their European legal order. This prestige is not superfluous, it is

fundamental: a law without prestige is no law inasmuch as a law without sanction is no law. European law with little prestige hardly is law.

National Democracies and Supranational Governance: Challenges to the *Politeia*

If European law faces great challenges in terms of its acceptance and position with respect to the perception of national legal orders, it is also because European integration, again, has been predominantly about economic integration.

The failure of planning, the inability to foresee, the absence of courage, and the reluctance to go supranational have impeded the necessary process of polity-building required by any market-building process. Indeed, the *Politeia* (the Greek word for "polity") defines a specific form of government into which citizens can believe be part of. It does not necessarily mean "democratic" as contemporary understanding evoke; neither does it evolves the notion of Republic nor of nation-state. It only refers to the requirement of having clear and intelligible institutions that citizens recognize as legitimate. Unfortunately, despite efforts made by the Member States and the European institutions to "democratize" the EU by building a supranational polity, the EU institutions are still perceived as far from European citizens. Indeed, the organization since 1979 of direct elections for the European Parliament has not helped this core institution to get credit or to avoid lots of abstentions among voters who are European citizens. Also, the recent creation, with the Lisbon Treaty, of the President of Europe elected by the European Council has not helped European citizens to identify Herman Van Rompuy, the first European President, nor did it helped European citizens to feel he is their European representative (Weiler 2000).

Lots of academic debates are of little help when it comes to increasing the democratic legitimacy of European institutions and to focusing more on polity-building rather than on market-building. One trend of the literature, new functionalism, has argued that the step-by-step integration would "spill over" continuously so that, eventually, an integrated and federated Europe would come to the fore. Within that process, the rise of

a political Europe would emerge as economic integration cannot last without institutional integration (Haas 2004; Sandholtz and Stone-Sweet 1998; Stone Sweet and Fligstein 2001). This process is triggered principally by supranational institutions who act with great independence from Member States. Another trend of literature, labeled “liberal intergovernmentalism,” argues contrariwise that Member States have retained, throughout the European construction, considerable latitude over the European institutions’ powers (Moravcsik 1993). The dynamics of European integration has been, according to that perspective, aligned with the powerful Member States’ interests who succeeded in deepening European integration while preserving their powers.

Beyond this academic debate lays the discussion on a polity-building at the European level which could, if achieved successfully, deliver to the market-building process an enhanced democratic legitimacy. The failure to envision such polity-building or the reluctance to achieve it can be explained by the very nature of the EU. The EU actually is, as I coin, a “soft federation.”

The Political Economy of the Current European Integration: The Soft Federation Hypothesis

In law, soft law is a well-known expression used to designate some quasi-normative rules that are neither hard law because they lack the enforceability of legal rules nor no law because they carry some normative guidelines of certain legal valence.

This concept of soft law can be extrapolated to political sciences and defines a federation that is both insufficiently powerful to be designated as a “hard federation” and insufficiently powerless to be designated as a confederation.

The establishment of a federation generally comes with both the exclusive recognizance of the federal state as depository of sovereignty (both internally and externally), with the creation of police and military powers, the exclusive power over monetary powers, and a preeminence over budgetary powers, and the exclusive power to carry out diplomatic activities. Contrariwise, the federation can intervene less often and deeply on

economic and social matters, the setting up of normative standards, and the policy interventions in fields such as education and research, environmental protection, health policy, etc. But on the other hand, the EU can no longer be considered as a confederation since majority voting has been generalized, whereas confederation are defined by unanimity voting in order to respect each Member State’s sovereignty.

Therefore, an inverse institutional architecture appears compared to classic federation framework: the most integrated parts within the EU are the least one in a federation, while the least integrated parts of the EU are at the essence of the federal level. Therefore, the EU is a federation without the common powers of a federation – it is a “soft federation.”

This hybrid situation is not a middle-ground situation: it may very well last for long, and hence, the “soft federation” format may become a new form of government rather than a temporary one if the “soft federation” hypothesis is confirmed over time.

Conclusion

European integration has been a great success in terms of enhancing the economic efficiency through the process of market-building, while it has remained a great disillusionment in terms of enhancing democratic legitimacy through the process of polity-building.

The two major trends identified in European integration lead to formulate a dual hypothesis: an efficiency hypothesis whereby supranational actors push for internal market rules to promote economic efficiency and a “soft federation” hypothesis whereby Member States push for the emergence of a peculiar federation which protect the core of national sovereignties.

Overall, European integration is a unique process successful enough to witness trials of replications throughout the world (e.g., African Union, Union of South American Nations, Arab League, etc.). This is certainly the strongest proxy for evaluating the results of 60 years of European integration.

References

- Bernard C (2010) *The substantive law of the EU: the four freedoms*, 3rd edn. Oxford University Press, Oxford
- Duchene F (1994) *Jean Monnet. The first statesman of interdependence*. Norton, New York
- Haas EB (2004) Introduction: institutionalisation or constructivism? In: *The uniting of Europe: political, social and economic forces 1950–1957*, 3rd edn. University of Notre Dame Press, Notre Dame, pp xiii–lvi
- Kant I (1992) *Perpetual peace. A philosophical essay*. Thoemmes Press, Bristol
- Montesquieu B (1758) *Oeuvres Complètes: De l'Esprit des Lois (Complete works: spirit of the laws)*. Gallimard/Pléiade, Paris (1951)
- Moravcsik A (1993) Preferences and power in the European community: a liberal intergovernmentalist approach. *J Common Mark Stud* 31:473–524
- Pelkmans J (2006) *European integration, methods and economic analysis*. Pearson Education/Financial Times, Harlow/Prentice Hall
- Portuese A (2011) The principle of subsidiarity as a principle of economic efficiency. *Columbia J Eur Law* 17:231
- Portuese A (2012) The principle of economic efficiency in the ECJ case-law (*Le Principe d'Efficiencce Economique dans la Jurisprudence Européenne*). Thèse de l'Université Paris II Panthéon-Assas, Paris. Accessible à: <https://docassas.u-paris2.fr/nuxeo/site/esupversions/c51d2b4b-4981-4c62-8531-285d0804dc05>
- Portuese A (2013) Principle of proportionality as principle of economic efficiency. *Eur Law J* 19:612–635
- Portuese A (2014a) The principle of legal certainty as a principle of economic efficiency. *Eur J Law Econ*. <https://doi.org/10.1007/s10657-014-9435-2>
- Portuese A (2014b) The case for a principled approach to law and economics: efficiency analysis and general principles of EU law. In: Mathis K (ed) *Law and economics in Europe, Foundations and applications*. Springer Books, Dordrecht, pp 275–328
- Stone Sweet A, Fligstein N (2001) *The institutionalisation of Europe*. Oxford University Press, Oxford
- Toth AG (2008) *The Oxford encyclopaedia of European community law*, vol II: *The law of the internal market* (2005); vol III: *Competition law and policy*. Oxford University Press, Oxford
- Turk A (2010) *Judicial review in the EU*. Elgar Publishing, London
- Weiler JHH (1991) The transformation of Europe. *Yale Law J* 100:2403–2483
- Weiler JHH (2000) Epilogue The Fischer Debate In: Joerges C, Mény Y, Weiler JHH (eds) *What kind of constitution for what kind of polity?: responses to Joschka Fischer*. Robert Schuman Centre, EUI
- Weiler JHH (2003) In: Begg I, Peterson J, Weiler JHH (eds) *Integration in an expanding European Union: reassessing the fundamentals*. Blackwell Publishing, London
- ## Further Reading
- Armstrong KA (2003) Governance and the single European market. In: Craig P (ed) *The evolution of EU law*. Oxford University Press, Oxford, pp 745–789
- Baldwin R, Wyplosz C (2009) *Economics of European integration*, 3 rev edn. McGraw Hill Higher Education, London
- Cecchini P (1988) *The European challenge 1992*. Wildwood House, Aldershot
- CEPR (1993) *Making sense of subsidiarity: how much centralisation for Europe?* London (study group Begg et al)
- CEPR (2003) *Built to last: a political architecture for Europe*. London
- Cohn-Bendit D, Verhofstadt G (2012) *For Europe! Manifesto for a postnational revolution in Europe*. Carl Hanser Verlag, München
- Craig P, De Burca G (2011) *EU Law. Text, cases and materials*, 5th edn. Oxford University Press, Oxford
- Diez T, Wiener A (2004) *European integration theory*. Oxford University Press, Oxford
- Eeckhout P (2005) *External relations of the European Union: legal and constitutional foundations*. Oxford University Press, Oxford
- Gilbert M (2012) *European integration: a concise history*. Rowman & Littlefield, London
- Hix S (2005) *The political system of the European Union*, 2nd edn. Palgrave, London
- Holle W (2006) *The economics of European integration: theory, practice, policy*. Ashgate, London
- Howells G (2002) Federalism in USA and EC – the scope for harmonised legislative activity compared. *ERPL* 10:601–622
- Olsen JP (2001) Organizing European institutions of governance: a prelude to an institutional account of political integration. In: Wallace H (ed) *Interlocking dimensions of European integration*. Palgrave, London
- Pelkmans J (2002) Mutual recognition in goods and services: an economic perspective. College of Europe, Bruges. BEEP briefings no 2, December. www.coleurop.be
- Pelkmans J, Labory S, Majone G (2000) Better EU regulatory quality: assessing current initiatives and new proposals. In: Galli G, Pelkmans J (eds) *Regulatory reform and competitiveness in Europe*, vol 1. E. Elgar, Cheltenham
- Sandholtz W, Stone-Sweet A (1998) *European integration and supranational governance*. Oxford University Press, Oxford/New York
- Smith A, Venables A (1991) Economic integration and market access. *Eur Econ Rev* 35:388–395
- Weatherill S (1995) *Law and integration in the European Union*. Clarendon, London
- Weatherill S (2005) *Harmonisation: how much, how little*. EBLR 533
- Woods L (2004) *Free movement of goods and services within the European community*. Ashgate, London

European Law

► European Community Law

European Nationality

Paul-Ludwig Weinacht
Güntersleben, Germany
Würzburg University, Würzburg, Germany

Synonyms

EU citizenship; European identity;
Europeanization

Definition

European Nationality (EN) is a concept with two meanings: the present legal, political, and cultural reality of the EU (“actual real concept”) and the possible futures of the EU (“target oriented concept”). Both sides of the concept, the “actual real” and “target oriented” ones, are employed in the ongoing process of enlargement and consolidation processes of the 28 (27) EU member states.

An Actual Concept

As an *actual concept* of our legal reality, EN means the EU-citizenship which since 1991 is tied to the primary national citizenship of an EU member state as a secondary one. In the treaty of Maastricht, all citizens of EU member states received the status of general and equal EU-citizenship (Art 17 sq EC). It does not replace the national citizenship, who’s lending or withdrawal remains in the sovereignty of member states – however it complements it in several aspects. The EU-citizenship offers firstly the freedom to settle down and to establish oneself in all member states (Art. 18 EC); secondly it gives the

right to vote in local elections of the state in which someone is legally settled and the right to vote for the European Parliament (Art. 19 EC). Thirdly, the right of diplomatic and consular protection (Art. 21 EC), the right of petition, the right to turn in writing to the organs of the community, p. e. to the representative of the citizens, and to get an answer in the chosen language (Art. 21 EC).

The program by the EP “Europe for the citizens” (2007–2013) intends “to integrate the citizens better in the process of the European unification”; it promotes “the co-operation between the citizens and their organisations from different nations.” The Single European Market is the most integrated institution in Europe. Here the famous “four freedoms” are granted since 1987: free traffic of persons, of services, of goods, of capital. These freedoms are limiting more and more the sovereign action of member states. They act – if necessary by the European Court – as engines for the strengthening of the political system of the EU.

Concerning the political and sociocultural meaning of EN, the EU-citizenship with its rights in annex is largely pre-determined by national preconditions. Conflicts in the context with the banks crisis and in the Brexit-movement in Great Britain showed it. Until now, European citizens as a whole do not form a nation. They do not fulfil the criteria of a nation: neither in the sense of the German or slave concept of nation (origin, history, language, etc.) nor in the French understanding of the word (support for the constitution of one and un-dividable republic).

The EU-citizens in their respective states have a pragmatic relation to the question of more intensive adherence to their nation or to Europe. The answers measured by the Eurobarometer-Survey are often: adherent to collective identities. This refers to “multiples” collective identities in the EU (Trenz et al. 2003, p 17). The mind of the citizen is influenced by the preparation and implementation of the election to the EP and – independently from the election campaign – intensified by culture programs of the EC (2007–2013). “Integration” and “participation” are headwords for the construction of a

well-united European Community, whose identity is based on commonly accepted values (<http://europa.eu/legislation>).

A Target-Oriented Concept

As a *target-oriented concept*, EN is precarious in so far as it applies to the controversial project of a “sovereign European state.” Pro-European politicians like to symbolize the way to this kind of state by a dynamic concept which embraces different fields: “Europeanization.” If the European citizenship would accept criteria of a sovereign state, it would be superior to the national citizenship. In the face of a sovereign position of the EU, some authors mean that common values could be effective in favor of the enlargement of European politics. A specialist means that strong structures and a dense net of competences in European institutions favor the “European identity and public spheres” (Risse 2010).

Prototypical forms of nation-building could serve as an historical example: the French nation as a “product of the state,” because its origin is a sovereign monarchic rule; the German nation as “the base of the state,” because it was – as a type of “community of cultural origin” (Weinacht 1994) – historically prior to the state (Brubaker 1994). The idea of Europeanization could be discussed too at the historic example of England: The political institutions created the English nation, and the political institutions were legitimated by the national English culture (Schulze 2004). It is questionable, if the insinuated lines of nation and institution building will result in a clear definite idea of state and if such an objective will be generally accepted. It is not without reason that the EU-Commission avoids fixing its institutional targets on a definite idea of state (“finality”) and speaks about “an ever closer union among the peoples of Europe” (Treatise of Maastricht 1992/93).

The “Europe of citizens,” which is intended, grows for the time being beside the nations in the form of cross-border sectors of a civil society, which develop both pro-European

and anti-European tendencies. Supporters of this phenomenon try to see in a European “active society of citizen” a new model (Münch 1999).

Is EN Desirable?

Concerning the desirability of European nationality, the Treaty of Maastricht (1992/93) defines the target of the EU as an “ever closer union among the peoples of Europe.” However some Eastern European (“Visegrad”) States have lost this consensus and tend – like France under de Gaulle (“Europe des patries”) – to a “Europe of nations.” With the referendum from June 2016, Great Britain cancelled the treaty of Maastricht in favor of an independent future for its nations. However, the aftereffects of Brexit are controversial. Britons who feel and wish to be part of the European project will loose with European citizenship the privileges inherent in it (especially “freedom of movement”). Many of them have begun to apply for the citizenship of their host country. The Irish embassy in London received 8017 applications for Irish nationality, compared to only 689 applications in 2015. The reason was not the search for a new identity, but the desire to maintain security of residence and the right to travel and live in the 27 countries that will remain members of the EU after the UK leaves (Stone 2017). Politicians and journalists made proposals in favor of a European associate citizenship “for nationals of a former member state” (Henley 2016; Stone 2016).

A European Citizens’ Initiative (ECI) came out informally known as “Flock Brexit” with a Facebook group titled “EU Citizenship for Europeans: United in Diversity in Spite of jus soli and jus sanguineus.” It demonstrates not only the pure pragmatic character of EU Citizenship but Citizenship without political Membership. Although political rights are demanded sometimes, European citizenship is no more than a title associated with the citizenship of an EU member state.

In legal terms, which are underlying the status of European citizen, EN is an actual concept. Its

actual value remains in transnational civil rights, especially the right of movement and entitlement to the same treatment including welfare rights, as the citizens of the country of residence. Prudent politicians declare that the European Union does not substitute the sovereign nations but gives them forward compatibility. That's why EN would remain for a long period a secondary identity.

The transition from the formal status of EU-citizen to a substantial European nationality, realizing what the concept oriented to targets means, is in a quite distant future.

Cross-References

- [European Community Law](#)
- [European Integration](#)

References

- Brubaker R (1994) *Citizenship and nationhood in France and Germany*. Harvard University Press, Cambridge, MA
- Henley J (2016) EU citizenship proposal could guarantee rights in Europe after Brexit. *The Guardian*, 8 November
- Münch R (1999) The transformation of citizenship in the global age. In: Kriesi H et al (eds) *Nation and national identity*. Verlag Rüegg, Chur/Zürich, pp 109–125
- Risse T (2010) *A community of Europeans?* Cornell University Press, Ithaca/London
- Schulze H (2004) *States, nations and nationalism: from the Middle Ages to the present*. Blackwell, Oxford
- Stone J (2016) How to keep your EU citizenship after Brexit, British nationals are set to lose out, but there are some workarounds, 14 June, BST
- Stone J (2017) Brexit vote creates surge in EU citizenship applications. *The Guardian*, 3 October
- Trenz H-J et al (2003) Demokratie-, Öffentlichkeits- und Identitätsdefizite in der EU. In: Klein A, Koopmans R, Trenc H-J et al (eds) *Bürgerschaft, Öffentlichkeit und Demokratie in Europa*. Leske und Budrich, Opladen pp 7–19
- Weinacht P-L (1994) Nation als Integral moderner Gesellschaft. In: Gebhardt J, Schmalz-Bruns R (eds) *Demokratie, Verfassung und Nation. Die politische Integration moderner Gesellschaft*. Nomos Verl.-Ges, Baden-Baden, pp 102–122

European Patent System

Bruno van Pottelsberghe de la Potterie
Solvay Brussels School of Economics and Management, Université libre de Bruxelles, Brussels, Belgium

Definition

The European patent system is defined as the policy mechanisms, jurisdictions, and institutions in Europe which allow inventors to acquire and enforce industrial property rights over their inventions.

Introduction

The European patent system is a terminology which is often oversimplified and described through the European Patent Convention (EPC). This is a multilateral treaty signed in October 1973 which essentially consists in creating the European Patent Organisation and in providing a legal system for the granting of European Patents. EPC patents, or so-called “European” Patents, are actually not really European, in the sense that they have to be translated and validated in national patent offices in order to be enforceable in national jurisdictions. The European patent system is more complex than the EPC, as it includes three layers of legal rights (national patents, the current European Patent, and the forthcoming Unitary Patent). Two types of patent offices grant these rights: national patent offices (which grant national patents) and the European Patent Office for the granting process of European and Unitary Patents. Litigations currently take place at the national level, and for unitary patents patent litigations should be treated by a Unified Patent Court, whose design is currently being framed and negotiated, as of January 2015. This entry aims at providing a wide understanding of the European patent system, its strengths, its weaknesses, and its challenges.

Chronology of Three Patent “Layers”

As of the late nineteenth century, all European countries had a national patent office that would grant national *patents*. The broad patentability conditions, found in most patent systems around the world, include novelty and inventiveness. In order to be granted, a patent must be novel (i.e., it should not have been published or made public prior to the application), and it must be inventive (there must be an inventive step, or non-obviousness, with respect to the prior art). If the patent owner aimed at extending its geographical coverage, it had to file subsequent applications in other European countries. This was particularly complex and costly, as it led to

additional translation costs, application fees, and examination fees, not to mention renewal fees to maintain the patent in force once it was granted, for a maximum duration of 20 years from the first date of application. According to the Paris Convention (March 1883), which is still in force today, patent owners have maximum 1 year after their first application to extend their patent abroad while not being refused on the ground of the novelty condition (a patent published elsewhere would constitute a prior art) (Table 1).

A second layer of patent protection was activated in 1978. It is known as the “European Patent” and was set up through the creation of the EPC, signed by 38 countries, as of June 2014 (including all countries of the European Union).

European Patent System, Table 1 Patent applications by patent office and origin, 2013

	Applications		Patents granted	Patents in force	Cumulated fees (4Y)
Name	Total	% nonresident (%)	Total	Total	EUR
Austria	2,406	10	1,256	110,202	830
Belgium	876	18	745	..	485
Bulgaria	297	5	125	1,431	n.a.
Croatia	253	9	159	4,243	n.a.
Czech Republic	1,081	9	611	7,780	n.a.
Denmark	1,534	13	309	51,277	1,398
Finland	1,737	8	711	47,058	1,890
France	16,886	13	11,405	500,114	728
Germany	63,167	25	13,858	569,340	988
Greece	717	3	282	2,966	513
Hungary	708	9	1,351	5,237	807
Ireland	390	15	214	108,218	700
Italy	9,212	10	8,114	68,000	1,160
Luxembourg	169	33	112	20,421	334
Netherlands	2,764	16	2,029	12,704	252
Poland	4,411	4	2,804	47,610	235
Portugal	669	3	130	36,782	200
Romania	1,046	5	451	17,100	550
Slovakia	210	12	115	2,755	n.a.
Spain	3,244	7	3,004	36,893	510
Sweden	2,495	7	685	14,539	658
UK	22,938	35	5,235	469,941	229
Total national offices (EU)	132,921	22	49,056	2,134,611	n.r.
European Patent office	147,987	50	66,696	n.r.	4,309

Source: figures on patent applications, granted and in force, are from WIPO statistical series released in 2014. Fees are from de Rassenfosse and van Pottelsberghe (2013) and represent cumulated fees from application to grant, in EUR

In short, once granted by the European Patent Office, a patent must be managed and enforced at the national level. This “European” terminology is therefore often confusing, and certainly inadequate, as the “European” dimension only takes place for the processing of the search report and the substantive examination of patent applications. They are centrally granted by the European Patent Office. Once granted these patents must still be validated and renewed in the national patent offices that fit with the desired market coverage of the patent owner. The cost of a European Patent is therefore relatively high and depends on its geographical coverage, as inventors must pay their validation fees and several other costs (intermediates, patent attorneys, renewal fees). An opposition procedure before the EPO can be initiated by any person or institution during a period of 9 months after the date of decision to grant a patent. Any subsequent litigation is operated at the national level (see Guellec and van Pottelsberghe 2007 for a broad introduction to the economics of the European patent system).

Table 1 presents these first two “layers” of patent protection. National patent offices receive national patent applications (ranging from a few hundreds in small countries to more than 60,000 in Germany and 23,000 in the UK). The cumulated number of patent applications at national patent offices (about 132,000) is smaller than the 150,000 patent applications filed at the EPO. Interestingly, the EPO is particularly targeted by foreign companies, as 51% of its patent applications are filed by non-EU residents. National patent offices are less targeted by foreign companies (only 21% of patent applications are filed by nonresidents).

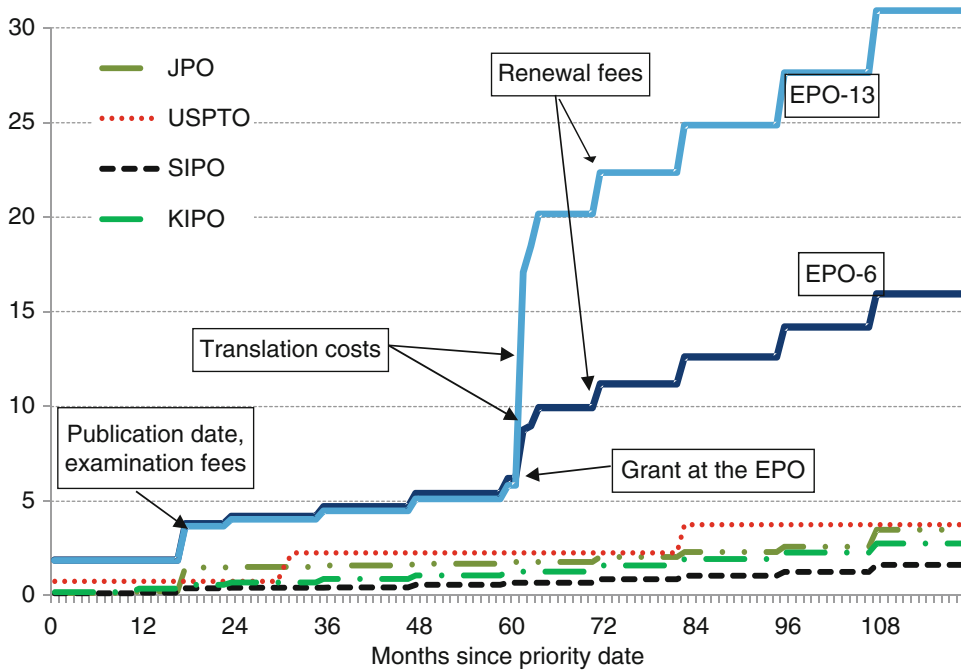
Countries inside Europe have adopted different fee structures and policies. The last column of Table 1 shows that the cumulated fees from application to grant fluctuate from less than 300 EUR in the UK or the Netherlands to more than 1,000 EUR in Italy, Finland, and Denmark. When it comes to the European Patent, the fees for patent prosecution before the EPO are much higher than in national patent offices, with more than 4,000 EUR of cumulated fees, from application to grant. And these cumulated costs do not reflect the

maintenance rates of European Patents in each national jurisdiction.

The fragmentation of the European market is ultimately illustrated by the very strong heterogeneity in the number of patent in force across countries. Germany, France, and the UK have by far the largest number of patents in force, more than 450,000 (be they granted by the EPO or by their national patent offices). All other EU countries have much less than 100,000 patents in force, due to their smaller size and hence lower market attractiveness. It is worth noticing that a European Patent becomes particularly expensive once it is granted, due to the translation costs and cumulated renewal fees. This is well illustrated in Fig. 1, which shows the evolution of cumulated costs for patents that aim at being put in force in 6 or in 13 European countries, as compared with Chinese, Japanese, South Korean, or US patents.

The main consequences of the system that combines national and “European” Patents are described in Mejer and van Pottelsberghe (2012) and in van Pottelsberghe (2010). The most visible one is related to the prohibitive costs of patent protection in Europe, as compared to several other regions in the world. A European Patent with protection secured for 10 years in 13 countries will cost more than 30,000 EUR, against less than 5,000 in the USA, Canada, China, Brazil, Japan, or South Korea.

Another important consequence of such fragmentation is related to the broad legal uncertainty associated with the European Patent, especially for technology-based spin-offs. Legal uncertainty is the outcome of a dual system in which the EPO grants patents centrally but where national patent systems have the ultimate power to validate, invalidate, and assess infringement proceedings relevant to their own jurisdiction. European Patents which are particularly valuable have a high propensity to be litigated, in several countries. Parallel litigations are extremely expensive and time-consuming, especially for small entities. Then come the incongruities, whereby a patent can be maintained valid in one jurisdiction and invalidated in another. And even if a patent is maintained valid in several jurisdictions, they



European Patent System, Fig. 1 Ten-year cumulated patent fees for a European Patent (validated in 6 and 13 countries), compared with the US and three Asian countries (Data source: Mejer and van Pottelsberghe

(2011); cumulated costs (in Euro) from application to grant and renewal in 6 or 13 EU countries (for the EPO) and in the USPTO (USA), KIPO (South Korea), JPO (Japan), and SIPO (China))

may reach opposite conclusions regarding infringement proceedings. Parallel imports are also relatively easy, as infringing products might enter Europe through a country with no patent being enforced and then easily being distributed over Europe. An additional source of inconsistency deserves attention: the fact that national patent offices may grant a national patent even if it has not been applied for at the EPO or if it has been refused by the EPO. Procedurally, it is perfectly permissible to make simultaneous filings at one or several national patent offices and at the EPO. In other words, the granting process orchestrated by the EPO can be “bypassed” if one or more applications are made directly to national patent offices. This practice may have a number of explanations, some innocent (the applicant being interested by only one or two national markets), some less so (a perception that some national offices are a “soft touch” for applications compared with the EPO, and hence applications in the gray zone regarding their patentability can

legitimately be filed at the central and national levels).

These prohibitive costs and legal uncertainty are the main reasons which have prompted policy-makers, for more than 50 years, to try to create a truly European Patent, first called “Community” patent and later entitled “Unitary” Patent, a patent that would bring protection over the whole European Union. This *third layer* has been voted in 2013 at the EU Competitiveness Council at the Ministerial level (Regulation EU No. 1257/2012 for the introduction of the European Patent with unitary effect or the Unitary Patent and Regulation EU No. 1260/2012 for the translation arrangements for the Unitary Patent) and ratified by 25 countries (Croatia, Italy, and Spain did not sign the treaty). The regulations have been adopted by way of enhanced cooperation – a tool that allows a group of EU member states to go ahead with integration and legislative projects when the required majority in the Council cannot be achieved; it is used only as

a last resort. It consists in creating a Unitary Patent, which would be automatically valid in the 25 member states of the EU, at the request of the patent owner and once the patent is granted by the EPO. The Unitary Patent is therefore a new option for the holder of a European Patent, who will choose between the classical “European” route of selecting the desired states for protection and a unitary protection covering 25 member states. During a transitional period of 12 years, the patents granted in German or French will have to be translated into English, and the patents granted in English will have to be translated into German and French. The translations will not have any legal effect; they are for information purposes only.

In a nutshell, the legal architecture of the Unitary Patent includes the “Unitary Patent Regulation” which consists in the introduction of the European Patent with unitary effect. Then comes the “Translation Regulation,” which describes the translation arrangements for the European Patent with unitary effect. These two regulations will be applicable for a maximum of 25 EU member states only on the date of entry into force of the Agreement on a Unified Patent Court (UPC) and only in the countries which will ratify it. As of January 2015, the Unitary Patent is therefore far from being operational, several details of its implementation having still to be framed and negotiated, and the number of countries it will actually cover is still unknown. These details are the medium-term challenge of the European patent system.

Medium-Term Challenge: Designing the Right System

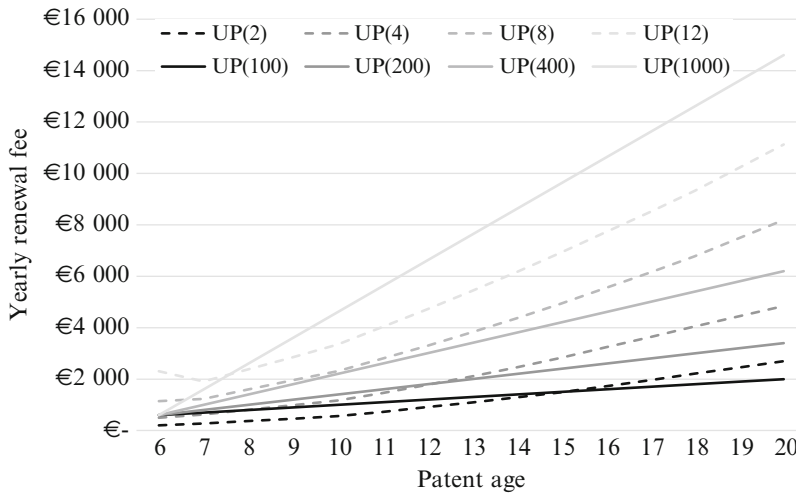
“Challenge” is an appropriate wording, because the decisions that will be taken by policy-makers, hopefully before June 2015, will substantially affect the use of the Unitary Patent and the broad effectiveness of the European patent system. Two important facets have to be framed: the level and structure of renewal fees and the setting up of a centralized litigation system, orchestrated by the Unified Patent Court.

Renewal Fees... For Who?

The Unitary Patent renewal fee structure has been the subject of intense negotiations for several years. This is a highly sensitive issue, as it has financial consequences for both the EPO and the national patent offices (NPOs), as well as for all the stakeholders of the European patent system: inventors, patent attorneys, lawyers, or translators. The renewal fees paid to NPOs by the owners of European Patents are split into two equal parts. Half the amount stays in the NPOs budget, and the other half is transferred to the EPO (cf. Danguy and van Pottelsberghe 2011a, b, 2014). For the NPOs of large frequently targeted countries, the patents granted by the European Patent Office generate massive resources (more than EUR 100 million in Germany and about EUR 50 million for France and the UK), equivalent to several times the annual working budget of their NPO. Even for smaller countries, the income generated by the renewal fees of European Patents exceeds by far the budget of their NPO induced by the proceeding of national patents. With the Unitary Patent, renewal fees would be collected by the EPO, and half the amount would be distributed across NPOs. The “renewal fee” debate is therefore not only related to the level these fees, but as well to the distribution key (what percentage will be allocated to each country, according to which criterion?). Some NPOs defend the position of setting up very high renewal fees, as they expect it will have a minor effect on their income (most patent owners would opt for the current “European” route, and those who will opt for the “Unitary” route will generate more resource).

In summary, as of January 2015 a self-fulfilling prophecy is in the air, whereby the decision taken by policy-makers on the renewal fees of the forthcoming Unitary Patent will affect its success and its perceived effectiveness (cf. Danguy and van Pottelsberghe 2014).

Two renewal fee schedules can be considered: (1) summing up the current European Patent fees of several countries, for instance, those in which granted patents are most frequently validated, like Germany, France, the UK, etc., and (2) adding a fixed increment each additional year of the patent



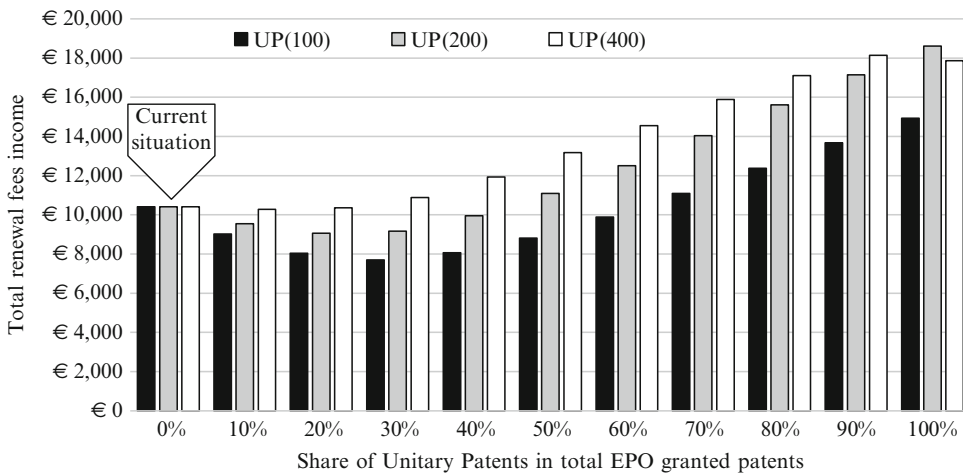
European Patent System, Fig. 2 Potential renewal fee schedules for the Unitary Patent (Note: *Dashed lines* show fee structures which sum the fees of X countries that are the most frequently targeted for protection. *Plain lines* show the fee structures with fixed annual increment (described

between parentheses, starting fee at EUR 600 on year 6, and then an increment of 100 EUR (200 EUR, 400 EUR, or 1,000 EUR) each additional year. Source: Danguy and van Pottelsberghe (2014))

life. Figure 2 illustrates four alternatives for the former structure – the sum of fees of the first 2 (4, 8, or 12) countries – and four alternatives for the latter one. The UP(200) fee schedule could be considered as the most appropriate because it is simpler than the additive fee structures, and it corresponds to what the business sector is currently paying. Indeed, van Pottelsberghe and van Zeebroeck (2008) showed that the average geographical scope of protection is about four countries – UP(4) is similar to UP(200) – for the patents granted 15 years ago by the EPO. With UP(200), an applicant would pay cumulated renewal fees of about 15,000 EUR to keep its patent enforced for 20 years in the 25 EU member states. This absolute cost is affordable in relative terms, given the large geographical scope of protection provided by the Unitary Patent. The EUR 30,000 figure is similar to what a patent holder must be ready to pay in the current European patent system for less than 10 years of protection in 13 countries.

But the important question from the patent offices' viewpoint is whether the Unitary Patent would generate the same resources with its central renewal fees for the EPO and for all NPOs.

Danguy and van Pottelsberghe (2014) provide some answer, by simulating the net present value of renewal fee income for the average patents granted by the EPO, within the current system and within a two-layers system. Figure 3 shows the sensitivity of these simulations to different Unitary Patent fee schedules and to the expected share of Unitary Patents in total EPO-granted patents. Two observations might be drawn from these simulations. First, the higher the level of Unitary Patent renewal fees, the higher the total renewal fee income per average patent granted by the EPO, independently from the share of Unitary Patent. In other words there is a natural temptation to set high renewal fees. Second, with low Unitary Patent renewal fees, there is a U-shaped relationship between the share of UP and the total renewal fee income per average patent. This is due to the substitution effect between the two types of patent. The total renewal fee income collected by patent offices could actually be lower than the income in the current situation if the unitary patent system is not attractive enough (i.e., low share of UP) and has low renewal fees. At first glance these simulation results strengthen the argument for high UP renewal fees, so that they generate more



European Patent System, Fig. 3 Simulated total renewal fee income per average patent (Note: the bars represent the net present value of a patent granted by the

EPO and which would opt for the Unitary Patent route, taking into account its expected maintenance rate. Source: Danguy and van Pottelsberghe (2014))

income than under the current situation. However, very high fees would lead to a low use of the new unitary patent.

But even with aggregate renewal fee income stable or larger than the current system, patent offices logically worry about the share of this income that will be allocated to them. Danguy and van Pottelsberghe (2014) have simulated the revenue stream of national patent offices under the dual system. Their results show that only a handful of NPOs could be negatively impacted by the creation of the unitary patent system. Except for Germany, the budgetary losses for these patent offices are very low and only occur with relatively low shares of Unitary Patents in total EPO-granted patents. Improving the revenue prospects could also be achieved if the number of patents granted by the EPO increase and if national renewal fees would increase.

Which Design for the Unified Patent Court?

The second medium-term challenge relates to the design of the forthcoming Unified Patent Court (UPC), aiming at proceeding unitary patent litigations in Europe. In order to come into effect, the UPC Agreement must be ratified by at least 13 member states (including Germany, France, and the UK). A logical consequence is that the

Unitary Patent will cover only those member states where the UPC Agreement is in force. There is therefore a high probability that the Unitary Patent starts with much less than 25 member states, because converging to a common patent court denominator might prove to be complex and lengthy.

Indeed, the current national litigation systems are highly heterogeneous, reflecting important differences across national jurisdictions. For instance, some countries have technically qualified judges and others not, and the wage of judges is particularly high in the UK. Total litigation costs vary significantly across countries. The UK is by far the most expensive jurisdiction, with costs that are nearly as high as the cumulated costs in France, Germany, and the Netherlands. In case of multiple parallel litigations across jurisdictions, cumulated costs vary from 310 thousands euros before the four courts of first instance up to 3.6 million euros when taking account of the cost of appeal at second instance. Multiple litigations are particularly prohibitive in Europe, especially for individuals and SMEs, and can be more than twice as high as in the USA (see Mejer and van Pottelsberghe 2012). The challenges here will be to find a proper balance and a design that rally a majority of national litigation systems.

It is the responsibility of the Contracting States to set up the Unified Patent Court. A Preparatory Committee has been created in September 2013 to set out a road map for the establishment and coming into operation of the court. It has five main working areas, including the legal dimension, the financial sustainability, the human resource and training components, the design of the information system, and the facilities. It seems highly probable that there will be a central division of the Unified Patent Courts, with three Divisions, specialized in specific technological areas, and based in Paris, London, and Munich. Then, there will be regional divisions and local divisions, covering one to several countries, depending on the geographical areas and country size. The envisaged local divisions and their working language between parentheses would be based in “England and Wales” (English), the Netherlands (Dutch and English), France (French), Germany (German), and Belgium (Dutch, French, German, English). Envisaged regional divisions include one for Romania, Bulgaria, Cyprus, and Greece (all official languages plus French and English); one for Nordic countries (Denmark, Sweden, Finland, Estonia, Latvia, and Lithuania, with English as the main language); and one for Hungary, Czech Republic, and Slovakia (language undecided as of January 2015).

A Drafting Committee, composed of highly experienced judges and lawyers, is in charge of preparing the final draft. As of January 2015, the 17th version has been released (up-to-date information can be found on www.unified-patent-court.org). Areas of concern include a potential transitional regime, especially regarding an opt-out (from the Unitary Patent) provision; the possibility or not to bifurcate; the working languages at various divisional levels; the training, qualifications, and experience of judges; and court fees, recoverable costs, and numerous further procedural details. If bifurcation is allowed, there will be four different ways to enforce patent protection in Europe (Hilty et al. 2012):

1. Patents granted by national patent offices and enforced through national courts
2. European Patents granted by the EPO and enforced through the Unified Patent Court system
3. European Patents granted by the EPO and enforced through national courts
4. Unitary Patents granted by the EPO and enforced through the Unified Patent Court system

It will be possible to switch from option 3 to option 2 as long as the patent has not been subject to litigation in a national court. This new enforcement system and its flexibility will pave the way for more strategic behavior by patent owners. Unitary Patents might be allowed to bifurcate toward European Patents and hence might actually not be enforced only through the UPC. This bifurcation issue is, as of January 2015, still subject to change.

Longer-Term Challenge: Make It Work. . .

The European patent system which is being envisaged by policy-makers will include three layers of patents (national, European, and Unitary) and four enforcement mechanisms. Its cost in terms of renewal fees could be particularly high, not to mention the litigation costs and the complexity of the whole system. True, the Unitary Patent Package is in itself less complex and expensive in relative terms (i.e., per capita or per market unit), but it is being built on top of a two-layer system.

One can hardly disagree with the fact that this system, in its current format, will probably not better fulfill its ultimate objective of stimulating innovation in Europe: three layers make it more complex, and it might end up being quite expensive. This three-layer system is to be compared with a less expensive one-layer system in the rest of the world. But this is certainly not a reason to discontinue the Unitary Patent and Unified Patent Court projects. They are crucial for the construction of the European patent system, and much political power or courage will be needed to

achieve a coherent and effective system. In order to build an effective and truly European patent system, the following milestone will have to be considered:

1. National patent offices should stop granting patents on their own, which does not preclude playing an important role in their national innovation system, to continue to receive national priority filings and to perform search for prior art and preliminary assessments.
2. The European Patent should be void, and a phasing-out agenda should be established.
3. Small and medium entities should have substantially lower fees and litigation costs.

References

- Danguy J, van Pottelsberghe de la Potterie B (2011a) Cost-benefit analysis of the community patent. *J Benefit Cost Anal* (Berkeley Electronic Press) 2(2):1–41
- Danguy J, van Pottelsberghe de la Potterie B (2011b) Patent fees for a sustainable EU patent system. *World Patent Inf* 33(3):240–247
- Danguy J, van Pottelsberghe de la Potterie B (2014) The policy dilemma of the unitary patent. Bruegel working paper 2014/13, 27 Nov 2014
- de Rassenfosse G, van Pottelsberghe de la Potterie B (2013) The role of fees in patent systems: theory and evidence. *J Econ Surv* 27(4):696–716
- Guellec D, van Pottelsberghe de la Potterie B (2007) The economics of the European patent system. Oxford University Press, Oxford, 250 p
- Hilty R, Jaeger T, Lamping M, Ulrich H (2012) The unitary patent package: twelve reasons for concerns. Max Planck Institute for Intellectual property and competition law research paper, 17 Oct 2012
- Mejer M, van Pottelsberghe de la Potterie B (2011) Patent backlogs at USPTO and EPO: Systemic failure vs deliberate delays. *World Patent Information*, 33 (2):122–127
- Mejer M, van Pottelsberghe de la Potterie B (2012) Economic incongruities in the European patent system. *Eur J Law Econ* 34(1):215–234
- van Pottelsberghe de la Potterie B (2010) Europe should stop taxing innovation. Bruegel policy brief, 2010/2, 8 p
- Van Pottelsberghe B, van Zeebroeck N (2008) A brief history of space and time: the scope-year index as a patent value indicator based on families and renewals. *Scientometrics* 75(2):319–338

European Union Anti-cartel Policy

J. M. Ordóñez-de-Haro¹, J. R. Borrell^{2,3} and J. L. Jiménez⁴

¹Departamento de Teoría e Historia Económica y Cátedra de Política de Competencia, Universidad de Málaga, Málaga, Spain

²Secció de Polítiques Públiques, Dep. d'Econometria, Estadística i Economia Aplicada, Grup de Govern i Mercats (GiM), Institut d'Economia Aplicada (IREA), Universitat de Barcelona, Barcelona, Spain

³IESE Business School, Public-Private Sector Research Center, University of Navarra, Navarra, Spain

⁴Facultad de Economía, Empresa y Turismo, Departamento de Análisis Económico Aplicado. Despacho D.2-12, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Abstract

The fight against cartels started in European Union since its foundation in 1957 and the passing of the Regulation empowering the Commission to enforce competition rules since 1962. However, anti-cartel policy was very ineffective in its early years from 1962 to 1980. It also had a lot of enforcement problems to uncover cartels until 1995. It was in 1996, when the leniency program was set up, when it truly became an increasingly effective policy. The leniency program is a mechanism by which infringing firms that have been active in a cartel can obtain fine reductions by providing hard evidence to the Commission about the existence and functioning of any cartel. The improving of the leniency program in 2002 and 2006, and the adoption of tougher fining policies and settlement procedure, has made cartel-busting policy much more effective in the last decade, but there is still much uncertainty to what extent anti-cartel policy will keep this trend of being more effective in the future.

EU Anti-cartel policy stages

The fight against cartels started in European Union (EU) since its foundation in 1957 and the passing of the Regulation empowering the Commission to enforce competition rules since 1962. Though in the last decade a clear advance in the detection, destabilization, and fining of cartels operating in the European Union has been noticed, the first years of anti-cartel policy were flawed. In fact, as Ordóñez de Haro et al. (2016) argue, four stages in EU anti-cartel policy can be highlighted.

First Stage (1957–1980)

The origin of the EU competition policy is found in the relevant provisions of the 1957 Treaty of Rome. The Article 85 of this Treaty (now Article 101 of the Treaty on the Functioning of the European Union) bans agreements among firms that restrain competition in the internal market.

The actual enforcement of these articles 85 and 86 became effective as of 21 February 1962, with the entry into force of the Council Regulation n° 17/62. This Regulation gave the Commission a central role as the authority charged with enforcing those articles, recognizing its power to open investigations, to adopt decisions and impose appropriate sanctions and remedies for the competition rules' infringements (McGowan 2009; Carree et al. 2010).

The enforcement system was based on the Commission's centralized control of the application of Article 85 and the requirement of prior notification by the parties of their agreements, decisions, and practices to the Commission which after examination of the notification could authorize them based on the proper exemptions contained in paragraph 3 of Article 85. This system caused serious delays in the procedural treatment and completion of the files and the consequent backlog of cases, since the Commission devoted a large proportion of its resources to deal with notifications.

To address this problem, the Commission tried to reduce the number of cases and to speed up the decision-making procedure by undertaking several initiatives, including the adoption and application

of several block exemption regulations, the use of the so-called comfort letters, or the introduction of notices on agreements of minor importance which do not have sufficient impact on competition.

At the end of this stage, in its decision in case *IV/29.595 – Pioneer Hi-Fi Equipment*, the Commission expressed its determination to shift its fining policy toward harsher sanctions for competition law infringements (Geradin and Henry 2005).

Seven cartels were sanctioned in this early stage of the European anti-cartel policy, totaling 55.53 million constant 2010 euro in fines (Ordóñez de Haro et al. 2016). The Commission just completed 0.4 sanctioned cases per year since 1962–1980 and imposed just fines of 600,000 euros per participating parent firm on average (constant 2010).

Second Stage (1981–1995)

It was a transitional stage during which the first consequences of that shift in Community fining policy were seen in the general level of fines imposed on cartels. The Commission reiterated its intention to move towards a tougher sanctioning policy in its Thirteenth Report on Competition Policy (1984).

Although there were no relevant legislative or institutional developments during this period, in December 1995, the European Commission published a draft notice concerning the non-imposition or mitigation of fines in cartel cases where undertakings cooperate in the preliminary investigation or proceedings in respect to an infringement. The subsequent adoption of this notice in 1996 represents a major milestone in EU competition law enforcement against cartels.

Over these 15 years, the Commission sanctioned 32 cartel cases. The total amount of fines exceeded 1122 million euro (constant 2010) in this period. The Commission just completed 2.1 sanctioned cases per year in this period and imposed fines of 4 million euro per participating parent firm on average (constant 2010).

Third Stage (1996–2005)

The introduction of the first leniency program in 1996 marks the start of a third stage which covers

all the steps taken toward the modernization of the EU competition policy up to 2005.

The first Community leniency program started on 18 July 1996. It was inspired by the US program in force since 1993, sought to encourage the breakdown of the “code of silence” among the members of the cartels and to become a successful tool that would significantly increase the effectiveness of competition policy (see Borrell et al. 2014).

Although this first leniency system is characterized by introducing some of the basic principles that would guide subsequent versions of the program, it lacked many enforcement details that were fixed in the 2002 and 2006 reforms of the leniency program.

Another significant change in this period was the publication by the Commission in January 1998 of its “*Guidelines on the method of setting fines imposed pursuant to Article 15 (2) of Regulation No 17 and Article 65 (5) of the ECSC Treaty*.” The European Commission, for the first time, provided a clear method for determining the final amount of a fine explaining the successive steps and criteria that would be taken in order to obtain that amount.

The creation of the first “anti-cartel unit” within DG IV (directorate later known as DG Competition), in December 1998, represented an important reinforcement of the particular fight against cartels. The progressive increase in resources culminated in the creation of a second “anti-cartel unit” in 2002. Nevertheless, the two anti-cartel units ceased to exist as they were conceived as a result of an internal reorganization of the DG Competition in July 2003. From then on, the resources and staff to combat cartels were then allocated to different units in several key sector directorates (Lowe 2008).

The numerous criticisms of the manifest deficiencies in the 1996 *Leniency Notice* yield to the publication, on 19 February 2002, of the new 2002 *Leniency Notice* which brought about significant changes to the procedure and specified requirements to benefit from the program (Arp and Swaak 2003; Borrell et al. 2015).

During the period covered by this stage, the provisions contained in the Council Regulation n°

17/1962 were valid until May 2004 when the *Council Regulation (EC) No 1/2003 of 16 December 2002 on the implementation of the rules on competition laid down in Articles 81 and 82 of the Treaty* entered into force. The most significant changes introduced by the Regulation 1/2003 mean the simplification of the administrative procedures and decentralization of the application of the competition rules in the EU. This reform implies that the system of notification and authorization was completely abandoned to be replaced by a directly applicable exception system. The Commission no longer had to deal with notifications from firms taking part in agreements, but the firms were directly liable of their conducts before the Commission, the National Antitrust Authorities, the European Courts, and the National Courts if they infringe the provisions of the Treaty reducing the Commission’s workload significantly (McGowan 2005).

From 1 June 2005 on, in response to the greater emphasis on combating cartels, a new Directorate in the DG Competition devoted exclusively to the fight against cartels became operational. The Cartels Directorate is responsible for carrying most of the cartel cases and plays a leading role, in close cooperation with the Directorate for Policy and Strategic Support in developing the policy to apply in the cartels arena.

Additionally, in 2005, a two-stage procedure is introduced by the DG Competition. So, from then on, all antitrust cases start with a first phase of investigation after which the Commission adopts a decision concerning the theory of identified harm and whether there is reason to pursue the case as a matter of priority. If the case is considered a priority, the Commission takes a decision in principle to initiate proceedings and then carries out a thorough investigation. This procedure is intended to spend less time and resources on cases that do not merit special attention.

To conclude this third stage, we should note the first initiative which tried to identify the main obstacles faced by the operation of a more efficient system of claims for damages which resulted from breaches of European Union competition laws: on 19 December 2005, the Commission

published the *Green Paper – Damages actions for breach of the EC antitrust rules* which includes the proposal of a series of measures aimed at encouraging victims of infringements of Articles 81 and 82 of the Treaty to exercise their right to claim for damages before Member States' courts (Pheasant 2006). Consequently, this initiative is designed to boost the interaction between public and private enforcement of antitrust rules (Diemer 2006).

In this period, a total of 44 cartel cases were sanctioned, 46% of them uncovered by a leniency application. Total fines imposed reach 5491.74 million constant euro. The Commission completed 4.4 sanctioned cases per year in this period (more than doubling the number of the previous stage) and imposed fines of 22 million constant euro per participating parent firm on average, multiplying the average fine by more than five times with respect the previous period.

Fourth Stage (2006–)

The last stage began in 2006, with several legislative reforms that aim to improve and consolidate those tools which proved to be valuable for fighting cartels, such as the sanctioning mechanisms, as well as the framework for cooperation between cartel participants and the Commission, in order to streamline procedures and to enhance the transparency and predictability.

The first of these improvements took place with the publication, on 1 September 2006, of the *2006 Fine Setting Guidelines*, which the Commission is currently putting in place, and they introduce some important new points (Barbier de La Serre and Lagathu 2013).

On 8 December 2006, the *2002 Leniency Notice* was replaced by the new *2006 Leniency Notice* (OJ C 298/17. 8.12.2006.). This is the second reform of the leniency program and aimed to provide for more clarity and transparency in its requirements and how to proceed in this program, as well as make it more attractive to potential cooperators (Sandhu 2007; Borrell et al. 2015).

In early July 2008, the Commission adopted the so-called Settlements Package. It consists of two documents: the *Commission Regulation*

(EC) No 622/2008 of 30 June 2008 amending Regulation (EC) No 773/2004, as regards the conduct of settlement procedures in cartel cases and the *Commission Notice on the conduct of settlement procedures in view of the adoption of Decisions pursuant to Article 7 and Article 23 of Council Regulation (EC) No 1/2003 in cartel cases*.

The *Settlement Notice* details the settlement procedure, interprets the new provisions contained in the *Settlement Regulation*, and provides further guidance on the first steps of the settlement procedure, the settlement discussions, the statement of objections, and the adoption of the final decision in such a procedure.

This package enables the Commission and parties to proceedings to follow a more simplified and simpler procedure when cartel participants, having seen the evidence in the Commission file, acknowledge their involvement in the cartel and their liability for the infringement (Mehta and Tierno 2008).

Finally, it is essential to stress the possible consequences for the fight against cartels that the adoption of the *Damages Directive* has and could have. Member States have until 27 December 2016 to transpose the provisions of the Directive into their legal systems (*Directive 2014/104/EU of the European Parliament and the Council of 26 November 2014 on certain rules governing actions for damages under national law for infringements of the competition law provisions of the Member States and of the European Union*). It remains to be seen whether the setup of an EU wide common framework for claiming private damages will make this policy even more effective in the near future.

The Commission's sanctioning activity during this stage was the most prolific of all with 53 cartel decisions until the end of 2014. Total fines imposed reach 17,739.87 million euro constant 2010. The Commission completed 5.9 sanctioned cases per year in this period and imposed fines of 62 million euro per participating parent firm on average (constant 2010), almost multiplying the average fine by three times with respect the previous period.

References

- Arp DJ, Swaak CR (2003) A tempting offer: immunity from fines for cartel conduct under the European Commission's new leniency notice. *Eur Compet Law Rev* 24(1):9–18
- Barbier de La Serre E, Lagathu E (2013) The law on fines imposed in EU competition proceedings: faster, higher, harsher. *J Eur Compet Law Pract* 4(4): 325–344
- Borrell JR, Jiménez JL, García C (2014) Evaluating anti-trust leniency programs. *J Compet Law Econ* 10(1):107–136
- Borrell JR, Jiménez JL, Ordóñez de Haro JM (2015) The leniency programme: obstacles on the way to collude. *J Antitrust Enfor* 3(1):149–172
- Carree M, Günster A, Schinkel MP (2010) European anti-trust policy 1957–2004: an analysis of commission decisions. *Rev Ind Organ* 36:97–131
- Diemer C (2006) The green paper on damages actions for breach of the EC antitrust rules. *Eur Compet Law Rev* 27(3):309–316
- Geradin D, Henry D (2005) The EC fining policy for violations of competition law: an empirical review of the commission decisional practice and the community courts' judgments. *Eur Compet J* 1(2): 401–473
- Lowe P (2008) The design of competition policy institutions for the 21st century—the experience of the European Commission and DG Competition. *Compet Policy Newsl* 2008-3:1–11
- McGowan L (2005) Europeanization unleashed and rebounding: assessing the modernization of EU cartel policy. *J Eur Publ Policy* 12(6):986–1004
- McGowan L (2009) Any nearer to victory in the 50-year war? Assessing the European Commission's leadership, weapons and strategies towards combating cartels. *Perspect Eur Polit Soc* 10(3): 283–307
- Mehta K, Tierno ML (2008) Settlement procedure in EU cartel cases. *Compet Law Int* 4:11–16
- Ordóñez de Haro JM, Borrell JR, Jiménez JL (2016) European Commission's fight against cartels (1962–2014): a retrospective and forensic analysis. Mimeo
- Pheasant J (2006) Damages actions for breach of the EC antitrust rules: the European Commission's green paper. *Eur Compet Law Rev* 27(7):365–381
- Sandhu JS (2007) The European Commission's leniency policy: a success? *Eur Compet Law Rev* 28(3): 148–157

European Union Law

► [European Community Law](#)

Europeanization

► [European Nationality](#)

Experimental Law and Economics

Sven Hoeppe

Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium

Guest Researcher, Max Planck Institute for Research on Collective Goods, Bonn, Germany

Abstract

Experimental law and economics is the newest methodological development in law and economics research. Yet a lot of researchers – old and young – are not familiar with the foundations of the method. This may lead to skepticism. In this entry, I elaborate on the purpose of experiments and introduce building blocks of the method that tend to distinguish experimental law and economics from experimental methods in other disciplines. Moreover, I shortly discuss concerns about the external validity of the results obtained in a laboratory experiment. Finally, I introduce and invalidate the most common criticisms against experiments that most often advanced by scholars unfamiliar with the method and the discipline. As knowledge about the method's foundations is further spread, experimental law and economics will substantially contribute advancing the research frontier of the economic theory of law.

Definition

Experimental law and economics exposes theoretical research in law and in law and economics to methods from experimental economics, thereby testing existing theory, investigating observed

deviations from theory, establishing empirical regularities as the foundation of new theoretical constructs, and informing policy-making.

first reading for an audience interested – but yet unfamiliar – with experimental approaches to law (and economics).

Introduction

Not every lawyer, legal policy-maker, judge, or legal scholar conceives legal rules as a governance instrument that can be employed to coordinate and steer individual behavior in society. For those who do adopt such an *ex ante* perspective, however, it is of utmost importance to hypothesize about and investigate how law and individual human behavior are interrelated. To pierce the veil of ambivalent plausibility arguments and conflicting theories, it is even more important to understand what kind of behavior may be caused by alternative legal interventions.

Yet, legal scholars are often surprised and become quite reserved when they learn that a colleague is not extensively engaged in doctrinal legal scholarship, but rather adopts a social science approach. Caution usually rises to suspicion when the straying colleague conducts experiments, which seem rather alien to legal scholarship. But experiments can, *inter alia*, help to understand whether policy proposals – developed through, e.g., doctrinal legal scholarship – will function as intended.

Actually, experiments are not a novel method in legal scholarship. They have long been conducted in the law and psychology as well as in the criminology branches of legal research (Engel 2013a, b). Today, “[t]he only true novelty is the advent of experimental law and economics” (Engel 2013a, p. 7).

This entry is not written as a review of studies using an experimental approach to law and economics. Excellent reviews with a cornucopia of examples exist (cf. Croson 2002; Arlen and Talley 2008; Lawless et al. 2010). Rather, it is written with the intention to confront all too common misunderstanding about the experimental approach to law and to law and economics with information about the usefulness of experimental methodology for legal scholarship as well as policy-making. This entry should provide a good

Purpose

Generally speaking, in both disciplines, economics and law, there is a strong emphasis on deductive reasoning. Researchers start with a set of axioms, use rules of logic to manipulate them, and subsequently derive conclusions about their research question. For instance, in (neoclassical) economics, certain assumptions about individual behavior are used to mathematically model the theoretical construct “market” and to draw implications for the allocation of resources. As an example for deductive reasoning in law, one core assumption is that the law makes sense (*ratio legis*) and a lot of doctrinal legal scholarship is devoted to establish this inner logic of the law by means of interpretation and, furthermore, to match real-world facts to these abstract norms by analytical comparison (subsumption). Thus, doctrinal legal scholarship can conclude that legal consequences from an abstract, model-like norm are applicable to real-world facts.

A consequence of this heavyweight deductive reasoning is that theories become so complex and refined that simply observing natural data does often not address the pressing theoretical questions. Empirical tests of those theories are therefore significantly more difficult to perform. Experiments are one way to meet such a challenge (Croson 2002).

The more conducive purpose of conducting experiments in law, in general, and law and economics, in particular, is to exploit three advantages of the methodology. The first, and main advantage of experiments, lies in their possibility to solve the identification problem – the problem that arises when there may be bidirectional influence between dependent and independent variables that makes it difficult to separate cause and effect. For example, consider crime levels and the number of police: Does an increase in the level of crime cause a demand for more police, or do

fewer police cause more crime? Identification is notoriously difficult with field data. To establish causality outside of an experiment, at least three steps are necessary (Lawless et al. 2010). First, there needs to be a reliable association – or strong correlation – between the two variables. Second, one needs to credibly show that changes in the variable hypothesized to be the cause precede the supposed effect variable. Third and most problematic, other rival explanations for this observation (confounds) need to be ruled out, which also concerns the rather big problem of omitted-variable bias, where a driving variable is missing from the statistical model used for analysis. As a consequence, the effect of variables that enter the analysis is over- or underestimated.

To the end of separating cause and effect, a laboratory experiment generates an artificial environment in which participants are randomly assigned to different treatments. Random assignment facilitates control for unobservable individual differences among the participants. If carefully designed, the treatments in the experiment differ only by one dimension. Accordingly, if a treatment effect can be observed, it is a necessary result from the manipulation. With their distinguishing features – manipulation of one or more variables of interest, random assignment, and experimental control of other variables – experiments are specifically designed to test cause-effect relationships (Lawless et al. 2010). The very large degree of control can be used to generate a laboratory environment that tests theories and separates alternate theories that may not be separable with naturally occurring data (cf. Croson 2002).

The second main upside of laboratory experiments is their replicability. Because laboratory experiments are replicable, other researchers can reproduce the experiment and verify the findings independently. A stream of independent replications can test the robustness of findings. This is very difficult with field data because the environment is constantly changing such that each observation or dataset is just a snapshot of a certain combination of conditions.

Especially for legal scholars, experimentation brings about a third important benefit that is likely

to be equally important: experiments enable the study of problems that are hard or even impossible to observe in the field (Lawless et al. 2010), i.e., even if data are already available from other sources. How would addressees of a legal rule act if the existing rule would be or would have been replaced by an alternative rule? And what are the behavioral effects of different alternative rules that are under consideration for being implemented? Many a fact that is relevant for doctrinal lawyers and legal policy-makers are counterfactuals and never enter into field data (Engel 2013a). Experimenters can observe the unobservable (Falk and Heckman 2009).

Building Blocks

The basic logic of an experiment is to create different treatment groups (or experimental conditions) by manipulating one or more variables of interest, but always only one at a time, while every other variable is experimentally controlled for. The treatment without any variable change serves as the control group. This experimental manipulation creates desired differences between the groups, which enables the isolation of the different variables of interest, reduces confound, and ultimately facilitates identification of cause-and-effect relations. Participants are randomly assigned to these treatment groups. Randomization may occur for a lot of aspects in a specific experimental design. With regard to randomly assigning participants to experimental condition, random assignment, e.g., assures the initial comparability of the different experimental groups by equalizing the distribution of non-manipulated participant characteristics, such as age, income, gender, and education, across the experimental conditions. While participants are then being exposed to the experimental stimuli, their decision behavior is measured. Afterward, the different measurements can be compared across treatment groups. Because the only difference between the groups is their exposure to the treatment, any measured difference between the treatment and control groups can be attributed to the treatment.

Beyond this general categorization, experimental methodologies and their best practices – even within the social sciences – are quite different. Because of different underlying theories, experiments in a law and economics paradigm differ in their methodology from the earlier experimental approaches to law in psychology and criminology. Experimental law and economics is based on the methodology of experimental economics (cf. Hoffman and Spitzer 1985; Davis and Holt 1993; Friedman and Sunder 1994; Kagel and Roth 1995).

Although experimental economics pushes more and more to the field (Levitt and List 2007, 2009), the standard tool in experimental economics is the laboratory experiment. The lab offers the possibility to effectively employ controlled variation at relatively low cost (Hoffman and Spitzer 1985; Friedman and Sunder 1994). Decision environments can be controlled in ways that are very difficult to duplicate in the field. In the lab, the experimenter knows and controls participants' material payoffs, the order in which the different participants can act and interact, and – maybe most importantly – the informational situation for each single choice.

In experimental economics, one can generally differentiate four types of experiments for specific purposes: some experiments test theory, others investigate observed deviations from theory (anomalies), again others establish empirical regularities as the foundation of new theories, and finally experiments inform policy-making or test-bed new policy proposals (cf. Hoffman and Spitzer 1985; Roth 1986; Smith 1994). Each of these types faces particular methodological challenges (cf. Croson 2002).

The decisive difference between experimental economics – and by extension experimental law and economics – and the other experimental approaches to law involves a number of dimensions (cf. Davis and Holt 1993; Friedman and Sunder 1994; Engel 2013b).

1. Participants are exposed to a specific incentive structure that matches the payoffs in the theory underlying the experiment. The choices the participants make in the experiment determine

their compensation. This ensures that participants seriously contemplate their decision because value is induced (Smith 1976; Friedman and Sunder 1994). This is the main reason why experimenters in economics trust that “behavior in the laboratory is reliable and real: Participants in the lab are human beings who perceive their behavior as relevant, experience real emotions, and take decisions with real economic consequences” (Falk and Heckman 2009, p. 536).

2. Since real money is at stake, participants do not face a hypothetical, but rather a real situation. Note that in some cases and contradictory to established experimental economics, it may make a lot of sense to not incentivize a participant. This is the case, for instance, if the participant mimics the role of a judge who has, ideally, no monetary incentive in the single decision in real life (cf. Engel and Kurschilgen 2011).
3. The experiment is usually free from context. Economic experiments are very abstract to facilitate identification (cf. Ariely and Norton 2007). Firstly, context may add variance to the data and blur otherwise clear results. More worrisome, context may create systematically biased results or demand effects, i.e., when the participant tries to act according to the perceived expectations of the experimenter. As a result, the generated dataset would not be reliable. Lastly, avoiding context is, of course, a requisite when the underlying theories are also free from contextual cues. Note that context, however, may be a treatment variable as well. If so, the manipulation of context is the premier goal of the experiment.
4. Another distinguishing feature is that experimental economists generally take great care to not deceive their subjects. This is different in psychology experiments where deception is commonplace, although not undisputed (cf. Hertwig and Ortmann 2008). In experimental economics, the validity of the results depends on the link between expected payoffs and individual behavior. If participants are deceived about variables establishing this link, the validity of their decisions is

questionable. Moreover, experimental economists usually avoid tainting the participant pool. They are “concerned about developing and maintaining a reputation among the student population for honesty in order to ensure that subject actions are motivated by the induced monetary rewards rather than by psychological reactions to suspected manipulation” (Davis and Holt 1993, pp. 23–24). Admittedly, however, there is an intense, contemporary discussion about whether the costs of deception in experimental economics exceed the benefits (cf. Bonetti 1998; Hertwig and Ortmann 2008; Jamison et al. 2008; Alberti and Güth 2013; Krawczyk 2013).

5. Economic experiments are oftentimes interactive in nature. The reason for interaction between participants is straightforward. Economics experiments mostly employ some form of social dilemma game as experimental workhorse, and theoretical predictions focus on the results of social interaction.
6. Usually, experimenters in economics take great effort to ensure complete anonymity. As with all the other features, this is done with the aim to facilitate identification, i.e., to avoid confounds. To provide a counterexample, the experimentalist may not care about absolute anonymity, if the research question addresses reputation effects.
7. Often, the trials in the experiment are repeated to allow for learning and to study interactive dynamics.

These features facilitate what an experiment is ultimately all about: complete control. Only in an environment with extensive control can the claim become credible that an observed difference in behavior is indeed caused by the treatment manipulation and not merely the result of correlation.

Caution

However promising and rewarding the clear separation of cause and effect, experimenters face validity concerns. In general, the concept of validity refers to the extent to which an empirical study

produces accurate and credible data. Validity is assessed on three dimensions. First, internal validity refers to the extent to which the research design allows making inferences about the relationship of different variables. Internal validity involves that the experimenter must ensure that the choices faced by the control and treatment groups differ in the ways the experimenter hypothesizes, but do not differ in some other ways. This implies accounting for alternative explanations and omitted variables. Second, external validity speaks to the degree to which the research results can be generalized to a population beyond the particular study, i.e., to different people, different settings, different times, and also different measures. Third, construct validity refers to the extent to which the measures used to observe certain variables sufficiently capture the construct that the empiricist wants to study (cf. Lawless et al. 2010).

In order to isolate variables and to identify cause-effect relationships, experimentalists necessarily trade off external validity for internal validity. The artificial environment in a laboratory experiment allows the experimenter to make clean-cut inferences, but raises questions about its congruence with the social phenomenon that is to be studied. By design, what is measured in a laboratory experiment is only analogous to what the experimenter wants to understand (Smith 2010; Engel 2013b). Just because a hypothesis is falsified in the lab does not necessarily suggest that it will also be falsified in the field because there are so many intervening variables that cannot be controlled for. However, Plott (1982) counters concerns about external validity by pointing out that a theory will likely not predict outcomes in the real world if it does not (at least) predict outcomes in an idealized, controlled laboratory environment. Falk and Heckman (2009) argue that external validity problems are not a problem unique to experimental methodology.

Especially when legal scholars are on the hunt for cause-and-effect relationships between legal rules and individual behavior, they pay an increased price because they tend to distance themselves even more from the traditional doctrinal legal discourse (Engel 2013a).

Criticism and Concern

Particularly when scholars or members of another audience are unfamiliar with experimental work in general and the specific methodology in particular, a standard set of criticisms is often held against the experimental results.

Probably the most notorious objection involves limited external validity (see above). The gist of this critique is that the experiment cannot tell anything about the real world because the decisions are made in an artificial environment. As mentioned above, this is indeed reason to be very cautious to interpret experimental findings. However, this typical reason to disapprove of experimental results is misguided if the experiment satisfies the assumptions of the tested theory. Theoretical predictions rest on the assumptions of that theory irrespective of where the assumptions are met – in the lab or in the field (Plott 1982). The objection about limited external validity may have more grips, when the experiment concerns legal policy-making. In this context, an isolated cause-effect relationship matters less than entire institutions where certainly many of those relationships work with or against each other. Nevertheless, this implies the need to understand complex institutions by disentangling the muddle of causes and effects. Alternatively, legal experimenters may implement sequential designs that investigate behavior before and after the implementation of the entire legal institution (Engel 2013b). Lastly, empirical scholars know that likely none of the available methods will ever fully address reality. Falk and Heckman (2009) emphasize that it is not so obvious whether laboratory data or the field data are more informative. The scientific value of a method ultimately depends on the underlying research question. The empirical methodology notwithstanding, “eventually, reality is too complex, too little orderly, to be studied in an objective way” (Engel 2013a, p. 18). The concern about external validity certainly puts another emphasis on the complementary nature of different empirical methods, but it provides no grounds to completely dismiss one of them.

A second standard objection against experimental results concerns the participant pool.

Novices in experimental methodology often criticize the fact that experiments in law and economics more often than not rely on university students. They argue that students are certainly very different from the population in the real world and that they make decisions differently from professional decision-makers outside the lab. Fortunately, experimental research on this subject is on the rise. Oftentimes, there is no discernible difference in decision-making between students and professionals. If there is, however, students sometimes are more successful than professionals – although not always. There are a number of reasons why this objection seems to have a weak foundation (cf. Croson 2002; Falk and Heckman 2009). A student sample is as representative as any other sample, as long as it does not consist of only students with the same academic background, if this is part of the manipulation (cf. Frank et al. 1993, 1996). Most importantly, most often, the experiment is not reliant on any specific knowledge or experience. By contrast, there are also risks in bringing professionals to the lab. For instance, they may bring in incentives, knowledge, and experience that do not exist in the experimental design and therefore systematically bias results. Nevertheless, it is surprising that people often do not complain about seemingly weak participant pools in other areas of science. This is nicely illustrated with a not so ironic insight from Gary E. Bolton that he shared with me and others in July 2013: “In experimental economics, we typically begin our studies with students, pretty much for the same reasons medical experimenters begin with mice. First, they work cheap. Second, we can do what we want with them (within institutional rules). And third, they are genetically similar to real human beings”.

Another prominent critique of experimental work in law and economics is based on the size of the payoffs in the experiment. It is often argued that the size of payoffs is too small to induce meaningful behavior. However, experiments have been conducted with low and high payoffs and, in fact, have even tested differences between low and high stakes in experimental settings. All in all, there are very few differences on average behavior when the size of the stakes varies (Smith

and Walker 1993; Beattie and Loomes 1997; Camerer and Hogarth 1999).

Taken together, these common objections can be summarized as the “lack of realism” challenge. This notion, however, is based on an implicit assumption about the hierarchy of how relevant data are generated, with field data being deemed superior to data from the laboratory. Ultimately, however, the issue of realism is not a question of laboratory versus field data. The real issue is determining the best way to identify the causal effects in question. In this context, it is also important to stress that empirical methods are complements to each other (cf. Arlen and Talley 2008), not substitutes, as the discussion about the “lack of realism” suggests. Paradoxically, however, many of these objections also suggest conducting more experiments instead of fewer (Falk and Heckman 2009), because the critique itself is often based on a variable that can be manipulated in a new set of experiments to test the critique. This is, for instance, true for the aforementioned criticism that stakes are trivial and that participants are inexperienced students. However, this is not to suggest that experimentalists should not take these common concerns seriously. On the contrary, experimentalists should be their own most austere critics in order to prevent automatic, arbitrary experimental design that would threaten the credibility of experimental results.

In addition to these standard objections to the experimental method, Engel (2013a) specifically discusses sets of concerns prevalent in the legal academic arena. He distinguishes the rather diffuse concerns into five common categories of specific objections from legal scholars. To traditional lawyers, the experimental method appears to be (1) too scientific, (2) too individualistic, (3) too narrow, (4) too anxious, and (5) too small. While cleanly dissecting and acknowledging the foundations of these concerns, Engel (2013a) humbly refrains from completely resolving – or, for that matter, rejecting – them. Also, doctrinal legal scholarship and experimental law and economics can achieve more, if they work hand in hand.

Summary

This entry has introduced the purpose of conducting experiments – also in a legal context – and the basic building blocks of the experimental law and economics paradigm. A cautionary emphasis has been put on the validity trade-off that is inherent in all experimentation. Finally, criticism and objections have been briefly surveyed. These need to be taken – objectively – into account, if experimentalists care for the scientific value of their own method.

As legal scholars, “[w]e will quite likely have to live with some tension between science and doctrine, between individualistic and social constructions of legal problems, between the conditions for causal inference and lumpy institutional choice, between the responsibility for legal development and the competitive pressure of peer-review, between intuition and explication” (Engel 2013a, p. 29). Yet, experimenters can bring to light scientific knowledge that matters for the adoption of new rules, the further development of existing rules, and even the decision of concrete cases. Therefore, doctrinal legal scholars, lawyers, judges, and regulators should find it at least useful and relevant to have the generic knowledge available that has been generated in the laboratory.

In offering a mere introduction and a tertiary source, this entry necessarily stayed at the surface of what can only be called a fascinating methodological challenge. If this text kindled an initial spark among the audience, it was utterly successful. As is the case for the other empirical research methods, however, experiments are not to replace traditional legal discourse. Rather, they are promising and enlightening complements to doctrinal legal scholarship and the continuous evolution of the law and its economic theory.

Cross-References

- [Empirical Analysis](#)
- [Games](#)
- [Good Faith and Game Theory](#)
- [Knowledge](#)
- [Panel Data Analysis](#)

References

- Alberti F, Güth W (2013) Studying deception without deceiving participants: an experiment of deception experiments. *J Econ Behav Organ* 93:196–204
- Ariely D, Norton MI (2007) Psychology and experimental economics: a gap in abstraction. *Curr Dir Psychol Sci* 16:336–339
- Arlen JH, Talley EL (2008) *Experimental law and economics*. Edward Elgar, Cheltenham
- Beattie J, Loomes G (1997) The impact of incentives upon risky choice experiments. *J Risk Uncertain* 14:155–168
- Bonetti S (1998) Experimental economics and deception. *J Econ Psychol* 19:377–395
- Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: a review and capital-labor-production framework. *J Risk Uncertain* 19:7–42
- Crosen R (2002) Why and how to experiment: methodologies from experimental economics. *Univ Ill Law Rev* 2002:921–945
- Davis DD, Holt CA (1993) *Experimental economics*. Princeton University Press, Princeton
- Engel C (2013a) Legal experiments: mission impossible? Eleven International, The Hague
- Engel C (2013b) Behavioral law and economics: empirical methods. Preprints of the Max Planck Institute for Research on Collective Goods Bonn, 2013/1
- Engel C, Kurschilgen M (2011) Fairness ex ante and ex post: experimentally testing ex post judicial intervention into blockbuster deals. *J Empir Leg Stud* 8:682–708
- Falk A, Heckman JJ (2009) Lab experiments are a major source of knowledge in the social sciences. *Science* 326:535–538
- Frank RH, Gilovich T, Regan D (1993) Does studying economics inhibit cooperation? *J Econ Perspect* 7:159–171
- Frank RH, Gilovich T, Regan D (1996) Do economists make bad citizens? *J Econ Perspect* 10:187–192
- Friedman D, Sunder S (1994) *Experimental methods: a primer for economists*. Cambridge University Press, Cambridge
- Hertwig R, Ortmann A (2008) Deception in experiments: revisiting the arguments in its defense. *Ethics Behav* 18:59–92
- Hoffman E, Spitzer ML (1985) *Experimental law and economics: an introduction*. *C Law Rev* 85:991–1036
- Jamison J, Karlan D, Schechter L (2008) To deceive or not to deceive: the effect of deception on behavior in future laboratory experiments. *J Econ Behav Organ* 68:477–488
- Kagel JH, Roth AE (1995) *The handbook of experimental economics*. Princeton University Press, Princeton
- Krawczyk M (2013) Delineating deception in experimental economics: researchers' and subjects' views. University of Warsaw Faculty of Economic Science Working Papers, no 11/2013 (96). Online available at: http://www.wne.uw.edu.pl/inf/wyd/WP/WNE_WP96.pdf. Retrieved 05 June 2014
- Lawless RM, Robbennolt JK, Ulen TS (2010) *Empirical methods in law*. Aspen, New York
- Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? *J Econ Perspect* 21:153–174
- Levitt SD, List JA (2009) Field experiments in economics: the past, the present, and the future. *Eur Econ Rev* 53:1–18
- Plott CR (1982) Industrial organization theory and experimental economics. *J Econ Lit* 20:1485–1527
- Roth AE (1986) Laboratory experimentation and economics. *Econ Philos* 2:245–273
- Smith VL (1976) Experimental economics: induced value theory. *Am Econ Rev* 66:274–279
- Smith VL (1994) Economics in the laboratory. *J Econ Perspect* 8:113–131
- Smith VL (2010) Theory and experiment: what are the questions? *J Econ Behav Organ* 73:3–15
- Smith VL, Walker JM (1993) Monetary rewards and decision costs in experimental economics. *Econ Inq* 31:245–261

Exploring the Deterrent Impact of Financial Supervisory Liability

Robert J. Dijkstra

Tilburg Institute for Private Law, Tilburg University, Tilburg, The Netherlands

Abstract

This entry provides a law and economics analysis of financial supervisory liability. It discusses the deterrent impact of financial supervisory liability by using existing law and economics theory and empirical evidence.

JEL Classification

K13

Introduction

Financial supervisory authorities have the complex task of supervising the financial markets to safeguard the stability of the system as well as to ensure an effective consumer protection. Their main activities consist of licensing of financial institutions (safeguarding entry to the market),

the ongoing monitoring of the health of financial institutions, sanctioning in case of noncompliance with law and regulations, crisis management, and conduct of business supervision (Lastra 2006). When they fail in their activities, they run the risk of being held liable by both third parties and the financial institutions subject to their supervision (Tison 2003; Athanassiou 2011).

In general, public authorities are liable in the same way as any private individual or company. This position follows from the historical fact that the Diceyan conception of the rule of law did not distinguish between public and private. Where a claimant wishes to sue another party, it should, in theory, not matter whether the defendant is a public authority or a private individual or company. It is therefore interesting to notice that nowadays more than 60% of the member states of the European Union have limited the liability of their financial supervisory authorities (Dijkstra 2012).

Why should financial supervisory authorities not be treated in the same way as any other defendant? What, if anything, justifies such a special treatment? The main policy argument, commonly used in favor of shielding financial supervisory authorities from liability is the “defensive conduct” argument. This argument assumes that the imposition of liability will inhibit the effective operations of a financial supervisory authority. The fear of being held liable is so severe that authorities start to act with too much caution when dealing with the supervisee.

On the other hand, submitting financial supervisory authorities to normal liability rules can be based on the preventive function of tort law, in which the threat of being held liable gives financial supervisory authorities incentives to perform their tasks with care (Dijkstra 2009). In this way, financial supervisory liability improves financial supervision by deterring the careless execution of financial supervisory tasks.

Both sides of the debate over the impact of financial supervisory liability can be examined using a law and economics approach. This entry explores the insights that existing law and economics theory and empirical evidence offer about the most likely deterrent impact of financial supervisory liability.

The remainder of the entry is structured as follows. Section “[Can We Apply Economic Analysis to Financial Supervisory Authorities?](#)” discusses the economic analysis of public authority liability. It makes clear to what extent traditional economic analysis can be applied to public authorities and, more specifically, to financial supervisory authorities. Next, section “[To What Extent Does Financial Supervisory Liability Deter?](#)” explores the impact of liability on the behavior of financial supervisory authorities by considering the context in which these authorities operate. It describes specific factors that are likely to influence the deterrent level of financial supervisory liability. As the actual deterrent impact of financial supervisory liability is at heart an empirical question, section “[Empirical Evidence](#)” presents an overview of the existing empirical evidence. Section “[Conclusion](#)” presents the conclusion of this entry.

Can We Apply Economic Analysis to Financial Supervisory Authorities?

The prediction of the effect of a liability rule is sensitive to assumptions about the underlying behavioral model (Spitzer 1977). The party subject to liability needs to be responsive to liability claims for liability claims to have an effect on their behavior at all. Traditionally, law and economics assumes that parties behave rational and want to maximize their utility. By internalizing the costs of their actions via liability rules, they will be motivated to take optimal care and/or engage in less dangerous activities, to prevent damage from happening (Faure 2008). This model seems to fit well to private companies. These companies are generally incentivized to produce at minimum average costs in order to maximize their profits. Faced with liability, they will take preventive measures in order to avoid being held liable and experience a decrease of profits. But what about public authorities? To what extent can the traditional law and economics model also be applied to public authorities? Their objective, unlike those of private companies, is not profit maximization. Legal literature is divided when it comes to answering this question.

Some scholars have argued that public authorities face budget constraints that could result in a more or less similar response to liability as private companies (e.g., Niskanen 1971; Rosenthal 2006). This perspective can be considered as an application of the rational choice model of bureaucratic behavior which assumes that a bureaucrat wants to maximize the size of her agency's budget. Maximizing the agency's budget would provide bureaucrats with other perks they value, such as, for example, prestige, compensation, and career prospects. It has therefore been suggested that public authorities may respond to liability with behavior approximating cost minimization (Kramer and Sykes 1987).

Other scholars are more critical regarding the application of the traditional law and economics model to public authorities. Rosenthal (2006) and Levinson (2000) argue that public authorities will respond to political rather than market incentives. When the political costs of diverting public resources to loss prevention are too high, the government will not make the investment even if it is economically justified (Rosenthal 2006). According to these scholars, liability cannot be expected to promote efficient governmental investment in loss prevention. One could however argue that liability imposes, at least, some political costs. After all, tort liability diverts public funds from other activities that are politically more profitable in terms of public perception and voter's appeal. Public authority liability may, at least potentially, thus create some political incentives to prevent careless behavior. We should however take into consideration that political incentives that may be powerful at the level of elected politicians and officials might not be that powerful for the nonelected bureaucrats employed in public authorities (Dari-Mattiacci et al. 2010).

The idea that liability results in some political incentives might not be completely true when it comes to financial supervisory authorities. While most public authorities are funded from a central, governmental budget, financial supervisory authorities are, more and more, funded from fees imposed on the supervised institutions. This means that any increase of supervisory costs due to law cases against financial supervisory

authorities is ultimately not fully born by the taxpayer but more and more by the financial industry, thereby diluting the impact of political incentives. One could thus question the ability of liability to motivate behavioral changes given the financial authorities' ability to pass (all or at least a significant part of) their costs on to the financial institutions subject to their supervision.

In addition to market and political incentives, one might also want to consider the effect of reputational deterrence (Gold 2016). Reputational deterrence occurs because employees want to protect their organization's pride as a form of protecting self-identity. They thus avoid actions that could expose their organization to lawsuits that can negatively affect their organizational pride. The question remains whether it is the lawsuit or the incident itself that generates negative publicity and hence negatively affects reputation.

There is obviously no easy answer to the question whether and how financial supervisory authorities respond to liability claims. Financial supervisory authorities are probably not solely self-interested or only motivated by reputational incentives, political incentives, or incentives provided by liability (Dijkstra 2010). Their behavior is likely to be influenced by their employees' own morality, social norms, organizational culture, and much more. So, at best, the behavioral consequences of imposing damages liability on public authorities are uncertain. However, the fact that more governments limit the liability of their financial supervisory authorities based on behavioral consequences makes it at least appropriate to investigate the deterrent impact of financial supervisory liability.

To What Extent Does Financial Supervisory Liability Deter?

Conditions for Deterrence

For liability to effectively deter unlawful conduct, a number of conditions have to be met. First, the targets of liability rules must be aware of and understand the rules. Secondly, they need to have the willingness to follow the rules. This will, among other factors, depend on the deterrent

level of that liability claim. In general, the deterrent level of a liability claim can be defined as the capacity to dissuade a potential tort-feasor from behaving carelessly and depends on the one hand on the size of a sanction and on the other hand on the probability of being sanctioned (Baker et al. 2004). In combination, these two variables constitute the expected sanction, and the expected sanction is what influences behavior. The more the public authority believes that it will be sued and punished in a way that has a significant impact, the greater the deterrent effect. The third condition that needs to be met is the ability to conform conduct to the requirements of the rules as actors make decisions about their activities (Robbenolt and Hans 2016).

Various circumstances, often arising from the context in which actors operate, can influence these conditions resulting in either over- or under-deterrence. In a situation of under-deterrence, the deterrent capacity of liability is too low to motivate parties to take optimal care while exercising their activities. In other words, potential tort-feasors do not internalize the full costs of their behavior. The opposite occurs in a situation of over-deterrence. In this case, the fear of being held liable is so severe that a potential tort-feasor engages into so-called defensive conduct. Defensive conduct refers to any act or omission that is performed solely to avoid liability or to provide a good legal defense against a liability claim because of over-deterrence (Hauser et al. 1991).

To determine whether liability can adequately deter undesirable activities, encourage defensive conduct, or has no impact at all, it is necessary to take a closer look at the specific context in which financial supervisory authorities operate.

Willingness to Follow the Rules: The Probability of Being Caught, Sued, and Held Liable

The higher the probability of being caught and sued, the more likely it is that liability will deter potential tort-feasors. So, what is the probability that a financial supervisory authority is caught, sued, and held liable by the court for careless behavior resulting in damage? To answer this

question, we need to make a distinction between two categories of financial supervisory liability (Tison 2003).

First, financial supervisory authorities can be held liable by third parties. In most cases, this will happen after the bankruptcy of a financial institution. Creditors of the financial institution will then try to receive compensation for their losses by suing the financial authority based on shortcomings in performing their tasks. Without a bankruptcy of a financial institution, the financial supervisory authority does not face a big risk of being held liable as the victims will first try to get compensation from the primary wrongdoer, the financial institution itself. The probability of third-party financial supervisory liability depends thus mainly on the probability of a financial institution going bankrupt.

Under normal circumstances, the chances of a financial institution going bankrupt are relatively small. Even when a financial supervisory authority acts negligent while performing its supervisory tasks, it is not likely that that causes financial institutions to go immediately bankrupt. It merely increases the probability of defaults in the future. The opposite is also true. Despite thorough financial supervision, there is still a possibility that a financial institution goes bankrupt. The bankruptcy of a financial institution depends namely on a wide variety of factors and circumstances of which financial supervision is merely one. The chances of an actual bankruptcy are further reduced by the fact that governments will most likely intervene when a financial institution is “too big to fail.” In the past, we have seen that governments will nationalize these troubled financial institutions.

But even if a financial institution goes bankrupt, there is a mechanism that limits the liability risk for a financial supervisory authority. Most countries, namely, have a deposit guarantee system in place that offers compensation to depositors in case a financial institution goes bankrupt (Dijkstra 2009). Only if their losses are (much) greater than the compensation from the deposit guarantee fund will they have an incentive for holding the financial supervisory authorities liable. This group of persons is likely to be limited as

a deposit guarantee fund, with a minimum coverage level of EUR 100,000, normally covers 95% of all deposits (Joint Research Centre 2010). We can thus conclude that the risk of third-party financial supervisory liability is limited, even if financial supervisory authorities are performing their tasks in a negligent manner.

The second category of financial supervisory liability seems to be more straightforward as it concerns a direct relationship between the financial supervisory authority and the victim (e.g., the financial institution subject to supervision). In this case, it is more likely that negligent behavior of a financial supervisory authority results in damage that is immediately detected by the victim. Liability in this category will mainly arise from wrongful actions, in most cases acting too strictly, aimed at the financial institutions. Examples include the wrongful rejection of a permit to operate (market entrance) and a wrongfully published sanction resulting in (reputation) damage for the financial institution. Compared to third-party liability due to too lenient supervision as discussed above, the probability of being held liable by financial institutions due to too stringent supervision is larger.

For both liability categories, it is important to mention that they are affected by the so-called availability heuristic. This behavioral economics term refers to a mental shortcut that relies on immediate examples that come to a person's mind when thinking about a specific topic. Judgments about the probability of being held liable are thus often affected by whether a recent event comes readily to mind. The key issue is thus not the objective risk of being sued and held liable but the perception of the severity of the risk that this will happen (Baker et al. 2004). That might also explain why the Dutch financial supervisory authorities changed their mindset about the threat of liability leading to defensive conduct after a few lawsuits arising from the financial crisis (Dijkstra 2017).

Willingness to Follow the Rules: The Size of the Sanction

As a general starting point, damages should fully compensate the victim for his losses because only then will the injurer internalize the negative

externalities that he has caused (Visscher 2009). So, to provide financial supervisory authorities with the correct incentives to behave carefully, damages should be based on the social losses caused by their negligent behavior. A number of factors influence the size of the sanction that financial supervisory authorities face when being held liable.

First, there is the earlier mentioned deposit guarantee system. This system compensates a significant part of the damage of depositors in case a financial institution goes bankrupt. Consequently, financial supervisory authorities will not face the full costs of their careless behavior. The existence of a deposit guarantee scheme is thus likely to decrease the deterrent level of third-party financial supervisory liability (Dijkstra 2009).

Secondly, financial supervisory authorities might have specific clauses in place that shield them from paying the full amount of damages themselves. These clauses often relate to all categories of financial supervisory liability. The Dutch financial supervisory authorities have, for instance, a safeguard clause in place with the Dutch Ministry of Finance. This safeguard clause limits their financial risk from liability claims to maximum 10% of their budget. In case a liability claim exceeds this maximum, the Dutch Ministry of Finance will cover the remainder of the damage compensation. This means that the Dutch financial authorities will then not internalize the full amount of the damage they have caused.

Although not all financial supervisory authorities have formal safeguard clauses in place, it is likely that implicit safeguard clauses exist. It is not realistic to assume that governments would endanger financial supervision by allowing liability claims to fully consume the budgets of their financial supervisory authorities. Furthermore, the fact that financial supervisory authorities are more and more funded by fees imposed on the financial institutions subject to supervision makes it less likely that the authorities will face the full monetary consequences of their careless behavior as they can pass their costs on to the financial industry.

In addition, financial supervisory authorities may also have the possibility to insure the

financial risks from liability claims, as is the case with the Dutch financial supervisory authorities. In general, law and economics theory predicts that liability insurance will reduce the incentives for potential tort-feasors (Cooter and Ulen 2016).

Awareness and Understanding of the Rules

Potential tort-feasors must be aware and understand the legal rules in order for liability to deter negligent conduct. The traditional economic model assumes that there is no uncertainty regarding the level of care needed to be compliant with the legal rules (Cooter and Ulen 2016). In the real world, however, legal standards are often uncertain. To determine the optimal level of due care, courts need complete and accurate information on the costs of care and the expected costs of accidents for each level of care. However, data necessary to set the optimal level of care will often be unavailable. In addition, courts will not always be able to properly observe the actual level of care exercised by the tort-feasor due to measurement errors, insufficient evidence, and misrepresentation about the actual level of care. Thus, only during a trial, when parties present the facts of the case, is the due level of care established in a more precise way. As a result, tort-feasors exercising a certain level of care might not know *ex ante* whether or not they will be found negligent.

This also applies to financial supervisory authorities. From case law and literature, it becomes clear that the standard for financial supervisory authorities under a negligent liability rule is “reasonable care.” This standard is surrounded by uncertainty, because financial supervisors have discretionary powers to fulfill their supervisory duties and face the difficult task of considering both the interests of the supervised institutions and the individual members of society. Furthermore, there are relatively few tort judgements in the area of financial supervisory liability which makes it difficult to know the exact content of the standard of care. What does this mean for the deterrent effect of tort law?

Uncertainty regarding the level of due care changes the deterrent impact of legal rules by creating two opposing effects (Craswell and Calfee 1986). The first effect is an incentive to

over-comply. Tort-feasors will, in this situation, take more care than is required by the legal standard of care to increase the chance that they will not be held liable. However, uncertainty also creates a chance that a tort-feasor will not be held liable, thus reducing the incentives to comply with the legal standard. The question thus created is which effect will prevail. Standard law and economic theory predict that over-deterrence (and thus defensive conduct) will occur (Craswell and Calfee 1986). Scholars have argued that this is even worse when a public authority is involved due to the fact that, unlike private tort-feasors, a public authority typically balances two external costs, as it does not bear the costs of over-precaution. Public authorities are therefore much more quickly inclined toward taking excessive care (e.g., De Geest 2011; De Mot and Faure 2012).

The existence of uncertainty regarding the standard of due care for financial supervisory authorities might explain why politicians and many others fear over-deterrence and thus defensive conduct on the side of financial supervisory authorities. By limiting the liability of financial supervisory authorities to cases of gross negligence and/or bad faith, they argue that the standard of care becomes clearer and hence would result in limiting over-deterrence and thus defensive conduct.

Ability to Conform Conduct

The third and last condition that needs to be met in order for liability to effectively deter is the ability of parties to conform their conduct to the requirements of the rules as they make decisions about activities in which to engage, the extent and location of those activities, and any precautions to undertake (Robbenolt and Hans 2016).

Law and economic scholars view organizations and thus also financial supervisory authorities as single economic and organic entities. However, it is important to note that public authorities themselves do not commit negligent acts; their employees do. Financial supervisory liability involves the imposition of liability on the organization itself due to the harmful behavior of its employees. Making the financial supervisory authority liable may however fail to provide

its employees with sufficient incentives to act carefully, as the sanctions imposed on the organization might not reach the responsible employees (Dijkstra 2009; Dari Mattiacci et al. 2010). The question of whether incentives are transferred from the financial supervisory authority to its employees is a manifestation of the well-known agency problem between organizations and their employees. The challenge for financial supervisory authorities is, therefore, one of overcoming principal–agent problems.

While most of the financial supervisory authorities have mechanisms in place to mitigate these problems (e.g., reward policies, recruitment policies, internal and external audits), it is realistic to assume that principal–agent problems dilute, to some extent, the deterrent effect of financial supervisory liability.

Over-deterrence, Under-deterrence, or No Effect at All?

The various factors from the previous paragraphs are more likely to result in under-deterrence than in over-deterrence. While a vague standard of care is likely to result in over-deterrence, several other factors contribute to under-deterrence. First, the threat of third-party liability seems to be limited as this liability category follows, in most cases, the bankruptcy of a financial institution which does not occur often. Second, in case a financial institution goes bankrupt, the existence of a deposit guarantee scheme limits the potential damage of third parties and thus also the potential number of claimants. Third, the existence of explicit or implicit safeguard clauses and the possibility to insure the financial risk will further limit the financial consequences for the negligent behaving financial supervisory authority mitigating the deterrent effect of liability. Furthermore, it is questionable whether the financial supervisory authority can effectively transfer the incentives from liability to their employees. Under-deterrence is therefore most likely to occur, thereby questioning the risk of defensive conduct.

It is however difficult, if not impossible, to accurately estimate the deterrent impact of financial supervisory liability from a theoretical perspective. Whether the imposition of liability

promotes more effective financial supervision, encourages defensive practices, or has no discernible effect is at heart an empirical question. The next paragraph presents therefore the outcome of empirical research on this topic.

Empirical Evidence

There is hardly any empirical research on the deterrent impact of financial supervisory liability available. The few studies that exist do show however that defensive conduct is not likely to occur, making under-deterrence more realistic.

Van Dam (2006) asked several Dutch national supervisory authorities, including the financial supervisory authorities, whether their behavior is influenced by the fear of liability. At that time, the supervisory authorities claimed that they did not change their policy out of fear for liability claims. Based on their statements, Van Dam concluded there was no indication for defensive conduct. A couple of years later, the financial supervisory authorities, most likely due to the influence of increased liability claims because of the financial crisis, seemed to have changed their minds and argued for a limitation of their liability. The attitude of the Dutch financial supervisory authorities toward liability claims changed since the financial crisis had put them more into the spotlights. This could indicate that perceptions regarding liability are likely to change when confronted with more liability claims that generate (negative) publicity.

In another study, Trebus and Van Dijk (2014) carried out interviews with four senior employees of the Dutch Financial Markets Authority to examine the impact of liability on their behavior. None of the respondents mentioned any form of defensive behavior in response to liability claims nor did they feel threatened by liability in the exercise of their daily supervisory activities. Based on their research, Trebus and Van Dijk concluded that financial supervisory liability has almost no impact on the behavior of financial supervisors.

A more comprehensive empirical research was carried out in 2015. Dijkstra (2017) conducted a

survey among 500 financial supervisors active in financial supervisory authorities in the member states of the European Union. The majority of the respondents classified the impact of financial supervisory liability as either neutral or positive. The findings of this study therefore imply a modest degree of deterrence. Furthermore, the relative neutral or positive attitude suggests that financial supervisors do not consider financial supervisory liability a burden for executing effective financial supervision. In addition, the survey did not find differences between those respondents who perceive the liability of their organization as limited and those who do not. This suggests that limiting financial supervisory liability does not affect perceptions of the impact of financial supervisory liability or possibly even the behavior of financial supervisors. Therefore, the study calls into question the widely accepted argument of defensive conduct as a reason for limiting the liability of financial supervisory authorities.

This limited empirical research can, however, not be seen as overwhelming empirical evidence regarding the impact of financial supervisory liability. One could further argue that the value of these types of studies (surveys) is limited because they measure perceptions and not actual behavior. They do however raise serious doubts on whether financial supervisory liability will result in over-deterrence and thus defensive conduct.

Conclusion

The theoretical literature is divided regarding the deterrent impact of financial supervisory liability. The picture that emerges from this entry, while lacking detail in many spots, should at least inspire skepticism about the risk of over-deterrence. Our theoretical analysis shows that there are many factors influencing the deterrent level of financial supervisory liability. Most of them tend to lead to under-deterrence, making it hard to believe that financial supervisory liability will result in over-deterrence. This outcome is supported by limited empirical evidence showing only a modest degree of deterrence.

References

- Athanassiou P (2011) Financial sector supervisors accountability: a European perspective, European central bank legal working paper series, No. 12
- Baker T, Harel A, Kugler T (2004) The virtue of uncertainty in law, an experimental approach. *Iowa Law Rev* 89:443–487
- Cooter R, Ulen T (2016) *Law & economics*, 6th edn. Addison-Wesley, California
- Craswell R, Calfee JE (1986) Deterrence and uncertain legal standards. *J Law Econ Org* 2(2):279–303
- Dari Mattiacci G, Garoupa NM, Gomez-Pomar F (2010) State liability. *Euro Rev Private Law* 18(4):773–811
- de Mot J, Faure M (2012) Report for the European group on tort law: state liability: an economic analysis. Ghent University Academic Bibliography, Ghent
- Dijkstra RJ (2009) Liability of financial regulators: defensive conduct or careful supervision? *J Bank Regul* 11(2):269–284
- Dijkstra RJ (2010) Accountability of financial supervisory agencies: an incentive approach. *J Bank Regul* 3(3):346–377
- Dijkstra RJ (2012) Liability of financial supervisory agencies in the European Union. *J Euro Tort Law* 3(3):346–377
- Dijkstra RJ (2017) Is limiting liability a way to prevent defensive conduct? *Eur J Law Econ* 43(1):59–81
- Faure MG (2008) Calabresi and Behavioural tort law and economics. *Erasmus Law Rev* 1(4):75–102
- Geest G de (2011) Who should be immune from tort liability? Washington University in St. Louis Legal Studies Research Paper No. 11-02-04
- Gold RM (2016) Compensation's role in deterrence. *Notre Dame Law Rev* 91(5):1997–2048
- Hauser MJ, Commons ML, Bursztajn HJ, Gutheil TH (1991) Fear of malpractice liability and its role in clinical decision making. In: Gutheil TG, Bursztajn HJ, Brodsky A, Alexander V. (eds) 1991 *Decision making in psychiatry and the law*. Williams & Wilkins Co, Baltimore, pp 209–226
- Joint Research Center (2010) Impact assessment to the commission proposal for the recast of the Deposit Guarantee Schemes Directive
- Kramer L, Sykes AO (1987) Municipal liability under § 1983: a legal and economic analysis. *Supreme Court Rev*:249–301
- Lastra RM (2006) *Legal foundations of international monetary stability*. Oxford University Press, Oxford
- Levinson D (2000) Making government pay: markets, politics and the allocation of constitutional costs. *Univ Chicago Law Rev* 67:345–420
- Niskanen WA Jr (1971) *Bureaucracy and representative government*. Aldine, Atherton, Chicago
- Robbenolt JK, Hans VP (2016) *The psychology of tort law*. New York University Press
- Rosenthal L (2006) A theory of governmental damages liability: torts, constitutional torts, and takings. *J Const Law* 9:1–73

- Spitzer ML (1977) An economic analysis of sovereign immunity in tort. *South Calif Law Rev* 50:515–548
- Tison M (2003) Challenging the prudential supervisor: liability versus (regulatory) immunity, *Financial Law Institute Working Paper* 2003-04
- Trebus J, van Dijk G (2014) Effecten van aansprakelijkheid op het handelen van de AFM empirisch onderzocht (Empirical research of the impact of liability on the Dutch financial markets authority). *Aansprakelijkheid, verzekering en schade* 14(4): 99–107
- van Dam CC (2006) Liability of regulators. An analysis of the liability risks for regulators for inadequate supervision and enforcement, as well as some recommendations for future policy. *British Institute of Comparative and International Law*, London
- Visscher LT (2009) Tort Damages. In: Faure MG (ed) *Tort law and economics*, encyclopedia of law and economics. Edward Elgar, Cheltenham, pp 153–200

External Incentives

► Intrinsic and Extrinsic Motivation

Externalities

Andrew Torre
School of Accounting, Economics and Finance,
Deakin University, Melbourne, VIC, Australia

Abstract

Externalities and court output are closely related since the primary economic function of the courts is to price and thereby internalize external effects in missing markets. This is achieved using the complementary technologies of torts and criminal law. An externalized cost can be broken up into two components, the cost imposed on the unconsenting victim and the cost imposed on wider society, such as insecurity and fear of further random episodes of the same event. Damages awarded to the victim in torts, and criminal punishment inflicted upon the offender, value and internalize the first and second components, respectively. Unless damages and punishment accompany each other, the externality will not

be fully internalized. In addition, punishment produces absolute general deterrence, which is a positive externality, and in this sense negative and positive externalities are mirror images of each other, or perfect complements, in a legal economics framework. This framework also establishes the theoretical underpinnings and basis for valuing joint court output in the civil and criminal jurisdictions.

Synonyms

[Externalized benefits](#); [Externalized costs](#)

Definitions

Externalities are classified as either negative or positive.

A negative externality arises when a decision maker does not bear all of the costs of his or her decision. As a consequence, some costs are externalized onto others without compensating them. An example would be burglary. The private cost to the burglar is the dollar value of the alternative use of his or her time and the expected penalty if caught and convicted. However, the social cost is much higher. First, there is the cost imposed on the victim in the form of property damage and/or loss, and in addition, intangible costs such as fear of further incidents and distrust of others, leading to excessive precautions being taken to protect the property. Second, people other than the victim also incur the fear and self-protection costs. The social cost of burglary therefore exceeds the private cost.

A positive externality arises when a decision maker does not capture all of the benefits of his or her decision. If the burglar is caught and convicted, then a judge will impose punishment in the form of a monetary, noncustodial or custodial penalty. The sanction produces a positive externality in the form of absolute general deterrence, because punishing the burglar deters others from similar behavior in the future. There is a private benefit to the victim, since the penalty may prevent burglary of his or her property in the

future and a much wider benefit to everyone else for the same reason. Therefore, the social exceeds the private benefit.

Introduction

A useful starting point for this essay is to conceptualize the legal system as a production process. In economic theory, a production function summarizes the relationship between the inputs, labor and capital, and the corresponding outputs for a given technology. Similarly, the legal system utilizes a given technology to produce several intermediate outputs and two distinct final outputs from a bundle of inputs. Lawyers distinguish between civil and criminal cases. A case and its ingredients constitute the raw materials of the legal system. For a civil matter these are the cause of action(s), the nature and quality of the evidence and the services of solicitors and barristers, and in a criminal case the offense(s), police, prosecutor, solicitors, and barristers' services. The two final joint civil and criminal court outputs that can be valued in dollar terms are damages and absolute general deterrence, respectively. The latter is jointly produced by the police and courts. Intermediate civil outputs include the court's finding for the plaintiff or defendant, orders of specific performance, injunctions, and in the case of family law matters, custody and property division decisions, for example. Intermediate criminal outputs are bail decisions, committals, trials, sentencing hearings, and the verdict. Absolute general deterrence means that as a consequence of the police clearing up offenses and courts inflicting punishment in the form of fines and jail time on convicted defendants, the social cost of crime will be lower in the future. This is because some prospective offenders at the margin of legal and illegal activity will desist from the latter, since the risk of expected punishment does not make it worthwhile. For this reason, punishment that deters is sometimes called utilitarian punishment (Bagaric 2000). Society is unequivocally better off and therefore net happiness increases. It would only be the case that punishment never deters some people other than the offender, if

penalties for serious offenses were abolished and their incidence did not increase. This seems highly improbable.

The technology of the legal system comprises three interrelated parts. First in common law systems, there is a body of complex past decisions called precedents or more generally the common law. Its birth approximately occurred with the establishment of the royal courts or *curiae regis* by Henry 11, who reigned from 1154 to 1189 (Mendelson 2010, p. 12). Second, the stock of judicial precedents is supplemented by a voluminous compendium of statutory-based law. The equivalent in civil law systems is a set of codes, modeled on the classical Roman law. However, unlike in common law jurisdictions, codes in civil law ones are the most authoritative source of law with judicial decisions being subservient to them. Third, there are the judges who create, interpret, and apply the law.

A production process is said to be efficient if the cost of using the inputs to produce the outputs is minimized; the same reasoning applies to the legal system; damages and deterrence can be supplied efficiently or inefficiently, for a given technology. However, this use of the term efficiency is generally different from the way it is used in the law and economics literature. In the latter, it tends to be employed by some writers when analyzing bodies of substantive law, in particular torts, criminal, contract, and property law (Landes and Posner 1981, 1987; Cooter 1985; Posner 1985, 2010). According to this paradigm, torts legal rules, for example, are efficient if they prospectively minimize the sum of expected damage plus precaution or avoidance costs (Cooter and Ulen 2004, Kaplow and Shavell 1999, Posner 2010). This hypothesis has stimulated the production of a large critical literature. For example, applying it to legal negligence would necessitate the court engaging in a fairly complex calculation exercise, in its inquiry as to whether the injurer had taken cost-effective precautions against the legally caused harm. Hayek's fundamental contribution to the socialist calculation debate in the 1920s and 1930s, in which he argued vigorously against central planning as a substitute for the market mechanism in allocating scarce resources between

alternative uses, was to show that localized, widely dispersed information precluded rational economic calculation on the part of economic planners (Hayek 1945). Similarly, a judge asked to undertake an efficiency cost-benefit calculation would find him- or herself in the same position as a central planner. This led Leoni and Hayek to an alternative characterization of the legal system, particularly common law dynamics, as a spontaneous order derived from the adjudication of individual claims and not the result of human design that aims at any particular end such as wealth maximization or efficiency (Leoni 1972; Aranson 1988; Hayek 1973, Volume 1; Cheren 2012). It seems that this approach fits more comfortably with rights based non-efficiency theories such as corrective justice, which Burrows (1999) defines as “protection of arbitrary alterations to the initial distribution of property rights by restoration of plaintiffs as far as possible to their ex-ante position, thereby preventing defendants from benefiting from the actual harm they have caused.”

The Economic Nature of Externalities and Their Relationship to Law

The subject matter of this essay is externalities, which can be broadly classified into negative and positive (Gans et al. 2012). Textbooks commonly use pollution as an example of the former, and research and development expenditure is common for the latter. A firm that produces steel incurs private costs of \$100,000 (raw materials and labor costs). Residents living across the road suffer from asthma as a consequence of pollution from the factory and incur medical costs of \$10,000. While the total social cost of producing steel is \$110,000, market forces will only reflect the private cost of \$100,000, since the manufacturer externalizes \$10,000 on to the residents. Consequently, market forces overvalue steel and too much is produced and consumed. As the fundamental cause of the problem is an absence of enforceable property rights in the airspace through which the pollution is transmitted, property law provides the basis for a market or private solution.

If either a court or the legislature vests property rights in the airspace to the firm, then the affected residents may be able to bribe the firm to cut back its emissions of pollution in return for a payment. Alternatively if the residents own the airspace, then the firm may be able to successfully negotiate to pay the residents to accept a mutually agreed upon level of pollution. Irrespective of the initial assignment of property rights, transaction costs may impede a bargain from being successfully consummated (Cooter and Ulen 2004; Posner 2010). Sources of high transaction costs include emotional or psychological conflict between the parties, information about the damage cost not being fully available to both parties, disagreements about the cost estimates making it difficult to negotiate a price, or too many parties to deal with (Cooter and Ulen 2004; Posner 2010). The more people that are involved, the less likely it is that an agreement acceptable to everyone will be reached. If bargaining fails, then the standard solution advocated is to force the firm to take into account the \$10,000 externalized cost when it is computing its profit, by imposing a tax or a fine equal to this amount (Gans et al. 2012). This is only efficient if it would be more expensive for the affected residents to take corrective action themselves, for example, by relocating.

In contrast to pollution, research and development expenditure produces a positive externality; consequently market forces will undersupply it. Unimpeded market forces will lead private firms to underinvest in knowledge creation because they are unable to fully appropriate the benefits of their investment. For example, in the case of a new drug that is very expensive to develop, it would be relatively easy for a competitor to discover its molecular formula through reverse engineering. As a consequence drugs will be undersupplied and overvalued (priced too high) by market forces. Traditional solutions are government subsidies, and patenting of the intellectual property embodied in the drug, to rectify the market failure.

These two examples suggest some sort of relationship between legal technology, in particular physical and intellectual property law, and externalities; however, the relationship is at best

cursory. Yet using a legal economics framework, it is possible to see the centrality of torts and criminal law to the problem at hand, and their relationship to the two externality types, in a different light. Torts and criminal law are very much concerned with defining and internalizing negative externalities, and in addition, perhaps counterintuitively, criminal law generates a positive externality. An externalized cost can be broken up into two components, the cost imposed on the unconsenting victim, for example, violation of bodily integrity following a sexual assault, and the cost imposed on wider society, for example, insecurity and fear of further random episodes of the same event. A finding in favor of the victim, followed by an award of damages in torts, values and internalizes the first, while a guilty verdict, followed by punishment, values and internalizes the second component. Full internalization occurs because the injurer is forced to confront both components of the externalized cost. As well, punishment produces deterrence, which is a positive externality or externalized benefit. Since an award of damages without a criminal penalty will only achieve partial internalization, economically, torts and criminal law are complementary technologies, not substitutes.

Historically, this was not always the case, because there were alternative ways for a victim to pursue justice for the same wrongful act, a choice between “compensation or vengeance” (Seipp 1996). If compensation was chosen, it would be paid to the victim and the king, so that the externalized cost would be fully internalized. However, most victims tended to choose vengeance since not many wrongdoers had sufficient wealth to pay the necessary compensation and during the Middle Ages (fifth to the fifteenth century), revenge was considered a “higher right” (Frankel 1996). Professor Mendelson classifies modern torts into three categories: statutory, trespass, and action on the case (Mendelson 2010). Trespass comprises battery, assault, false imprisonment, variants of statutory trespass, trespass to land, trespass to goods, and cattle trespass, while case comprises, *inter alia*, deceit, detinue (unlawful deprivation of possession), conversion, nuisance, libel, slander, defamation, negligence,

passing off, and a series of miscellaneous intentional torts of action on the case for personal injury (*idem.* p. 8). These causes of action protect a multiplicity of entitlements, which include physical integrity or immunity from direct and indirect injury (battery and negligence); the right to use land, light, air, running water, the sea, and its shore (trespass to land, private and public nuisance, and negligence); and the right to corporeal and intellectual property (conversion, detinue, trespass to goods, passing off, misrepresentation, and injurious falsehood) (*idem.* p. 6). All of these causes of action have equivalent counterparts in the criminal law.

Two Illustrations of the Close Nexus Between Externalities and Torts and Criminal Law

Two detailed examples are now used to illustrate the themes explicated in this essay. The first is the case of a property owner who proposes buying an airspace lot above the roof of an apartment block, which is adjacent to his top floor apartment. The entire apartment block is owned by one person and the objective of the property owner’s purchase is to preserve his unimpeded city views. If this market transaction fails and the adjacent owner in any way obstructs the property owner’s view, she will externalize an uncompensated cost onto him. As in any market transaction, the property owner will have a maximum willingness to pay for unimpeded views and the apartment owner will have a minimum or reservation price that she will require before relinquishing the right to her air space lot. For a deal to be consummated, the negotiated price will need to lie between these two values. The two most likely stumbling blocks to a successful outcome would be agreeing on a mutually acceptable price and/or animosity between the parties. These potential problems are called transaction costs and can prevent the completion of a successful bargain that would internalize the externality (Cooter and Ulen 2004; Gans et al. 2012; Posner 2010).

If bargaining fails and the apartment owner subsequently erects a structure on her roof without

in anyway invading the property owner's airspace, then even though she is externalizing a cost on to him, no common law legal right would have been violated. This is because, as reaffirmed by the English House of Lords in 1997 in the case of *Hunter v Canary Wharf Ltd.*, the law of torts does not recognize a legal right to a view from one's property (Mendelson 2010, p. 668). If however the property owner ignored his neighbor's refusal to sell her air space lot and attempted to coerce her into changing her mind, by, for example, entering the balcony of the neighbor's top floor apartment without her permission, then this would infringe a legally recognized right. This externalized cost would be actionable at common law for trespass to the neighbor's land (Mendelson 2010, p. 138). While an injunction ordering the tortfeasor to desist from the trespass would enforce the neighbor's property right, it would not in the economic sense force the property owner to confront the full costs of his decision to bypass the market. The first step in achieving this, i.e., fully internalizing the externality, would be a finding for the victim and an award of compensatory damages, which measures the court's assessment of the externalized cost, imposed on the neighbor. The second step is a finding of guilt and the infliction of punishment. For example, section 9(1)(e) of the Victorian Summary Offences Act 1966 provides for a fine of \$3,500 or up to 6 months imprisonment for criminal trespass. In the absence of punishment, the rest of society would not be compensated for the costs imposed on it. Those social costs include apprehension, distrust of others, and excessive expenditure on self-protection. The deterrence that follows punishment, as well as being an externalized benefit or positive externality, is a public good, because everyone in society is able to consume it simultaneously and no one can be excluded on the basis of willingness and ability to pay. Punitive or exemplary damages would be an imperfect substitute for punishment, even though it is unlikely that they could be insured against, because as already noted, the ability of the injurer to pay can substantially impact their deterrence value. This is particularly the case in countries such as Australia where a defendant cannot

be punished twice by an award of punitive damages and criminal conviction (Mendelson 2010, p. 48).

Legal negligence provides the background for the second example. Mendelson writes: "the advent of railways in the early 1830s, which brought in its wake an unprecedented toll of accidental injuries and death, provided the impetus for the development of negligence as a separate tort" (Mendelson 2010, p. 277). The entitlement not to be injured or killed by negligent driving is one of the rights recognized and protected by this tort and the criminal law. Protecting road and street users' rights not to be interfered with by negligent driving using the market mechanism or Coase bargains would entail very high transaction costs and thus market failure (Coase 1937, 1960). Drivers would have to locate and then contact every potential victim of their negligent driving and then negotiate a price at which each victim would be prepared to relinquish the right not to be negligently harmed. Alternatively, every potential victim would have to locate and contact every potential negligent driver, and then negotiate the price to buy the right not to be harmed. For both injurer and victim, it would not be economical to use cars, streets, or the roads because the search and bargaining costs, (transaction costs) would be too high (Posner 2010). In this case however, unlike the first where the property owner is compelled to bargain with his neighbor if he does not want his view to be obstructed, transaction costs are too high to justify a market solution.

Legal negligence requires the victim to establish that the injurer breached a legally recognized duty of care, which factually causes reasonably foreseeable damage (Mendelson 2010, p. 281). Similar elements constitute the criminal counterpart of the tort of negligent driving. For example, S.318(2)(b) of the Victorian Crimes Act 1958 provides that:

Any person who by the culpable driving of a motor vehicle causes the death of another person shall be guilty of an indictable offence and shall be liable to level 3 imprisonment (20 years maximum) or a level 3 fine or both. For the purposes of subsection (1) a person drives a motor vehicle culpably if he drives the motor vehicle negligently, that is to say, if he fails unjustifiably and to a gross degree to

observe the standard of care which a reasonable man would have observed in all the circumstances of the case.

In both of these examples, the cause of the externalized cost problem is not an absence of property rights, since these are well defined in both cases. The respective sources of the problem are market bypass where a market transaction would have been relatively cheap and failure to observe a legal standard of care in the face of prohibitively high transaction costs that make market bypass impossible (Posner 2010).

Judicial Valuation of Externalities

The notion of “price lists” for externalized harms can be found very early on in the history of Anglo-Saxon and Germanic law. An example is “the code of Æthelberht, the King of Wessex, which contains elaborate tariffs of fines for breach of the peace” (Mendelson 2010, p. 10). More generally, “wergeld” tables served two functions. First, they provided for a fixed scale of compensation, which was determined by the victim’s social and legal status. For example, the “wergeld” of princes and free lords was 360 shillings, the judicial class 30 shillings, rent-paying tenants and other free-men 15 shillings, and a day laborer’s “wergeld” was paid in wheat (Mendelson 2010, p. 34). Second, the tables put a monetary value on injuries to different parts of the body, which was proportional to the victim’s “wergeld”; for example, the price of permanent injury to the mouth, nose, eyes, tongue, ears, male sexual organs, hands, and feet was set equal to one half of the victim’s “wergeld” (Mendelson 2010, p. 35).

In contemporary times judicial valuations of externalized costs can be inferred from the average quantum of compensation they award in torts to victims and the average quantum of punishment they impose on offenders. The averages aggregate and summarize the dispersed or localized information about the distribution of diverse cases for a particular cause of action and offense. They are socially optimal because the judges’ comparative advantage is in producing justice according to

law. This requires access to all of the relevant information about the case, and knowledge of the relevant substantive law, which they only possess. The last statement needs to be somewhat qualified when considering the ‘optimality’ of court awarded tortious damages. In Australia, following reform legislation early this century, a distinction is now made between intentional and unintentional wrongs; the former are still governed by the common law of damages, while the latter are governed by statute (Mendelson 2010). Statutory based damages are subject to statutory thresholds and capping (Mendelson 2010). Consequently, judicially determined awards in the case of negligence for example, are constrained optima. However common law principles also considerably constrain the courts’ assessment of the victim’s disability. In estimating ‘the quantum of damages for past and future pain and suffering, loss of enjoyment of life and loss of earning capacity’ courts apply the ‘once for all’ rule established in 1699 (Mendelson 2010). Compensation is a lump sum, which cannot be subsequently changed, if the plaintiff’s injury worsens or develops into something more serious down the track (Mendelson, 2010). Similarly, when victims sustain very serious and permanent injuries, for example quadriplegia, courts face insuperable obstacles in forecasting the victim’s future loss of earning capacity and nursing requirements. As noted by Mendelson (2010), this situation makes it very likely that the gravity of the victim’s injury and injurer advantage are positively correlated. Average damages denoted by AD measure the price of the externalized harm borne by the victim. The supply of absolute general deterrence plus the judicial valuation of the externalized cost to society implied by the optimal punishment is found by minimizing Eqs. 1 and 2 with respect to X^* and F^* , the optimal jail sentence and fine, respectively:

$$\begin{aligned} \text{MIN.E}[C] &= E[D[O[X^*]]] + P X^* b O[X^*] \\ &= D[O[X^*]] + E P X^* b O[X^*] \end{aligned} \quad (1)$$

$$\begin{aligned} \text{MIN.E}[C] &= E[D[O[F^*]]] + P F^* O[F^*] \\ &= D[O[F^*]] + E P F^* O[F^*] \end{aligned} \quad (2)$$

where $E[C]$ is expected cost, D is harm, O is offenses, EP is the expected probability of the offender being jailed or fined, b is the cost of incarcerating an offender each time period, and X^* and F^* are the average jail sentence and fine for the offense. The asterisk indicates that they are exogenous and optimal in the sense explained earlier. The solutions are given by Eqs. 3 and 4, respectively:

$$\frac{F^*EP}{e} + F^*EP \quad (3)$$

$$\frac{X^*b^*EP}{e} + X^*b^*EP \quad (4)$$

The first term on the RHS of Eqs. 3 and 4 is the value of the positive externality corresponding to the optimal fine F^* and term of imprisonment X^* , given EP the probability of the sanction being imposed and e the absolute general deterrence elasticity, whose value will lie between 0 and -1 . A value of -0.5 , for example, indicates that a 1% (10%) increase in the average fine or jail sentence lowers the expected future costs of crime by 0.5% (5%), etc. A value of -1 indicates perfect absolute general deterrence. The second term of Eqs. 3 and 4 is the court's valuation of the externalized social cost of the offense, i.e., the burden placed on wider society:

$$V_N = AD + F^*EP \quad (5)$$

$$V_N = AD + X^*b^*EP \quad (6)$$

Expressions 5 and 6 give the dollar value (V_N) of the court's valuation of the negative externality implicit in the award of tortious damages (AD) and the punishment imposed for the corresponding offense, while Eqs. 7 and 8 give the value of the positive externality (V_P) from deterrence:

$$V_P = \frac{F^*EP}{e} \quad (7)$$

$$V_P = \frac{X^*b^*EP}{e} \quad (8)$$

The two cases discussed in the third section of this essay can be used to illustrate how these

expressions would be translated into dollar values. The parameter values used are deterrence elasticity -0.3 , the probability of a fine or imprisonment 0.5 , and the annual cost of incarceration $\$80,000$.

Case 1

	Damages	Fine (average value)	Imprisonment
Tort of trespass to land	\$5,000		
Offense of trespass to land		\$5,000	6 months
Judicial value of negative externality			
(i) Damages + fine		\$7,500	
(ii) Damages + imprisonment			\$15,000
Judicial value of positive externality			
(i) Fine		\$8,333	
(ii) Imprisonment			\$33,333
Total: negative and positive externality		\$15,833	\$48,333
Civil output	\$5,000		
Criminal output		\$10,833	\$53,333
(Assumes only 1 case: civil and corresponding criminal matter)			

Case 2

	Damages	Fine (average value)	Imprisonment
Tort of negligence	\$20,000		
Offense of culpable driving		\$20,000	1 year
Judicial value of negative externality			
(i) Damages + fine		\$30,000	
(ii) Damages + imprisonment			\$60,000
Judicial value of positive externality			
(i) Fine		\$33,333	
(ii) Imprisonment			\$133,333
Total: negative and positive externality		\$63,333	\$193,333
Civil output	\$20,000		
Criminal output		\$43,333	\$173,333
(Assumes only 1 case: civil and corresponding criminal matter)			

Summary

The production of goods and services that are sold in the marketplace increases social welfare because it generates a social surplus, profit plus consumer surplus. The latter is the difference between the average prices paid for an item and buyers' maximum willingness to pay and is a measure of consumer welfare. Profit measures business surplus, which is the difference between revenue and the opportunity cost of all scarce resources used to generate it. Conceptually

valuing the output of products that are priced in the market is straightforward. This is not the case however for nonprofit services that are not priced by the market mechanism.

Final output valuation in these cases requires a well-developed theory of the organization's economic functions. In relation to the courts, the position adopted in this essay is that they price external effects in missing markets, in both the civil and criminal jurisdictions. The valuation of their final output follows logically from this. Average civil and criminal jurisdiction output is equal to expressions $5 + 7$, where fines are used as punishment, and expressions $6 + 8$, where imprisonment is used. Total civil jurisdiction output is average damages multiplied by the number of cases (the first term in Eqs. 5 and 6), and total criminal jurisdiction output is the second terms of Eqs. 5 and 6 plus Eqs. 7 and 8 multiplied by the number of cases.

As the two examples show, the pricing exercise is related to the sanction used; the negative externality of land trespass is judicially valued at \$7,500 (damages + fine) and \$15,000 (damages + imprisonment), while the corresponding figures for negligent driving are \$30,000 and \$60,000, respectively. The value of the positive externality or public good of deterrence is \$8,333 (fine), \$33,333 (imprisonment) for land trespass, and \$33,333 and \$133,333 for negligent driving.

References

- Aranson P (1988) Bruno Leoni in retrospect. *Harv J Law Publ Policy* 11:661–711
- Bagaric M (2000) Punishment and sentencing: a rational approach. Cavendish Publishing, London
- Burrows P (1999) A deferential role for efficiency theory in analysing causation-based tort law. *Eur J Law Econ* 8:29–49
- Cheren R (2012) Tragic parlor pigs and comedic rascally rabbits: why common law nuisance exceptions refute

- Coase's economic analysis of law. *Case West Law Rev* 63(2):556–596
- Coase R (1937) The nature of the firm. *Economica* 16(4):386–405
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cooter R (1985) Unity in tort, contract and property: the model of precaution. *Calif Law Rev* 73(1):1–51
- Cooter R, Ulen T (2004) Law and economics, 6th edn. Pearson Addison Wesley, Boston
- Frankel T (1996) Lessons from the past: revenge yesterday and today. *Boston Univ Law Rev* 76:89–102
- Gans J, King S, Gregory Mankiw N (2012) Principles of microeconomics. Cengage Learning Victoria, Australia.
- Hayek F (1945) Use of knowledge in society. *Am Econ Rev* 35(4):519–530
- Hayek F (1973) Law legislation and liberty. Rules and order, Vol 1. The University of Chicago Press, Chicago.
- Kaplow L, Shavell S (1999) Economic analysis of law. Working paper 6960. National Bureau of Economic Research, Cambridge, Mass.
- Landes W, Posner R (1981) The positive economic theory of tort law. *Ga Law Rev* 15:851–921
- Landes W, Posner R (1987) The economic structure of tort law. Harvard University Press, Cambridge, Mass.
- Leoni B (1972) Freedom and the law. Nash Publishing, Los Angeles
- Mendelson D (2010) The new law of torts, 2nd edn. Oxford University Press, Australia and New Zealand.
- Posner R (1985) An economic theory of the criminal law. *Columbia Law Rev* 85(6):1193–1231
- Posner R (2010) Economic analysis of law, 8th edn. Aspen Publishers, New York
- Seipp D (1996) The distinction between crime and tort in the early common law. *Boston Univ Law Rev* 76:59–88

Externalized Benefits

- Externalities

Externalized Costs

- Externalities

F

Fake

- [Counterfeiting Models: Mathematical/Economic](#)

Family

- [Adoption](#)

Fiduciary Duties

Remus Valsan
Edinburgh Law School, University of Edinburgh,
Edinburgh, Scotland, UK

Definition

Fiduciary duties arise in legal relations where one party, the fiduciary, acquires decision-making authority over the interests of another party, the beneficiary. The first party becomes bound by a set of duties aimed at ensuring that he exercises his discretion in the best interests of the beneficiary, to the exclusion of his own interests or the interests of third parties. Fiduciary duties are strictly enforced by the courts, in order to ensure a proper exercise of discretion by the fiduciary and to preserve the utility of fiduciary relations.

Fiduciary Relations

Fiduciary duties arise in fiduciary relations. Fiduciary relations exist in multiple forms in different legal areas, such as private law, commercial and corporate law, family law, and, in some cases, public law. The concept of fiduciary duties is sometimes regarded as a purely common law concept, historically intertwined with the institution of trust and with the equitable jurisdiction of the English Court of Chancery (Sealy 1962). Looking at the substance, rather than the history or technical details of fiduciary relations, however, reveals that such relations exist across all legal traditions. The common law and mixed jurisdictions have developed an elaborate set of principles which constitute a body of fiduciary law. In civil law countries, in contrast, fiduciary relations have developed in a more isolated and fragmentary manner (Graziadei 2014). For this reason, most of the relevant theoretical literature comes from common law or mixed jurisdiction authors. Because the underlying features and issues are similar in all fiduciary relations, many of the insights discussed below are not jurisdiction specific.

It is commonly stated that the family of fiduciary relations is open ended. It comprises established (or *per se*) fiduciary relations and *ad hoc* fiduciary relations. Examples of established relations include trustee-beneficiary, agent-principal, tutor-ward, director-company, or solicitor-client. In some jurisdictions, especially Canada and the USA, the list of

established categories is longer: it includes relations such as physician-patient, clergy-parishioner, parent-child, or Crown-aboriginal peoples. Fiduciary duties may also arise *ad hoc* in legal relations that do not belong to an established category, based on the concrete factual circumstances. In the *per se* fiduciary relations, fiduciary duties are presumed to exist (rebuttable presumption), whereas in the *ad hoc* relations, the existence of fiduciary duties must be proved, based on relevant indicia (Shepherd 1981).

The recognition of the open ended nature of the family of fiduciary relations has created the need to identify the core indicia or elements that trigger the application of fiduciary duties in new circumstances. Beyond the general consensus concerning the two types of scenarios where fiduciary duties may arise, courts and theorists have expressed many different views with respect to the necessary and sufficient core elements that attract fiduciary duties. Two elements are increasingly recognized as indispensable for fiduciary duties to exist: an undertaking to act for the interests of another coupled with power or discretion to affect the other's legal or practical interests (Valsan 2016).

The requirement of undertaking to act for another signifies that fiduciary duties are triggered voluntarily. They are enforceable only against those persons who undertook to do something for the benefit of another person or for an abstract purpose. This means that fiduciary duties cannot be imposed by courts *ex post* in an instrumental way, in order to achieve a certain outcome or trigger a desirable legal remedy, if there is no undertaking from the fiduciary to act in the beneficiary's interests. Furthermore, the fiduciary undertaking to act in the interests of another is not a duty to achieve a certain result, *i.e.*, a binding obligation to achieve the best result available for the beneficiary. Such an obligation would minimize or remove the role of discretion and would render the fiduciary office very cumbersome. Moreover, it is established that courts will not second-guess a fiduciary's judgment *ex post*, based on the results achieved. This rule, known in the company laws of certain jurisdictions, such as Delaware, as the business judgment rule,

protects fiduciaries that exercise their judgment in good faith, with due care and without conflicts of interest, from liability based on the results of their decisions (Valsan and Yahya 2007). The requirement of power, or decision-making authority involving exercise of judgment, indicates that fiduciary duties are imposed only when a party has scope to exercise discretion over the interests of another (Thomas and Hudson 2010). In contrast, when a contractual party has an obligation to act in the other party's interest in a clearly prescribed manner or following the other party's directions, contractual, rather than fiduciary liability, will normally arise.

Fiduciary Duties

When a party undertakes to act in the interests of another and acquires decision-making authority with regard to how to pursue the other party's interests, fiduciary duties are imposed in order to ensure the integrity of the exercise of discretion. Which duties are properly regarded as fiduciary is a matter of some controversy. All authors agree that the proscriptive (negative) duties forbidding fiduciaries to act in a conflict of interest or to seek unauthorized benefits are fiduciary duties. Some authors add to these proscriptive duties a set of positive (or prescriptive) duties, such as a duty to act in good faith, a duty of reasonable care and diligence, a duty to disclose relevant information, or a duty not to disclose confidential information (Finn 1977). The fiduciary status of these positive duties is disputed. The preferable view is that, although duties of good faith, care, confidentiality, or disclosure are often associated with a fiduciary position, they apply to a wide spectrum of non-fiduciary legal actors as well and therefore should not be labeled fiduciary (Conaglen 2010).

The proscriptive fiduciary duties are commonly expressed as four fiduciary rules: the no-profit rule, the no-conflict rule, the self-dealing rule, and the fair dealing rule. The no-profit rule forbids a fiduciary from retaining an unauthorized benefit acquired by virtue of his fiduciary position. The no-conflict rule states that a fiduciary is not allowed to place himself in a position where

his personal interest, or interest in another fiduciary capacity, conflicts or possibly may conflict with his duty. The self-dealing rule renders voidable, at the beneficiary's will, purchases by a fiduciary, in his personal capacity, of property under his administration, irrespective of the honesty of the transaction. The fair dealing rule renders voidable the purchase by a fiduciary of the beneficiary's interest, unless the fiduciary demonstrates that the transaction is entirely fair and honest and that the beneficiary gave his informed consent (Smith 2003).

A defining feature of the proscriptive fiduciary duties is their particular strictness. The degree of strictness varies with the type of fiduciary relation and jurisdiction. For instance, fiduciary duties of trustees are generally more demanding than those of company directors, and US courts are more permissive than UK courts when it comes to self-dealing by directors (Kershaw 2012). Notwithstanding these differences, fiduciary duties are often regarded as stricter than other private law obligations. One manifestation of this strictness is the reprehensibility of the appearance of self-interested conduct. A fiduciary will be liable for breach of the no-conflict rule not only in case of an actual conflict between interest and duty but also when he acts in a situation of potential conflict of interest. A potential conflict arises when there is a real sensible possibility of conflict, and the fiduciary is liable if he acts in these circumstances irrespective of his good faith or proper motivation. Moreover, when the proscriptive duties are breached, a fiduciary is generally liable irrespective of whether the beneficiary suffered any loss or even obtained a benefit following the conflicted transaction. Furthermore, a fiduciary is forbidden from taking a business opportunity he came across while exercising his role, even if the opportunity is not available to the beneficiary (Conaglen 2010).

Remedies for Breach of Fiduciary Duties

Breach of fiduciary duties attracts personal and proprietary remedies. Some remedies are aimed to recover the profits that the fiduciary obtained

through breach of fiduciary duty. They focus on undoing the wrong rather than on compensating the beneficiary for his loss. Disgorgement of profits may be sought under a personal claim in the form of account of profits, or under a proprietary claim, in the forms of constructive trust or equitable lien. Other remedies aim to compensate the beneficiary for his loss: the equitable compensation or damages. Another potential remedy, applied more rarely and when the breach was in bad faith, is the imposition of exemplary (or punitive) damages, over and above disgorgement of profits, or compensation for loss. Furthermore, a transaction entered into in breach of fiduciary duty is voidable at the instance of the beneficiary. This remedy usually leads to rescission, where the transaction is unwound and the parties are restored to their previous positions (McGhee et al. 2015).

These remedies could be generous to beneficiaries and therefore are extremely attractive. The beneficiary may bring a claim for disgorgement of profits without the need to prove loss; a proprietary claim will shelter the beneficiary against the consequences of fiduciary's insolvency; a claimant can seek compensation for loss without being restrained, in form at least, by the common law rules of causation, remoteness of damage, or contributory negligence. The desire to have access to these remedies is one of the reasons why claims for breach of fiduciary duty are very popular (Millet 1998).

When an actual or potential breach occurs, a fiduciary has the option to escape liability by fully disclosing the relevant information (e.g., the actual or potential conflict of interest or the benefit that was not authorized *ex ante* by the relationship) to all relevant beneficiaries and seek their informed consent.

The Purpose of Fiduciary Duties

Fiduciary Duties as Gap Fillers

In a view that dominates the law and economics literature in this area, the purpose of fiduciary duties is to fill in the gaps in the fiduciary contract. Because the fiduciary relation engenders

unusually high costs of specification and monitoring, parties specify *ex ante* a generic duty of loyalty (the fiduciary's duty to act in the best interests of the beneficiary), and courts spell it out *ex post* by identifying what the parties would have agreed on if they had written a fully specified contract (Easterbrook and Fischel 1996).

The gap filling approach, also known as the contractarian theory of fiduciary duties, dominates the current law and economics fiduciary theory. Fiduciary relations are regarded as a particular type of contract, one that is characterized by large gaps in the parties' agreement (Cooter and Freedman 1991). Although most contracts are incomplete, fiduciary relations have a particularly high degree of incompleteness. There are two main causes for this. First, parties to a fiduciary contract seldom label their relation as fiduciary, rarely include terms from the standard fiduciary vocabulary, such as loyalty or unselfishness; and seldom bargain *ex ante* over all circumstances that may constitute a conflict of interest or an unauthorized benefit (Duggan 2010). Second, at a more fundamental level, the high incompleteness of the fiduciary contract is due to its essential feature: the beneficiary (the principal) hires an expert fiduciary (the agent) and entrusts him with management and control of an asset or with another task involving exercise of discretion. The agent acquires an open-ended power over the asset and undertakes imperfectly observable discretionary actions that affect the principal's wealth. Because the agent's expertise and professional judgment are a key justification for the creation of the fiduciary relation, the principal has neither the ability nor the incentive exhaustively to specify *ex ante* what the agent should do (Jensen and Meckling 1976).

Fiduciary duties are, thus, a tool used to fill in the contractual gaps in the fiduciary contract. The completion of the fiduciary contract is often delegated to courts. Courts supply the missing contractual provisions by identifying the terms the parties themselves would have agreed to, had they negotiated about the unanticipated circumstance *ex ante* (Alces 2015). In the contractarian approach, fiduciary duties are mere standard terms in contracts, derived and enforced in the same way

as other contractual undertakings. The implications of this approach are manifold. First, contractarians deny that fiduciary duties serve a moral or public interest function. They have the same objective as contractual obligations in general, namely, to implement the parties' own perception of their joint welfare (Easterbrook and Fischel 1996). Second, the court's role in enforcing a fiduciary contract is limited to supplying default rules that parties are free to contract around *ex ante*. Parties who are able to determine *ex ante* that they do not want fiduciary duties to apply to certain aspects of their bargain are free to exclude such circumstances either generally or on a case-by-case basis, via an obligation of *ex ante* disclosure and informed consent. Consequently, the fundamentally contractual nature of fiduciary duties means that courts cannot override express or implied contractual terms in furtherance of other, higher-ranking, interests. When economic agents enjoy complete freedom and autonomy to decide the form and content of their bargains, they engage in wealth-maximizing exchanges that promote economic efficiency (Duggan 2010).

Fiduciary Duties as a Tool to Protect Relations of Trust and Confidence

Another view, sometimes referred as the anti-contractarian approach to fiduciary duties, is based on the idea that fiduciary relations have a special nature compared to regular private or commercial interactions. They are marked by a high degree of trust and confidence that one party places in the other, which causes the former to be particularly vulnerable to abuse of power by the latter (Flannigan 1989). Once the beneficiary relinquishes control over the entrusted asset or task, the fiduciary is presented with opportunities to misappropriate the asset or opportunities arising in relation to its management, whether by theft, conversion, self-dealing, or negligence (Frankel 1995). The idea of vulnerability to abuse is sometimes conveyed using the concept of fiduciary expectation. The most important indicator of a fiduciary relation is the fiduciary expectation, which entitles one party to expect that the other will act in the first party's interests for the purpose of the relationship to the exclusion

of his own personal interests. Due to this expectation, the trusting party relaxes his self-vigilance and independent judgment and becomes vulnerable to abuse. Consequently, in this view, courts impose *ex post* fiduciary duties not as an approximation of what parties would have agreed to, but in order to protect vulnerable people in power-dependency relationships (Finn 1989).

In the anti-contractarian view, fiduciary duties are an instrument of public policy used to maintain the integrity, credibility, and utility of relationships perceived to be of importance in society. Because such unequal power-dependency relations are pervasive and particularly valuable for the society, they have moral and public interest dimensions not captured by the contract law framework. Due to these specific dimensions, in certain cases fiduciary duties trump the individual interests of the parties.

Fiduciary relations and regular contractual relations are viewed as fundamentally different. In contract parties are usually on equal footing and expected to further their own interest, within the limits of good faith, unconscionability, and undue influence. Fiduciary relations, in contrast, have trust and confidence, vulnerability, and dependency of one party on another, at their core. The underlying rationale of the fiduciary obligation is not individualistic private ordering. The law in this area serves an educative or pedagogic function, aiming to heighten the morality of the marketplace by raising the standards of commercial dealings above ordinary market temptations.

Fiduciary Duties as a Tool to Protect the Exercise of Fiduciary Discretion

This view of fiduciary duties links the proscriptive fiduciary duties with the essential characteristic of a fiduciary relation, the decision-making authority over the interests of another. The content and purpose of duties specific to persons in a fiduciary position are easier to grasp if these duties are separated into two main groups. On the one hand, there are the traditional proscriptive fiduciary duties, usually articulated as the no-conflict and no-profit rules. On the other hand, there is a core duty binding on fiduciaries, which is different from the proscriptive duties and which justifies

their existence. The proscriptive duties are connected with the core duty in the sense that they play a protective or prophylactic role: they aim to prevent violations of the fundamental fiduciary duty. The core duty binding on a fiduciary is the duty to exercise discretion based on relevant considerations (Valsan 2016).

In general terms, a fiduciary is bound to exercise discretion within the objective limits of his powers and in what he believes to be the best interest of the beneficiary or the scope for which the power was granted. The determination of beneficiaries' best interests or of the purposes for which a power was granted allows the fiduciary a large degree of subjectivity, within a clearly defined legal perimeter. An appropriate exercise of discretion imposes on fiduciaries two requirements. First, a fiduciary must exercise active discretion, in the sense of applying his mind and reaching a conscious decision regarding the need for, and the implications of, exercising any power or discretion that he holds in fiduciary capacity. Second, if a fiduciary decides that it is opportune to exercise a power, he must decide where the best interests of the beneficiary lie, or what is the best way to achieve the purpose for which the power was given, depending on the type of power under exercise. The two aspects of the exercise of judgment involve a similar decision-making process: fiduciaries must decide based on relevant considerations (Thomas 2013).

The proper judgment duty has a procedural nature. It tells a fiduciary what to do when exercising discretion, rather than what is a relevant consideration for each decision. Determining the considerations that are relevant to an exercise of discretion is a matter for the fiduciary's judgment. Such considerations include the nature and the purpose of the particular power to be exercised; the relationship that the power has to the other powers and duties of the fiduciary; the nature of the transaction in which the fiduciary intends to perform; the wishes, circumstances, and needs of beneficiaries; fiscal considerations; and so on. The weight that each of these relevant factors should carry in determining the course of action is also a matter left to fiduciary's judgment. As long as fiduciaries apply their mind to the importance of

a relevant consideration for a particular decision, they comply with the duty of real and genuine consideration of relevant factors, irrespective of the actual outcome of their decision (Thomas 2013).

The proscriptive fiduciary duties protect the core duty to exercise judgment based on relevant considerations in two ways. First, they compel a fiduciary consciously to eliminate self-interest or the interest of a third party not relevant to the fiduciary relation from the list of considerations that a fiduciary is allowed to take into account when exercising judgment. Secondly, they protect the exercise of judgment against any potential indirect or nonconscious effect that the presence of an irrelevant interest may have on fiduciary's judgment. The mere presence of an actual or potential intruding interest affects the reliability of fiduciary's judgment in ways that cannot be measured and corrected against. Extraneous interests have the potential to affect the way in which the decision-maker evaluates the seriousness of various risks, the desirability of certain outcomes, or the perception of connections between cause and effect (Davis 2012). Conflict of interest situations are reprehensible because they create an unusual risk of error, thus rendering one's judgment less reliable (Norman and Macdonald 2010). Since the effects of an extraneous interest cannot be accurately measured, the reprehensibility of a fiduciary conflict of interest does not depend on the fiduciary's good faith or genuine desire to act solely in the beneficiary's best interests. A fiduciary is in breach of the no-conflict rule on the basis of being in a conflict situation, irrespective of his belief that he is capable of resisting the temptation or corrupting influence of the interest that could interfere with his judgment.

Conclusion

Fiduciary duties are obligations binding on persons occupying fiduciary positions, which compel such persons to refrain from adopting decisions in circumstances that may amount to conflict of interest, unauthorized benefits, self-dealing, or unfair dealing with the beneficiary. They guide

and protect the exercise of the fiduciary's decision-making authority by removing the actual or potential influence on fiduciary's discretion of factors or interests that are not relevant to the scope of fiduciary's authority. Fiduciary duties have been regarded as an efficient method to fill in the large gaps in fiduciary contracts by postponing or delegating to courts the specification of the actions that a fiduciary is allowed or prohibited in his mandate to promote the interests of the beneficiary. Because courts tend to have a strict approach to fiduciary's liability, fiduciary duties have also been regarded as a public policy tool that protects valuable social relations characterized by trust, confidence, power, and dependency.

Cross-References

► [Conflict of Interest](#)

References

- Alces K (2015) The fiduciary gap. *J Corp Law* 40:351
- Conaglen M (2010) *Fiduciary loyalty: protecting the due performance of non-fiduciary duties*. Hart Publishing, Oxford
- Cooter R, Freedman B (1991) The fiduciary relationship: its economic character and legal consequences. *N Y Univ Law Rev* 66:1045
- Davis M (2012) Empirical research on conflict of interest: a critical look. In: Peters A, Handschin L (eds) *Conflict of interest in global, public and corporate governance*. Cambridge University Press, Cambridge, p 54
- Duggan A (2010) Contracts, fiduciaries and the primacy of the deal. In: Bant E, Harding M (eds) *Exploring private law*. Cambridge University Press, Cambridge, p 275
- Easterbrook FH, Fischel DR (1996) *The economic structure of corporate law*. Harvard University Press, Harvard
- Finn PD (1977) *Fiduciary obligations*. Law Book Co, Sydney
- Finn PD (1989) The fiduciary principle. In: Youdan TG (ed) *Equity, fiduciaries and trust*. Carswell, Toronto
- Flannigan R (1989) The fiduciary obligation. *Oxf J Leg Stud* 9:285
- Frankel T (1995) Fiduciary duties as default rules. *Oregon Law Rev* 74:1209
- Graziadei M (2014) Virtue and utility: fiduciary law in civil law and common law jurisdictions. In: Gold A, Miller P (eds) *Philosophical foundations of fiduciary law*. Oxford University Press, Oxford, p 287

- Jensen M, Meckling W (1976) Theory of the firm: managerial behaviour, agency costs and ownership structure. *J Financ Econ* 3:305
- Kershaw D (2012) The path of corporate fiduciary law. *NYU J Law Bus* 8:395
- McGhee J et al (eds) (2015) *Snell's equity*, 33rd edn. Sweet & Maxwell, London
- Millett P (1998) Equity's place in the law of commerce. *Law Q Rev* 114:214
- Norman W, MacDonald C (2010) Conflicts of interest. In: Brenkert G, Beauchamp T (eds) *The Oxford handbook of business ethics*. Oxford University Press, Oxford, p 441
- Sealy LS (1962) Fiduciary relationships. *Camb Law J* 20:1
- Shepherd JC (1981) *The law of fiduciaries*. Carswell, Toronto
- Smith L (2003) The motive not the deed. In: Getzler J (ed) *Rationalizing property, equity, and trusts: essays in honour of Edward burn*. LexisNexis UK, London, p 53
- Thomas GW (2013) *Thomas on powers*, 2nd edn. Oxford University Press, Oxford
- Thomas G, Hudson A (2010) *The Law of Trusts*, 2nd ed. Oxford, Oxford University Press
- Valsan R (2016) Fiduciary duties, conflict of interest and proper exercise of judgment. *McGill Law J* 62:1
- Valsan, Yahya (2007) Shareholders, creditors, and directors' fiduciary duties: a law and finance approach. *Virginia Law Bus Rev* 2:1

Financial Education

Marco Novarese¹ and Viviana Di Giovinazzo²

¹Department of Law and Economics, University of Eastern Piedmont, Centre for Cognitive Economics, Alessandria, Italy

²Department of Sociology and Social Research, University of Milano Bicocca, Milan, Italy

Definition

Financial education is an area of research involving finance, psychology, and behavioral and cognitive economics that grew out of the need to improve the financial skills of citizens. It aims to promote consumer awareness concerning the functioning of financial markets through financial training and continuing education which help to develop the skills and knowledge that allow

individuals to make informed, effective decisions concerning their financial resources.

Historical Outline

Financial education was introduced into a number of American secondary schools from the late 1950s onward, with the aim of providing citizens with basic financial knowledge on incomes and savings, taxation, first home buying, insurance, and pensions. During the 1960s, reinforced by the Johnson Great Society Program and Ralph Nader's consumer protection campaigning, financial education programs multiplied and even became compulsory in some states. By the 1990s there was a growing awareness of the need for financial education, with citizens facing a rapid increase in the supply of products and financial services.

In 1995, the US Department of Labor and the Treasury, as well as 65 public and private organizations, organized the first American Savings Education Council. Surveys investigating household financial decision-making have been regularly conducted since 1997, especially by the Jump\$tart Coalition for Personal Financial Literacy. In 1998, the Securities and Exchange Commission set up the Facts on Savings and Investing Campaign. Its initial report states that "One of [financial education's] major goals is that all Americans are armed with the information they need to make sound financial decisions and protect their hard-earned savings" (SEC 1999). Most surveys report a progressive worsening in financial matters.

From the data it emerges, for example, that the average family debt with credit card companies rose from an average of \$2,985 in 1990 to an average of \$8,300 in 2002 (Fisher 2003). Other data indicate that in 2001, the average American saved 1.6% of his disposable income, compared to an aggregate personal saving rate of 7% in the period 1959/2001 (US Bureau of Economic Analysis 2002). In 2003, the share of households that declared bankruptcy reached an estimated record of 1.5% (Boushey and Weller 2006). The data report, furthermore, a high number of

individuals still not availing of banking services (Hilgert et al. 2003). Some research also shows that the higher the citizens' level of financial knowledge, the greater their likelihood of acquiring financial products and services as well as making more informed saving and investment choices. Various initiatives have therefore been undertaken to promote the financial inclusion of the poorer sections of the population in order to increase their autonomy.

Since the early 2000s, and particularly after the 2008 financial crisis, governments and financial institutions have insistently highlighted the need for financial education as a matter of first priority, partly with the aim of instilling new trust in markets and financial products. The literature, however, reveals contrasting opinions on the efficacy of the programs. Some studies do not find a direct link between financial education and changes in investment decision-making (Choi et al. 2004; Cronqvist and Thaler 2004). Others, such as Bernheim et al. (2001), identify a direct connection between financial education, higher saving rates, and higher net worth in states with compulsory financial education courses. Other studies suggest that counselling sessions can contribute to reducing the risk of delinquency. In a study carried out on health and retirement, Lusardi (2004) observes that attending seminars is also associated with an increase of both financial and total net worth.

Aims of Financial Education

According to experts like Lusardi and Mitchell (2007, 2011, 2014), the main aim of financial education is to make citizens become more rational economic agents – an essential condition for efficient markets – and also more autonomous in making decisions on purchases. They feel, therefore, that the general deterioration in the financial position of savers has been caused in the first place by the rapid circulation of financial products of greater complexity (e.g., credit cards, revolving credits, subprimes, student loans, payday loans, and tax refund loans), which, thanks to the technological progress (like home banking), have also reached small investors directly, making them

more vulnerable to predatory loans. In the second place, there is an even lower level of familiarity with financial matters, due in part to limited knowledge of mathematics.

Studies testing mathematical/financial skills have, in fact, revealed serious inadequacies in three areas of great importance: compound interest, inflation, and stock risk. Lusardi and Tufano, for example, surveyed 1,000 US residents and found that 36% of the respondents were able to do a compound interest calculation, 35% understood that making minimum credit card payments means essentially never paying the balance off, and only 7% understood that making a \$1,200 payment at the end of a year is more advantageous than making monthly payments of \$100. Similar surveys report that Europeans also need support to understand and make effective use of certain financial products.

In the European countries taking part in the OECD/INFE survey (International Network on Financial Education 2016), with the exception of Estonia, the Netherlands, and Norway, (where between 76% and 80% of the population answered correctly), the calculation of simple interest on savings posed a problem for at least 25% of respondents, rising to over 50% of the population in Albania and the Russian Federation. With respect to compound interest, the large majority of Europeans are unable to understand that the value of interest following 5 years' compounding is more than five times the simple interest. Again, the Netherlands and Norway stand out as exceptions. The concept of risk diversification appears to be even more challenging: over 30% of the population do not understand what it is, in Albania, Austria, Belgium, Croatia, Czech Republic, Estonia, Hungary, Latvia, the Netherlands, Norway, Poland, the Russian Federation, and the United Kingdom.

Cognitive and Behavioral Critique

The data also show that financial education can have the side effect of increasing the citizens' faith in their own financial abilities, even though no improvement has actually been made. Forty percent of the FINRA respondents declared they had

high or very high levels of financial knowledge, despite the fact they were proven to be limited (Lusardi and Tufano 2009). In a survey carried out in Australia, 67% of the respondents declared they knew what compound interest was, but only 28% were able to give the correct answer to a question on the subject (OECD 2006).

Behavioral researchers (i.e., Bernartzi and Thaler 2004) have investigated the possibility that increased financial knowledge has a positive effect on intended behavioral changes but without necessarily bringing them about. Choi et al. (2004) report that 35% of contributors interviewed about pre-retirement saving plans 401(k) declared they intended to make greater savings in the future. In the following 4 months, however, the majority (86%) had not made any changes in their personal saving habits. In a similar way, Clark et al. (2006) show a very weak correlation between good intentions and actual changes in saving habits, presumably caused by the intervention of psychological variables such as the peer effect. In contemporary consumer society, limited interest in savings may be caused by the pressures of conspicuous consumption.

Cognitive and behavioral economists find it useful to study the problems connected to financial education in the context of bounded rationality (Simon 1978) and cognitive bias (Tversky and Kahneman 1981, 1991; Elster 1996; Frederick et al. 2002; Gigerenzer 2007). In contrast with conventional, neoclassical theory, which attributes full rationality to the individual, behavioral finance holds that, given the available alternatives, people are unable to make the optimal choice. This occurs for multiple reasons, including asymmetric information and the fact that the rationality of decisional processes is altered by heuristic-induced, systematic computational errors (i.e., mental shortcuts adopted when faced with the excessive complexity of calculation), as well as personality (risk aversion, tendency to procrastinate) and emotive factors (anxiety, indecision, fear, euphoria).

Such factors are classified by the literature as problems of self-control, e.g., the tendency to overborrow/under-save, procrastinate, and/or manifest overconfidence. Other psychological

phenomena include hyperbolic discounting (e.g., preferring \$10 today to \$12 in a week's time but preferring \$12 in a year and a week to \$10 in a year), herding (imitating the decisions of others as a form of mental shortcut, whether based on the supposition that others know more or simply in order to conform), loss aversion, status quo bias (showing a preference for the status quo even when it does not increase material well-being), the framing effect (when a decision is influenced by the way various options are presented), anchoring (the tendency to base choices on reference points, such as peer advice, which is not objectively relevant), and nonstandard probabilistic thinking (e.g., lottery purchases).

Behavioral economists also show that financial education programs with a good content can fail because of the way in which they are structured or presented. For example, research has shown that women have different preferences for investments and respond differently to choice framing (Croson and Gneezy 2009). To be pigeonholed as poor can make those on low incomes reject a course and refuse to follow it (Ross and Nisbett 2011). Similarly, the elderly have different requirements and different attitudes toward phenomena like risk and uncertainty (Agarwal et al. 2009).

There are, indeed, deep-seated problems inherent in financial education. Firstly, research reveals persistent weaknesses in basic mathematical/financial knowledge. Secondly, it indicates the presence of systematic errors which affect decision-making processes and serve to limit rationality. Lastly, the efficacy of financial education depends largely on how the programs and courses are presented and the kind of clientele they target.

Given the widespread limited knowledge of mathematics, the OECD created in 2012 the Programme for International Student Assessment (PISA), the first large-scale international study, which examines the financial knowledge and competence acquired by 15-year-olds both within and outside school. From the 13 OCSE countries taking part in the initial enquiry, it emerged that only 10% of the students could analyze complex products and solve financial problems of above-average difficulty. The PISA enquiry also highlights the fact that an average student with a more

advantaged socioeconomic profile tends to score a higher mark than one from a poorer background. Such data appear to confirm the importance of financial inclusion as a means of reducing social inequality.

Research on psychological traps and systematic errors – Thaler and Sunstein (2008) – indicates the need for experts, such as brokers and financial advisors, to help citizens to make informed choices. To solve the problems of asymmetric information arising between citizens and experts, they deem it necessary to have policy-makers that safeguard individual choice with specific financial markets regulations, by indicating, for example, default options on mortgages, investment funds, and superannuation plans, which are not obligatory, but which become effective in cases of protracted indecision (i.e., libertarian paternalism).

Conclusion

Many of these problems involving financial education have led various researchers (Gale and Levine 2010; Willis 2011) to the conclusion that financial education is a necessary but ineffective instrument. The factors adduced are:

1. The complexity of financial decisions facing a saver who is not a professional investor.
2. The heterogeneity of consumer financial circumstances and values making it very difficult and expensive to form personalized plans.
3. The speed of change in product offerings and industry practices.
4. Individuals' lack of interest in taking part in programs that require a considerable investment of time and energy.
5. Persistent cognitive and behavioral biases in citizens who recognize the existence of psychological traps (Willis 2011: 429–430).

Finally, Willis highlights the high public costs of financial education programs and the fact that, paradoxically, they may actually limit individual autonomy. In order to keep up with the novelties and complexities of financial markets, compulsory lifelong learning would be required. Governments,

foundations, and international organizations thus need to weigh up the opportunity cost of financial education and remember that the limited efficacy of its programs often does not depend on factors related to the rationality of the individual citizen.

Cross-References

- [Behavioral Law and Economics](#)
- [Bounded Rationality](#)
- [Cognitive Law and Economics](#)
- [Experimental Law and Economics](#)

References

- Agarwal S, Driscoll JC, Gabaix X, Laibson D (2009) The age of reason: financial decisions. *Brook Pap Econ Act* 40(2):51–117
- Bernartzi S, Thaler R (2004) Save more tomorrow: using behavioral economics to increase employee savings. *J Polit Econ* 112(1):164–187
- Bernheim BD, Garrett DM, Maki DM (2001) Education and saving: the long-term effects of high school financial curriculum mandates. *J Public Econ* 80:435–465
- Boushey H, Weller CE (2006) Inequality and household economic hardship in the United States of America. DESA working paper 18
- Choi JJ, Laibson D, Madrian BC, Metrick A (2004) For better or for worse: default effects and 401(k) savings behaviour. In: Wise D (ed) *Perspectives in the economics of aging*. University of Chicago Press, Chicago, pp 81–121
- Clark RL, d'Ambrosio MB, McDermed A, Sawant K (2006) Retirement plans and saving decisions: the role of information and education. *J Pension Econ Financ* 5(1):45–67
- Cronqvist H, Thaler RH (2004) Design choices in privatized social-security systems: learning from the Swedish experience. *Am Econ Rev* 94(2):424–428
- Crosen R, Gneezy U (2009) Gender differences in preferences. *J Econ Lit* 47(2):448–474
- Elster J (1996) Rationality and the emotions. *Econ J* 106:136–197
- Fisher P (2003) Evaluating financial education: history, theory & application. Unpublished master's thesis, Ohio State University, Columbus
- Frederick S, Loewenstein G, O'Donoghue T (2002) Time discounting and time preference: a critical review. *J Econ Lit* 15:351–401
- Gale WG, Levine R (2010) Financial literacy: what works? How could it be more effective? Paper presented at the first annual conference of the Financial literacy research consortium, Washington, DC, 18 Nov 2010
- Gigerenzer G (2007) *Gut feelings: the intelligence of the unconscious*. Viking, New York

- Hilgert MA, Hogarth JM, Beverly SG (2003) Household financial management: the connection between knowledge and behavior. *Fed Reserv Bull* 89(7):309–322
- Lusardi A (2004) Saving and the effectiveness of financial education. In: Mitchell O, Utkus S (eds) *Pension design and structure: new lessons from behavioral finance*. Oxford University, New York, pp 157–184
- Lusardi A, Mitchell O (2007) Financial literacy and retirement preparedness: evidence and implications for financial education. *Bus Econ* 42(1):35–44
- Lusardi A, Mitchell O (2011) Financial literacy around the world: an overview. *J Pension Econ Financ* 10(4): 497–508
- Lusardi A, Mitchell O (2014) The economic importance of financial literacy: theory and evidence. *J Econ Lit* 52(1):5–44
- Lusardi A, Tufano P (2009) Debt literacy, financial experiences, and overindebtedness. NBER working paper 14808
- OECD (2006) The importance of financial education. Policy brief. OECD
- OECD (2016) Financial education in Europe: trends and recent developments. OECD
- Ross L, Nisbett RE (2011) *The Person and the situation: perspectives of social psychology*. McGraw-Hill, New York
- SEC (1999) The facts on saving and investing. Office of investor education and assistance securities and exchange commission
- Simon HA (1978) Rationality as a process and as a product of thought. *Am Econ Rev* 70:1–16
- Thaler R, Sunstein C (2008) *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven
- Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458
- Tversky A, Kahneman D (1991) Loss aversion in riskless choice: a reference-dependent model. *Q J Econ* 106(4): 1039–1061
- US Bureau of Economic Analysis (2002) Data on personal savings as a percentage of disposable personal income
- Willis LE (2011) The financial education fallacy. *Am Econ Rev Pap Proc* 101(3):429–434

Financial Regulation

Hadar Yoana Jabotinsky
Faculty of Law, Hebrew University of Jerusalem,
Jerusalem, Israel

Abstract

Financial markets do have special attributes which require regulatory intervention. They are complex markets which are abundant

with asymmetric information, moral hazard, externalities, and agency problems. They are markets in which products mature over a long period of time causing a need for regulatory monitoring which is exacerbated by consumer demand for regulation and economies of scale in monitoring. Moreover, the financial firms in these markets are crucially important from a systemic point of view to the health of the economy in general. Having said all that, financial regulation is costly. Regulation in general should only be enacted if the costs of implementing it are lower than the benefits derived from what it seeks to achieve. Regulation is not about quantity but about quality. The “right” kind of regulation gives the financial institutions the incentives to act in a way which enhances social welfare and reduces market failures.

Introduction

Regulation tends to disrupt the market process and changes opportunities and costs for entrepreneurial discovery and profits. If a free market is generally a desirable goal from an economic point of view, why not allow it in the financial service sector? If nothing is wrong with the free market, then financial regulation becomes worthless or even harmful. If there is something wrong with it, with regard to the financial sector, then what is it exactly about the financial sector that makes the free market inefficient from an economic point of view? Assuming that the financial sector does require specific regulation, the second question that has to be considered is what are the costs of such regulation? If the costs exceed the benefits of regulating, then regulating is not desirable as it causes social welfare to decrease.

The following pages will attempt to put together a comprehensive list of the reasons for financial regulation and its potential costs. Some of the costs are not quantifiable, but may have a strong impact on the efficiency of the financial regulation; others are quantifiable and are used in the Regulatory Impact Analysis conducted by regulators before issuing a new piece of regulation.

The Building Blocks

What Are the Rationales Behind Regulation?

Traditional economic approach lists three main purposes behind regulating markets (Brunnermeier et al. 2009):

1. Insuring competition, constraining the use of monopoly power, and preventing distortions to the market's integrity
2. Protecting consumers in cases where asymmetric information, which is costly to obtain, might harm them
3. Protecting against externalities where the cost of regulation is lower than the costs of the externalities

This traditional approach to regulation is called **the public interest approach** which assumes that a market economy may produce undesirable outcomes to consumers. A different and more modern approach to regulation, **the self-interest approach**, claims that regulation is made to serve the interest of the regulated group. In other words, the group which stands to benefit and the group which stands to be harmed both have an incentive to influence regulation in order to produce a better outcome for them (Stigler 1971; Peltzman 1976).

These considerations come to play also in the financial markets. However, as the financial sector has a few special attributes which make it more prone for market failures and misuse of consumers, the considerations for regulating the financial market are slightly different than those which exist in markets in general.

When thinking of modern financial regulation, one can identify three main goals:

1. Preventing systemic risk
2. Protecting consumers/investors
3. Helping design a framework for deciding monetary policy and determining exchange rates

The economic rationale for regulation and supervision in banking and financial services has long been known and debated. Generally, the need for financial regulation stems from addressing the concerns and needs listed below (Llewellyn 1999):

- Internalizing externalities
- Reduction of transaction costs for an efficient allocation of financial resources
- Enhancing consumers and investors' confidence and reliance and preventing a race to the bottom of risk management criteria
- Limiting and preventing unwanted herding directions
- Fighting crime and terror (e.g., anti-money laundering regulation)
- Correcting market failures (e.g., information asymmetries and agency problems)
- Achieving economies of scale in monitoring and regulation
- Correcting behavioral biases on behalf of the consumers
- Responding to consumer demand for regulation
- Reducing litigation costs by referring consumer complaints to the financial regulator

These rationales can be divided into two general types of regulation and supervision: prudential regulation and conduct of business regulation.

Prudential regulation assumes that consumers do not have enough information to assess the stability of the institution in which they place their money nor are they in a position to assess its **risk approach** (Armour et al. 2016). In this case, regulation is needed to assure that the financial institution does not take on excessive risk and endanger consumers' savings. Even if consumers are given information at the time contracts are signed, it is not enough to protect them down the road from risky behavior on behalf of the financial firm. If systemic risk factors are taken into account, the need for prudential supervision is paramount. One of the most important roles of financial regulation is to prevent or minimize systemic risks, i.e., reduce externalities. Systemic risk is the risk that an entire system or market might collapse. This risk is exacerbated by links and interdependencies, where the failure of a single entity or cluster of entities can cause a cascading failure (Acemoglu et al. 2015; Committee on Capital Markets Regulation 2009).

Conduct of business regulation focuses on protecting consumers during their day-to-day encounters with financial firms. Such regulation will generally cover proper disclosure rules, fair

treatment of customers, and competence of advisors and other service providers. Generally speaking, conduct of business regulation solves problems arising from asymmetric information and principal-agent relationships and ensures proper conduct when doing business with consumers.

Why Not Use Contracts?

The economic literature considers contracts preferable to regulation, as regulation is generally costly and is likely to yield a less efficient allocation of resources than bargaining. As described by Llewellyn (1999) in the case of financial services, it is likely that contracts will fail. Contract failure has many dimensions, such as (i) agency conflicts which may lead to bad advice to consumers, (ii) insolvency of the supplying firm prior to the delivery of the goods, (iii) mismatch between the consumers' expectations and the product or service delivered, (iv) fraud on behalf of the financial institution, (v) incompetence to supply the product in the expected standard, (vi) misunderstanding of the type of product or of its risk attributes by the consumer, and (vii) behavioral inclinations which offset rational decision making by consumers. As explained above, financial markets are highly complex and are prone to asymmetric information and agency problems. For these reasons, contracts and law enforcement by market participants tend to not be enough to ensure a well-functioning market, and regulatory intervention is needed (Enriques and Hertig 2011).

The Rationales for Prudential Regulation

Prudential regulation can be divided into micro-prudential regulation which concerns itself with the stability of the individual institutions and macro-prudential regulation which is concerned with the stability of the financial system as a whole. In general, the rationale for prudential regulation stems from addressing the following major points:

Reducing Externalities

Unlike the "perfect" market described in the economic literature, financial markets do, when unsupervised, allow for externalities (for a

discussion on externalities and transaction costs in general, see Coase 1960). This is mainly due to the fact that a financial firm takes into consideration solely its own risk without taking into account the risks that society might suffer as a whole from its malfunction (Dodd 2002). The rationale behind regulating externalities is that the true price of the product is not reflected in the price (Baldwin et al. 2012). The results of such externalities became evident during the 2007–2009 financial crisis and the large "bail out" schemes which followed; financial institutions were at large not held responsible for the risks they undertook, and the society as a whole had to pay the price in order to avoid an even larger turmoil. Moreover, as some countries lacked some or all of this money, they had to increase their national debt, which might have negative effects on the economy of these countries in the future such as inflation or fluctuation of currency or reduction of their ability to lend more money if needed (Reinhart and Rogoff 2009). In 2007 – 2009 the excessive risk taking on the US market has spread to nearly all markets around the world affecting them and bringing down firms which, at first glance, did not have anything to do with the excessive risk taking in the US market.

The problem with systemic risk unfolded in financial firms is that even if the risk of collapsing is small, its consequences might be devastating. Even with capital restrictions on some financial institutions, such as banks, they may still produce some externalities as capital requirements may limit the amount of direct exposure to default, but indirect exposure is still prevalent. If capital requirements cannot prevent all externalities, could government guarantees such as deposit insurance reduce concerns with regard to risk-related externalities? The idea behind government guarantees is that consumers should not be forced to face the consequences of actions that were not under their control. However, in order for deposit insurance to protect against a run on the financial institution, the coverage of the insurance has to be 100%. This is not the current situation in most countries (Cecchetti 1999). The problem with the idea of granting insurance coverage for deposits is

that it induces moral hazard problems – banks might be tempted to take on more risks and to operate with less capital. Depositors on the other hand might seek banks who take on more risk as they can receive higher interest rates as long as the bank is solvent and still be compensated if the bank goes bankrupt. But if the deposit insurance is anything short of 100%, the incentive for a run on the bank in specific circumstances remains. The situation can therefore be viewed as a trade-off between preventing bank runs and preventing moral hazard problems. As deposit insurance removes the incentives of liability holders in the financial institution to oversight their financial institutions, there is a need for regulatory intervention which will guarantee that the behavior of the insured institutions is not irresponsible. In a way, financial regulation is expected to bring a cure to their inherent moral hazard problem.

Externalities are also present with regard to pricing of some securities, such as derivatives, OTCs (over the counters), and other securities which are based on an underlying asset. It is thought that the price of securities reflects the risk levels unfolded in the underlying asset. A more “risky” security, i.e., the one which yields more variance, will have a lower price. However, that is only true for direct ownership of the security. The risk unfolded in risky securities extends beyond direct ownership. That extra risk is not priced nor calculated within the price of such securities.

What is special to the type of externalities in the financial market is that they cannot be solved by self-regulation even if the financial institutions agreed to it, as the single financial institution itself is not aware of their magnitude.

Controlling Herding

Prudential regulation is also needed in order to prevent and limit unwanted herding directions. It is thought that investors influence other investors and this influence has a first-order effect (Devenow and Welch 1996). Herding is a concept which is hard to define, yet when we refer to herding in the financial sector context, we refer to it as decision making by entire populations which can lead to systemic erroneous, namely, suboptimal choices. Herding is the power behind

bubbles (for definition of the term, please see Garber 1990), bank runs, noise trading, and other unwanted phenomena in the financial markets which lead to distraction of wealth.

Bankers and other financial employees might also suffer from herding when comparing their actions to the actions of other financial employees in their sector and mimicking them. Thus, in times of crisis, there might be unwanted behaviors on behalf of financial employees (such as shortage of credit in the market due to the fact that one bank decides to cut down on its loans and all other banks react and follow).

Herding does not require coordination, but simply an ability to collect information about what others are doing in the market. There are two views with regard to herding; the first claims that investors/financial employees are not rational and simply behave like cattle in a herd, blindly following the lead of others. The second views investors/financial employees as rational players and puts its focus on externalities: optimal decision making is thought to be distorted by lack of information or suboptimal incentives. Either way, one of the goals of financial regulators is to reduce unwanted herding to a minimum and to deviate the power of herding toward wealth maximizing directions by providing reliable information to the market, monitoring in order to try and prevent the unwanted effects of bubbles which are created due to herding, and solving credit crunches once they have already been formed.

Efficient Allocation of Financial Resources and Strengthening Investors’ Confidence

Prudential regulation is also necessary in order to allocate financial resources efficiently. The financial market and the institutions operating in this market are essential for economic growth. Banks, insurance companies, the stock exchange, and the likes allow for the concentration of savings and for the efficient allocation of these resources to investment projects that generate economic growth. Financial regulators play a crucial role in reducing information asymmetries with regard to products and providing for a quality label for the financial institutions and for the financial stability of the country in which these institutions

operate. This in turn strengthens investors' confidence and allows them to invest not only in the financial institutions but also in the country itself, knowing that, in high probability, their investment will be returned, sometimes with a profit.

Providing Information to the Market

A financial regulator plays an important role in providing information to the market, mainly through disclosure requirements, which in turn helps the market assign the right price tag to the products (Hayek 1945) and prevents the problem of a market for lemons (Akerlof 1970). In some cases, regulation sets minimum standards for products, and by doing so, it helps clean the market from lemons. Minimum standards are also needed in order to prevent adverse selection, i.e., to prevent "good" or "careful" firms from being driven out of the market.

As banks race for higher profits, they drive risk management criteria down, a situation which may lead to the collapsing of the system. Without regulation, the dominant strategy of each bank is not to invest in appropriate risk management due to the fact that risk management is costly as it restrains the business from acting more aggressively and therefore cuts down on short-term profits. The Nash equilibrium is then set on all banks not investing in appropriate risk management and eventually collapsing. Moreover, due to the systemic connections between banks, if one bank behaves irresponsibly and collapses it might bring down other banks, including those banks that have behaved responsibly in managing their risks while giving up on the extra profits attainable from high-risk, high-reward bets. Financial regulation is needed in order to solve this race to the bottom by setting common minimum standards and insuring compliance with the standards. Such standards will not always differ from the standards that would have been set by the industry if each financial institution could insure that its competitors will also follow these standards. In other words, financial regulation is sometimes useful in order to coordinate competitors in situations in which the Nash equilibrium dictates that each firm defects, even though it is in their interest to cooperate.

The Rationales for Conduct of Business Regulation

Over the years, scholars have played with the idea of an "efficient" financial market, i.e., a market in which there are no information gaps or asymmetries, in which the price of securities accurately reflects the value of the firm and in which investors have access to all the relevant and needed information and are able to analyze it properly (Sharpe 1970). In such a market, agency problems, externalities, and moral hazard would not exist. Therefore, in such a market, there would be no need for regulation as financial regulation is costly and therefore should be avoided whenever possible. The rationale for financial regulation, as for all regulation, is to correct such market failures and imperfections.

Asymmetric Information

The problem of asymmetric information and lack of ability to assess the financial product are enhanced by the existence of products which mature over a large number of years. Such products include pension funds, insurance policies, options with a long duration date, saving accounts which are closed for a long period of time, funds, current accounts, etc. Moral hazard issues may come into play causing the firm to behave differently prior to the purchase of the product then post the purchase. There is no way, other than regulation, to prevent this from occurring. For this reason, regulation enforcing disclosure is essential (Kurlat and Veldkamp 2015). But simply providing the consumers with information is not always enough. The existence of complex financial products makes it difficult for unprofessional customers to monitor the financial institution.

Monitoring

One of the goals of financial regulators is to monitor financial enterprises and assist in monitoring investments and management performance in these firms. Financial regulators are better equipped to monitor financial products, also due to the fact that they develop the relevant expertise in monitoring over time. Monitoring is important

in this market as one of the attributes of financial products is the fact that the contracts attached to the products are usually long-term contracts. This in turn creates several problems among which are principle-agent and monitoring problems. Another monitoring role of financial regulators comes into play with reducing information asymmetries with regard to risk. Due to the benefits of economies of scale, expertise, and the high cost of monitoring for private consumers, it is economically rational to leave the responsibility to monitor financial products partially in the hands of the financial regulators (Baldwin et al. 2012).

Consumers' Behavioral Biases

In consumer contracts, sophisticated firms will try and make use of consumers' behavioral biases in order to expropriate more profit. Competitive forces push sellers to take advantage of their consumers. Financial regulation is needed also in order to correct for such bias. The first thing that should be considered when we talk about consumer contracts is the existence of huge asymmetries between the sides of the contract. One such asymmetry is characterized by the existence of behavioral biases on the side of the consumer, while the other side is a sophisticated firm taking advantage of these behavioral human flaws (Bar-Gill 2004; Campbell 2016).

Consumers Demand for Regulation and Low-Cost Dispute Settlement Mechanism

Consumers themselves demand regulation in order to satisfy their need for quality reassurance. Consumers are aware of the fact that financial markets are highly complicated and require a degree of expertise; most consumers are aware of the fact that they themselves do not possess such expertise, and so most consumers would like an external regulator to monitor and set a certain level of standard to the financial industry.

Furthermore, in lack of a financial supervisor, each consumer/investor is left on his or her own to deal with injustices caused to him or her by the financial institution. The existence of a financial regulator provides the consumer/investor with an address to which he or she can turn to in order to

complain about unjust behavior on behalf of the financial institution. This in turn reduces the need to turn to courts in order to solve petty disputes. Moreover, the existence of a financial regulator enables all consumers to complain without distinguishing between them on the base of their wealth, providing them with low-cost dispute settlement mechanism. This in turn induces the financial institutions to treat their consumers fairly.

Interim Summary

Over the years, a number of positive theories have been developed in order to explain how and when does government intervention in markets occur and what drives changes in regulation. A few different approaches have been used to study this issue, one of which relates to "public interest": regulation is essential in order to correct market failures resulting from externalities and to fix the information gaps between the industry and the consumers. From this perspective, regulation is needed in order to enhance social welfare.

A key challenge to this approach lies in the fact that regulation is not always optimally designed to enhance social welfare – there are many cases in which designing the regulation differently would be more beneficial from a social welfare point of view, yet it is not done. Why? The simple answer would be due to costs.

What Are the Costs Associated with Financial Regulation?

There are many different types of costs which prevent the regulator from reaching an optimal solution. When we talk about the optimal design of the financial regulators, it is important that we take these costs into account (Enriques 2014). Costs are also related to the test of proportionality. In order to enact new regulation under any OECD country regime, there should be (i) a public interest which the regulation comes to

advance, (ii) a rule of law enabling the regulator to regulate, and (iii) the regulation should be proportionate to the goal it is trying to achieve. Among the costs of financial regulation, we can list the following:

Capture of the Financial Regulator

Regulation has major distributional effects and is costly to the regulated firms, because it restricts them from operating in a way which maximizes their profits and, if effective, makes them internalize their costs. Therefore, it is in the interest of the financial firms to effect the regulation they will have to comply with and limit what they perceive to be its “damages” to them. This is also known as the “private interest” theory of regulation or the economic theory of regulation (Bernstein 1955; Stigler 1971; Peltzman 1976; Enriques and Hertig 2011). This theory describes the regulatory process as a competition between two interest groups, in which the well-organized, well-coordinated group is able to extract rents at the expense of the more dispersed, less informed groups (Becker 1983). Under this theory, the strong organized interest group is able to capture the regulator and influence its regulation to promote the benefits of the regulated firms. A captured regulator will act according to the interests of the regulated firms rather than in accordance with its mandate which is to promote the common good. From a welfare perspective, a captured regulator might yield one of the most serious costs associated with financial regulation as a captured regulator has the capability to heavily damage the general welfare in order to promote other, sometimes personal, goals.

Cost of Mistakes

Another cost that is related to financial regulation is the cost of a mistake being made by the regulator. If the regulator issues regulation based on wrong perceptions of either a market failure or the approach which is needed to be taken in order to correct it, then such mistake will spread out to the entire regulated market (Romano 2014). This is sometimes referred to by the literature as “macroeconomics distortions.” Such distortions could increase existing market deficiencies and

undermine the objectives intended for the regulation in the first place. In some cases, mistakes in financial regulation may increase the magnitude of a financial crisis or even cause it.

Systemic Risk Arising from Financial Regulation

As discussed earlier, one of the major goals of financial regulation is to prevent systemic risk. However, financial regulation may by itself cause systemic risk if it is deficient and especially if it spreads to a global level. Treaties or global regulations are often a political compromise between the countries involved and thus are not easily adjustable for the needs of a specific market (Romano 2014).

Another source of systemic risk resulting from regulation might stem from a regulatory attitude promoting complex and detailed regulation. This attitude might cause financial institutions to rely solely on the regulation without using common sense to protect against dangers which were neglected by the regulation or unexpected changes in risks on the one hand, and cause consumers, investors, internal auditors, and financial regulators to feel overly confident with regard to the stability of the financial system on the other hand.

Distortion of Competition

Financial regulation often creates barriers to entry. The need for a bank license and collateral requirements are two prominent examples. As regulators are concerned with stability and preventing systemic risks, they might tend to be “overprotective” and put up demands which leave “large margins.” Such an attitude may prevent or discourage new firms from entering the market (Hendrickson 2011). Concerns of systemic risks may lead the financial regulator to keep financial institutions “alive” even in situations where otherwise, given a fully competitive market, would have led to the restructuring or removal of the financial institution from the market. Financial regulation may also interfere with competition within the market itself by demanding exorbitant disclosure – if all information is disclosed there is less room left for competition.

Costs of Fragmentation of the Regulatory Regime

Fragmentation in the context of regulation is a term used in order to describe a situation in which there is more than one supervisory authority active in the market. In such a case, we would consider the market to be “fragmented” from a regulatory point of view as there is more than one regulator imposing regulatory policies and demands on the regulated firms in the market. This situation may lead to severe cooperation issues between the regulators and cause problems with the flow of information between the different financial regulatory bodies (Jabotinsky 2017). Fragmentation incurs costs; the existence of several regulators acting without coordination in the market may lead to inefficiencies and cause regulatory arbitrage on the part of the regulated institutions. This in turn implies the formation of conflict or jurisdiction and lack of regulation or overlapping regulation. Moreover, the existence of several regulators increases the risk of incoherent regulation resulting in uncertainty on the part of market participants.

Others

Administrative costs – Regulatory agencies, like any agency, cost money. As financial regulation is a public good, the financial regulator is financed by the government using public money.

Cost of compliance on behalf of the financial institutions – Regulatory compliance demands place a heavy financial burden on financial institutions, especially on smaller market participants. The cost of regulation exhibits strong economies of scale, sometimes resulting in smaller banks being “penalized” twice as much as larger institutions.

Innovating around the regulator – A profit-seeking firm will go the length to extract more profit from the market; thus, it will go to great efforts to find a way to innovate around the regulation. In some cases, cases which do not provide the market with a new product or service that is materially different from the existing products on the market, innovating around the regulation is costly and does not generate greater social welfare.

Moral hazard resulting from a rigid regulatory regime – A rigid, detailed, and protective regulatory regime might remove the responsibility from the financial institution’s employees and transfer it to the employees of the financial regulator, thus causing the employees of the financial institution to behave recklessly.

Summary

As has been discussed above, financial markets do have special attributes which require regulatory intervention. They are complex markets which are abundant with asymmetric information, moral hazard, externalities, and agency problems. They are markets in which products mature over a long period of time causing a need for regulatory monitoring which is exacerbated by consumer demand for regulation and economies of scale in monitoring. Moreover, the financial firms in these markets are crucially important from a systemic point of view to the health of the economy in general.

Having said all that, financial regulation is costly. Regulation in general should only be enacted if the costs of implementing it are lower than the benefits derived from what it seeks to achieve. That is especially true for the financial markets, as the health of these markets affects the social welfare of society as a whole.

Moreover, with regard to financial supervision, it is crucial that the responsibility for the actions of the financial institutions and for compliance with the laws and regulations remains in the hands of the financial institutions’ employees and management. Leaving the responsibility in the hands of the financial institutions themselves is important also from the aspect of minimizing regulatory mistakes and systemic risk caused by regulation. If financial firms are provided with regulatory guidelines instead of strict rules, this helps in diversifying the market. In the era of global systemic risk, this is crucially important.

Consumers themselves should be entrusted with the responsibility to monitor what their financial institution is doing with their assets. In order to do so, financial regulation should force

financial institutions to provide consumers with easy to understand data.

Regulation is not about quantity but about quality. The “right” kind of regulation gives the financial institutions the incentives to act in a way which enhances social welfare and reduces market failures.

Cross-References

- [Central Bank](#)
- [Insurance Market Failures](#)
- [Law and Finance](#)
- [Lender of Last Resort](#)
- [Market Failure: Analysis](#)

References

- Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2015) Systemic risk and stability in financial networks. *Am Econ Rev* 105(2):564–608
- Akerlof G (1970) The market for lemons: quality uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Armour J, Awrey D, Davies PL, Enriques L, Gordon JN, Mayer C, Payne J (2016) *Principles of financial regulation*. Oxford University Press, Oxford
- Baldwin R, Cave M, Lodge M (2012) *Understanding regulation*. Oxford University Press, Oxford
- Bar-Gill O (2004) Seduction by plastic. *Northwest Univ Law Rev* 98:1373–1375
- Becker GS (1983) A theory of competition among pressure groups for political influence. *Q J Econ* 98(3):371–400
- Bernstein MH (1955) *Regulating business by independent commission*
- Brunnermeier M, Crocket A, Goodhart C, Hellwig M, Persaud AD, Shin H (2009) *The fundamental principles of financial regulation*. 11 Geneva report on the world economy
- Campbell JY (2016) Richard T. Ely lecture restoring rational choice: the challenge of consumer financial regulation. *Am Econ Rev* 106(5):1–30
- Cecchetti SG (1999) The future of financial intermediation and regulation: an overview. In: ‘Why and How Do We Regulate?’ current issues in economics and finance. Federal Reserve Bank of New York, NY
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Committee on Capital Markets Regulation (2009) *The global financial crisis. A plan for regulatory reform* http://www.europarl.europa.eu/meetdocs/2009_2014/documents/d-us/dv/d-us_tgfc-ccmr_executive_summa/d-us_tgfc-ccmr_executive_summary.pdf
- Devenow A, Welch I (1996) Rational herding in financial economics. *Eur Econ Rev* 40:603–615
- Dodd R (2002) Special policy report 12: the economic rationale for financial market regulation. Derivatives Study Center, Washington, DC
- Enriques L (2014) Regulators’ response to the current crisis and the upcoming reregulation of financial markets: one reluctant regulator’s view. *Univ Pennsylvania J Int Law* 30:1147–1155
- Enriques L, Hertig G (2011) Improving the governance of financial supervisors. *Eur Bus Organ Law Rev* 12:357–378
- Garber PH (1990) Famous first bubbles. *J Econ Perspect* 4(2):35–54
- Hayek F (1945) The use of knowledge in society. *Am Econ Rev* 35(4):519–530
- Hendrickson JM (2011) *Regulation and instability in U.S commercial banking, a history of crises* Palgrave. Macmillan studies in banking and financial institutions. Palgrave Macmillan, Basingstoke
- Jabotinsky HY (2017) The federal structure of financial supervision: a story of information-flow. *Stanf J Law Bus Finance* 22:53–84
- Kurlat P, Veldkamp L (2015) Should we regulate financial information? *J Econ Theory* 158:697–720
- Llewellyn D (1999) The economic rationale for financial regulation. FSA occasional papers in financial regulation
- Peltzman S (1976) Toward a more general theory of regulation. *J Law Econ* 19:211
- Reinhart CM, Rogoff KS (2009) *This time is different, eight centuries of financial folly*. Princeton University Press, Princeton
- Romano R (2014) For diversity in the international regulation of financial institutions: critiquing and recalibrating the Basel architecture. *Yale J Regul* 31:1–76
- Sharpe WF (1970) Stock market price behavior. A discussion. *J Financ* 25(2):418–420
- Stigler GJ (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 1:3–21

Financial Risk

- [Risk Management, Optimal](#)

Firm

- [Organization](#)

Fiscal Federalism

Philip C. Hanke¹ and Klaus Heine²

¹Department of Public Law, University of Bern, Bern, Switzerland

²Rotterdam Institute of Law and Economics, Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Abstract

Fiscal federalism is concerned with the optimal level of centralization or decentralization of state activity. The early literature attributed the central government two functions: ensuring allocative and macroeconomic stability on the one hand and assistance of the poor (redistribution) on the other hand. A second generation of fiscal federalism went beyond the realm of public economics and added aspects of political economy and to the debate. In the context of European integration, these questions are of particular relevance, with migration and the sovereign debt crisis being two examples addressed in this entry.

Definition

Fiscal federalism is a subfield of public economics discussing which functions of the public sector and appropriate instruments to carry out these functions are best centralized and which should be allocated to decentralized levels of government (see, e.g., Kenyon and Kincaid 1991; Oates 1999).

Introduction

Roughly one half of the world's population lives in systems of government where sovereignty is constitutionally divided between a central government and constituent political entities such as states or provinces, which in turn can also share their competences with local units such as cities or counties. Some countries, although not

federations de jure, have devolved significant policy-making powers to local or regional levels of government (e.g., France or the United Kingdom). At the same time, the process of European integration is creating a federation sui generis. It can thus be studied how to best allocate functions of the public sector and policy instruments vertically to carry out these functions among the different levels of government.

Broadly speaking, the scholarship on fiscal federalism can be divided into a first and a second generation (Qian and Weingast 1997; Oates 2005). The earliest formulations of the first-generation theory of fiscal federalism (as in Musgrave 1959; Oates 1972) attributed to the central government two functions. One was the basic responsibility for macroeconomic stability. With highly open local economies, decentralized governments cannot affect local levels of employment and prices using the means of monetary policy. The other one was the assistance of the poor. Since individuals and businesses are mobile, redistribution should be done at higher levels of government to prevent net contributors from relocating to a jurisdiction with a lower tax burden. Second-generation economic theories of fiscal federalism expand earlier theories by adding components of political economy (public choice) and political science.

Consequently, this entry follows the historical development from first- (section "[First-Generation Fiscal Federalism](#)") to second-generation federalism (section "[Second-Generation Fiscal Federalism](#)"). Section "[Fiscal Federalism in the EU](#)" specifically discusses federalism within the context of the European Union and gives two concrete examples: the case of migration and the European sovereign debt crisis. Finally, section "[Conclusion](#)" concludes.

First-Generation Fiscal Federalism

The public finance literature in the 1950s and 1960s came to an important finding, namely, that there should be a "separate governmental institution for every collective good with a unique boundary, so that there can be a match between

those who receive the benefits of a collective good and those who pay for it" (Olson 1969, 483). First-generation fiscal federalism is thus primarily concerned with a fundamental trade-off: providing public goods and services through centralized policy-making is suboptimal because of the divergences in local tastes and conditions, while decentralization can lead to inefficiencies because local governments do not fully internalize interjurisdictional externalities.

The Tax Assignment Problem

The tax assignment problem, coined by McLure (1983), addresses the fundamental normative question which taxes are best suited at which level of government. It is assumed that people, goods, and resources can easily move between lower-level jurisdictions, while there is little or no mobility at the higher (e.g., national) level. At the starting point is the Tiebout (1956) model, in which mobile individuals can choose between bundles of taxes and public goods provided by a large number of local communities. Since people have different preferences regarding those bundles, they will move to the community that maximizes their personal utility. Through this mechanism, it is possible to provide public services at a decentralized level. Since the individual preferences are then matched with the local governments supplying exactly the public goods demanded, total welfare is increased. This proposition was formalized as the so-called decentralization theorem in Oates (1972). It follows that taxing mobile individuals through a decentralized system of government is impossible or causes distortions in resource allocation as individuals bear costs to avoid taxation. As the subsequent work of Oates and Schwab (1991) and Oates (1996) shows, thus mobile factors should be taxed with benefit levies. Non-benefit taxes, that is, those that usually have some redistributive effect, should be allocated to the central government. Nevertheless, many local jurisdictions still maintain local redistributive schemes.

The decentralization theorem also faced the question why a central government is inherently unable to produce the optimal outcome. The answer is twofold. First, there is an information

problem: local governments are closer to their constituencies and thus have a better knowledge of their residents' wishes and requirements, as well as of the costs involved in fulfilling them. Secondly, the idea that a central government fine-tunes its policies to the needs of local jurisdictions collides with the general principle of equal treatment at the national level, where a certain character of uniformity is expected from the central government.

Olson's (1969) concept of fiscal equivalence, where local public goods were provided in such a way that the geographical scope of benefits coincided with the boundaries of the jurisdiction financing them, led to a practical problem: designing a federal system without any spillovers between jurisdictions could hardly exist (Oates 2005). Thus, with decentralized provision of local public goods, it is necessary that the central government provides subsidies to the local jurisdictions in order to internalize the benefits of positive externalities between the local governments. Another role for the central level government is less guided by efficiency concerns than it is by equity considerations: by redistribution from rich to poor jurisdictions, it can contribute to the cohesion of economically differently developed regions.

The early literature on the tax assignment problem already recognized the importance of hard budget constraints, that is, that jurisdictions cannot export parts of their tax burden onto other jurisdictions. Otherwise, local budgets would be inflated beyond efficient levels. Nevertheless, vertical transfers among entities in a federal state are common. While some countries have equalizing transfers from richer states or provinces to poorer ones (through the federal level), some other countries such as the United States know them primarily at the level of the states, that is, there are equalizing grants among local jurisdictions. Through such equalizing grants, spillover benefits can be internalized. There are also equity considerations for equalizing transfers, while the verdict is still out on their efficiency effects. It is not clear whether they promote economic growth in poorer regions or whether they inhibit the adjustments necessary for economic development.

Furthermore, they play a role in ensuring a progressive tax system and correct the regressive effects of decentralized taxation.

The more recent literature on fiscal federalism has in further detail emphasized the importance of hard budget constraints (see below).

Economies of Scale

Providing public goods and services comes at a cost. If these costs have a fixed component (or more general have the feature of subadditivity), then decentralized provision means that these fixed costs have to be borne by each jurisdiction providing it and the total costs will be larger than if the central government incurred them. Thus, public goods and services with high fixed costs should rather be centralized, whereas goods and services with low fixed costs can be decentralized. This point can be applied not only to traditional public goods and services but also to law as a public good, which implies strong network externalities and may lead to path dependencies in the law provision (Heine and Kerber 2002).

Laboratory Federalism

Decentralized government can facilitate experimenting with new laws and regulations in order to assess their suitability. In a dissenting opinion to the US Supreme Court's 1932 *New State Ice Co. v. Liebmann* ruling, Justice Louis Brandeis argued in favor of legal experimentation: "It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country" (*New State Ice Co. v. Liebmann*, 285 U.S. 262 at 285). This idea has been taken up by theories of federalism. In this view, lower-level governments can experiment with new policies, which, if deemed successful, can later be implemented at the national level. Oates (1999) gives the examples of unemployment insurance and emissions trading. The recent decision of the Supreme Court on the unconstitutionality of any prohibition of same-sex marriages comes to mind as a more recent example of a case where experimentation at the state level led to subsequent complete adoption at the national level.

The diffusion of new policies does not necessarily have to be vertical, but can also be horizontal. A basic problem with such experimentation is that it is better not to be the first to implement a new policy, but to rather observe other jurisdictions innovate and then free ride on their learning experience. In other words, there is a positive information externality which potentially leads to too little experimentation (see, e.g., Rose-Ackerman 1980; Strumpf 2002). However, there are examples in which jurisdictions are very engaged in horizontal interjurisdictional competition, as the example of the US State Delaware showcases with regard to corporate law (Romano 1993).

Interjurisdictional Competition

With decentralization comes competition among lower-level governments. Through various policy instruments, such as tax rates but also, e.g., environmental regulation, they seek to attract capital investments.

It is still an open question under which conditions horizontal competition among governments is efficiency enhancing and when it is destructive. The models of Oates and Schwab (1988, 1991, 1996) liken interjurisdictional competition to perfect competition in the private sector. Local governments set efficient bundles of public goods and taxes in order to compete among each other for mobile capital. The result is a "race to the top." For these results to hold, a series of assumptions has to hold, namely, that jurisdictions behave as price takers in capital markets, that public officials act in the interest of the common good and are not self-interested, and that all necessary regulatory instruments are available to them in order to carry out the desired policy.

Other models, but also the Oates and Schwab models if the abovementioned assumptions are not met, emphasize the possible outcome of a "race to the bottom," in which competition is detrimental to total welfare. A typical example is the setting of environmental standards. If politicians are primarily concerned with attracting businesses and enlarging the local tax base, then they might set excessively lax environmental standards (Oates and Schwab 1988).

While mobility is an important factor in interjurisdictional competition, it has also been argued that the simple fact of observing policies in other jurisdictions – what has been called “yardstick competition” (Salmon 1987) – can be enough to exert pressure on politicians to implement similar ones. The presence of decentralized government influences decision-making and constrains the set of actions (e.g., the possible tax rates) even if local governments are not in a position strong enough to actively participate in the competition (Kenyon 1997).

Second-Generation Fiscal Federalism

While fiscal federalism is concerned with the optimal level of centralization or decentralization of state activity, it also provides a framework to explain and predict the constitutional arrangements found in practice. Second-generation fiscal federalism thus goes beyond the realm of public economics and draws its inspiration from public choice, information economics, the theory of the firm, organization theory, and contract theory (Oates 2005). It also profits from input from adjacent disciplines such as political science.

Market-Preserving Federalism

On a different level of analysis than the public economics approaches mentioned above, federalism can also be seen as an institution to preserve a market economy with well-defined property rights. In a strain of literature in the tradition of new institutional economics, federalism as a system can be designed as a mechanism limiting how far a country’s political class can encroach on its markets. In a seminal article, Weingast notes a “fundamental political dilemma of an economic system,” namely: “A government strong enough to protect property rights and enforce contracts is also strong enough to confiscate the wealth of its citizens” (Weingast 1995, 1). Property rights and the enforcement of contracts are thus only secure if the state is limited in its ability to confiscate wealth.

In subsequent articles, five conditions are established in order to create and to preserve market-preserving federalism. First, there “exists

a hierarchy of governments with a delineated scope of authority (e.g., between the national and subnational governments) so that each government is autonomous in its own sphere of authority.” Secondly, “the subnational governments have primary authority over the economy within their jurisdictions.” Thirdly, “the national government has the authority to police the common market and to ensure the mobility of goods and factors across subgovernment jurisdictions.” Fourthly, “revenue sharing among governments is limited and borrowing by governments is constrained so that all governments face hard budget constraints.” Finally, “the allocation of authority and responsibility has an institutionalized degree of durability so that it cannot be altered by the national government either unilaterally or under the pressures from subnational governments.” While the first condition is inherent to federalism, the other four ensure its market-preserving character.

In a somewhat similar vein, Inman and Rubinfeld (1998) argue that federalism is next to rights enumerated in the constitution, and the separation of powers between branches of the central government is the third possible line of defense in protecting rights of citizens. The federal form, they argue, can also encourage participation, as individual votes are more likely to be pivotal in small jurisdictions and accessing as well as monitoring politicians is easier.

Self-Interested Agents in Political Processes

An important expansion on the initial literature on the economics of federalism is the insight that the public officials do not necessarily act in the interest of an elusive general interest, but can also act as self-interested actors. The same also applies to voters who also maximize objective functions in the context of political processes. With these assumptions, the principal-agent model becomes the main approach. The relationship between the principal and the agent is characterized by asymmetric information and imperfect monitoring, while the outcome has a stochastic component. The contract between the two thus has to be based on the observed outcome. With multiple levels of government and different societal actors, there are

many possible interpretations of who the principal is and who the agent is. For instance, in what is referred to as “administrative federalism” (Inman 2003), the public officials of the local governments can be considered the agents of a central government, which has certain objectives. In other models of fiscal autonomy, the principal-agent conflict of interest is between the electorate and elected officials, and elections constitute incomplete contracts with unverifiable information. The conclusion from these models is usually that decentralization improves accountability and control (see, e.g., Seabright 1996; Tommasi and Weinschelbaum 2007).

Information Problems

The early literature on fiscal federalism already highlighted the information problem present in policy-making and understood it as an argument in favor of decentralization. Second-generation fiscal federalism elaborates a bit further on where this information comes from. At first, there is the question why the central government cannot, through various means, acquire detailed information on local needs. As Cremer et al. (1996) point out, such activities are costly, and obtaining such information is more important and valuable to local public officials than it is to those of the central government. This creates the fundamental problem that it is generally assumed that the central government has no significant issues collecting information on the needs regarding public goods and services provided at the national level. Additionally, the information problem creates the question how the central government is then able to set the Pigouvian subsidies to compensate local governments for positive interjurisdictional spillover effects. It can only do so if it has accurate information on the valuation of the spillovers at the local level. Nevertheless, it is argued (e.g., in Oates 2005) that an outcome with imperfect subsidies is still better than one where externalities are ignored altogether.

Functional, Overlapping, Competing Jurisdictions (FOCJ)

A proposal for a new kind of federalism lies in the concept of functional, overlapping, competing

jurisdictions. Advocated in the late 1990s by Bruno Frey and Reiner Eichenberger (1999), the government is divided into organizations called FOCUS. The idea is that every such entity is functional, that is, it deals with a specific, delimited matter, such as education, public safety, or infrastructure. As a result, there are several, overlapping FOCJ covering certain individuals or regions. Individuals can then choose which FOCUS applies to them for which policy field. If the function is territorially bound, then a local entity (“commune”), such as a town, determines democratically which FOCUS it belongs to. There is thus competition between FOCJ. A FOCUS, once chosen, has the power to levy taxes. It is thus a jurisdiction.

The advantage of these FOCJ, so the argument, is that the concept combines four important aspects of the economic theory of federalism. First, fiscal equivalence can be achieved. Secondly, the boundaries are well defined so that new members of these clubs can be charged optimally, namely, at the marginal costs. Thirdly, the voting by foot mechanism is enabled through exit and entry. Finally, there is political competition through democratic processes and thus a “voice” mechanism.

Such a system that relies less on geographical units can be found in several federal settings around the world and in history. The United States knows the concept of “special districts,” Swiss local units can and do build functional and overlapping special communes, Germany knows “special purpose associations” (*Zweckverbände*), and the European Union carries strong traits of FOCJ as well (e.g., the monetary union, the Schengen Area, and others).

Fiscal Federalism in the EU

There are two phenomena taking place at the same time in Europe. First, countries have devolved some competences from the central government to local or regional levels (e.g., the creation of *régions* in France and devolved governments in the United Kingdom), thus creating more decentralized policy-making. Secondly, there is,

at the same time, the process of European integration, which gradually allocates competences to the European institutions. What might appear as contradictory at the first glance can very well be interpreted as a more nuanced approach to fiscal federalism.

The Principles of Conferral, Subsidiarity, and Proportionality

The structure of European federalism is laid out in the Treaty on European Union (TEU): “The limits of Union competences are governed by the principle of conferral. The use of Union competences is governed by the principles of subsidiarity and proportionality” (Art. 5(1) TEU). Under this principle, “the Union shall act only within the limits of the competences conferred upon it by the Member States in the Treaties to attain the objectives set out therein. Competences not conferred upon the Union in the Treaties remain with the Member States” (Art. 5(2) TEU).

In other words, the EU does not have the authority to appropriate competences to itself. This distinguishes it from other federations, where in some fields the national legislature has the so-called competence-competence (the power to assign competences) and can reassign competence between the federal and the sub-federal level through a regular law and not necessarily through a constitutional amendment. The EU’s competences are conferred through the TEU; thus, any change would require a treaty change, which is an intergovernmental procedure involving all 28 Member States and therefore lengthy and inert. This inertia also leads to a problem in the other direction: once the EU has a certain competence, it is very difficult to devolve policy responsibilities from the supranational back to the national level (Kirchner 1998).

Article 5(3) states: “under the principle of subsidiarity, in areas which do not fall within its exclusive competence, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level...”

Finally, the principle of proportionality requires that “[...] the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties [...]”

Current Challenges

Migration

In the European Union, migration is regulated at several levels of government. While the freedom of movement of EU citizens is a fundamental freedom well entrenched in EU law, admission of third-country nations (TCNs) is primarily left to the member states. The *Schengen Area* has a joint visa policy, and the *Dublin system* constitutes a joint framework for a decentralized system of examining applications for asylum.

The decentralized treatment of TCNs entails significant restrictions on their mobility within the EU. For instance, a recognized refugee is not allowed to relocate to a different EU Member State within the first 5 years of residence in the country that granted the refugee status. In this sense, the current migration regime challenges the common assumption in economic theories of federalism of complete mobility within a federation. Asylum seekers, who according to the *Dublin system* are required to request asylum in the EU member state that they first set foot in, might thus be stuck in a place that is not optimal from an allocative perspective (e.g., if their skills are more sought after in a different Member State).

The European Sovereign Debt Crisis

During the debt crisis that started in 2009, several Eurozone member states (Cyprus, Greece, Ireland, Spain, and Portugal) were unable to fulfill their obligations toward their creditors. They thus required assistance of the European Central Bank (ECB), the International Monetary Fund (IMF), and the European Financial Stability Facility (EFSF), a special purpose vehicle established by the EU member states in 2010, which was replaced by the European Stability Mechanism (ESM) in 2013.

In terms of fiscal federalism, the debt crisis showed that the hard budget constraint in the relationship between member states and the EU as a federal system of open economies could not

be upheld. This does not necessarily mean that large-scale transfers among member states through the EU institutions are not economically justified. Furthermore, fiscal federalism theories in some instances do recognize the presence and necessity of interjurisdictional transfers.

Tax Harmonization

The European Union is currently discussing harmonizing taxes and tax bases within the Union. VAT harmonization has already taken place with the intention of creating a level playing field for firm operation across the EU. The currently principal legal source is Council Directive 2006/112/EC of 28 November 2006 on the common system of value-added tax. The EU harmonized not only VAT law but also limited the number of different VAT rates the Member States can set for various types of goods: there is a minimum 15 % “standard rate,” and countries can apply two reduced rates of at least 5 % on certain goods. The list of goods eligible for the reduced rates is set at the EU level. From a fiscal federalism point of view, it is debatable why the EU should regulate the VAT rates set by the Member States as it is not clear what the externalities in a nonregulated setting would be.

Another ongoing project is the establishment of a so-called Common Consolidated Corporate Tax Base (CCCTB). This proposal, which was developed by the European Commission, aims at formulating mandatory common rules on how to calculate the tax base and at allowing companies to report their tax results consolidated at the European level. The individual Member States would then still set the actual tax rate. From a federalism theory point of view, this measure seems to reduce the information problems present in intra-European trade while not limiting the ability of jurisdictions to set their own tax rates. However, one may conceive the CCCTB also as a first step toward the “cartelization” of tax policies of the Member States to the detriment of tax payers.

Conclusion

Taking the insights from economic theories of federalism seriously means to recognize that

both centralizing and decentralizing ideologies are wrong. As Olsen pointed out, “there is a case for every type of institution from the international organization to the smallest local government” (Olson 1969, 483). Theories of fiscal federalism can contribute to assign specific competences to the appropriate level of government. Second-generation federalism added to this insight that the separation of powers across different vertical levels of government is per se a precondition to sustain a functioning market economy that honors property rights.

References

- Cremer J, Estache A, Seabright P (1996) Decentralizing public services: what can we learn from the theory of the firm. *Rev Econ Polit* 106(1):37–60
- Frey BS, Eichenberger R (1999) The new democratic federalism for Europe: functional, overlapping, and competing jurisdictions. Edward Elgar Publishing, Cheltenham
- Heine K, Kerber W (2002) European corporate laws, regulatory competition and path dependence. *Eur J Law Econ* 13:47–71
- Inman RP (2003) Transfers and bailouts: enforcing local fiscal discipline with lessons from U.S. federalism. In: Rodden JA, Eskeland GS, Litvack J (eds) *Fiscal decentralization and the challenge of hard budget constraints*. MIT Press, Cambridge, MA, pp 35–83
- Inman RP, Rubinfeld DL (1998) Subsidiarity and the European union. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*. Macmillan, London, pp 545–551
- Kenyon DA (1997) Theories of interjurisdictional competition. *N Engl Econ Rev* 13–36
- Kenyon DA, Kincaid J (1991) Competition among states and local governments: efficiency and equity in American federalism. The Urban Institute, Washington, DC
- Kirchner C (1998) Competence catalogues and the principle of subsidiarity in a European constitution union. *Constit Polit Econ* 8:71–87
- McLure C (ed) (1983) *Tax assignment in federal countries*. Australian National University, Canberra
- Musgrave RA (1959) *The theory of public finance*. McGraw-Hill, New York
- Oates WE (1972) *Fiscal federalism*. Harcourt Brace Jovanovich, New York
- Oates WE (1996) Taxation in a federal system: the tax-assignment problem. *Public Econ Rev* 1:35–60
- Oates WE (1999) An essay on fiscal federalism. *J Econ Lit* 37(3):1120–1149
- Oates WE (2005) Toward a second-generation theory of fiscal federalism. *Int Tax Public Financ* 12(4): 349–373

- Oates WE, Schwab R (1988) Economic competition among jurisdictions: efficiency enhancing or distortion inducing? *J Public Econ* 35(April):333–354
- Oates WE, Schwab R (1991) The allocative and distributive implications of local fiscal competition. In: Kenyon DA, Kincaid J (eds) *Competition among states and local governments*. The Urban Institute, Washington, DC
- Oates WE, Schwab R (1996) The theory of regulatory federalism: the case of environmental management. In: Oates WE (ed) *The economics of environmental regulation*. Edward Elgar Publishing, Aldershot, pp 319–331
- Olson M (1969) The principle of ‘fiscal equivalence’: the division of responsibilities among different levels of government. *Am Econ Rev* 59(2):479–487
- Qian Y, Weingast BR (1997) Federalism as a commitment to preserving market incentives. *J Econ Perspect* 11(4):83–92
- Romano R (1993) *The genius of American corporate law*. AEI Press, Washington, DC
- Rose-Ackerman S (1980) Risk taking and reelection: does federalism promote innovation? *J Leg Stud* 9:593–616
- Salmon P (1987) Decentralization as an incentive scheme. *Oxf Rev Econ Policy* 3(2):24–43
- Seabright P (1996) Accountability and decentralization in government: an incomplete contracts model. *Eur Econ Rev* 40:61–89
- Strumpf KS (2002) Does government decentralization increase policy innovation? *J Public Econ Theory* 4(2):207–241
- Tiebout CM (1956) A pure theory of local expenditures. *J Polit Econ* 64(5):416–424
- Tommasi M, Weinschelbaum F (2007) Centralization vs. decentralization: a principal-agent analysis. *J Public Econ Theory* 9(2):369–389. <https://doi.org/10.1111/j.1467-9779.2007.00311.x>
- Weingast BR (1995) The economic role of political institutions: market-preserving federalism and economic development. *J Law Econ Org* 11(1):1–28

Fiscal System

Cécile Bazart
CEE-M – Univ Montpellier – CNRS – INRA –
SupAgro, University of Montpellier,
Montpellier, France
Laboratoire Montpellierain d’économie théorique
et appliquée (LAMETA), University of
Montpellier 1, Montpellier, France

Synonyms

Tax structure; Tax systems; Taxation

Definition

Fiscal systems gather all the taxes and contributions levied to fund the State. These taxes and contributions share the characteristics of being compulsory and unrequited payments to the government, in the OECD sense. Historically, fiscal systems have shown their scalable nature as they are embedded in larger economic, social, political, and cultural systems. It is possible to distinguish the positive analysis and the normative analysis of fiscal systems. The positive analysis enlightens fiscal systems characteristics in terms of tax structure, while the normative analysis questions the qualities of a *good* tax system.

Positive Analysis of Fiscal Systems

Fiscal systems are large collection processes aiming to fulfil the imperatives of financial returns, efficiency, and fairness associated with State intervention, as summarized by Musgrave in 1959, with the three-function framework. The first function, of resource allocation, is to see that resources, generated through taxes, are levied and used efficiently. The second function, of revenue redistribution, deals with the fair distribution of income. The third function of stabilization is to ensure the achievement of high employment, price stability, and growth. Positive analysis of fiscal systems describes their tax structures. From the nineteenth century, the observed growth of the State, as a result of the economic development and greater public intervention, give tracks to explain the extension and diversification of tax structures. Comparison of taxation patterns among countries enlightens the high degree of diversity existing in fiscal systems. Still, groups of fiscal systems with common characteristics can be distinguished. An immediate split would be between developed countries’ fiscal systems and those of developing countries. Developing countries’ systems differ due to limited economic activity and weak accountability standards that result in a restricted set of usable tax bases and low administrative power.

From the beginning of the twentieth century, a similar diversity in the types of taxes levied is to be

found in developed countries. This implies diversity both in terms of the tax base (commodity tax or excise, land taxes, custom duties, and income tax) and tax schedules (lump sum tax, proportional or progressive rates, brackets, exemptions). Still, both the world wars and the development of the welfare State by the middle of the twentieth century have changed the dimension of fiscal systems, with a large and permanent increase in taxes, especially income taxes. This growth of State activities is described by the Wagner law (Bird 1971), as well as the displacement and ratchet effect (Peacock and Wiseman 1961). Over the 1950s and the 1960s, in turn appeared the VAT tax that was implemented progressively in European countries and most developed countries. In the end, current OECD countries' tax systems show common trends in terms of the greater variety in the types of taxes used to collect a higher percentage of GDP of public funds; however, there also exists some degree of country-specific heterogeneity, notably in the total tax pressure and the share of tax revenue derived from each type of tax. Effectively, OECD countries can be separated into a low-tax group and a high-tax group, (Fig. 1, OECD 2017). All rely on the same set of taxes: income taxes, property tax,

consumption taxes, social contributions but with different weights. Another common feature is the fact that they take into account the specificities of family structures, while the way to deal with this issue varies widely across countries. Below table gives an overview of the different weights OECD countries attach to each type of tax. It shows that major differences are to be found in the weight given to social security contributions, overall weight and types of consumption taxes used, importance of capital taxation (corporate and property taxes).

Normative Analysis of Fiscal Systems

The normative analysis of fiscal systems details the imperatives for a *good* fiscal system and as a consequence throws another light on the trend towards diversification of tax structure. A *good* tax system should be able to collect private money in the most efficient but also fair manner, without weakening the legitimacy of the tax, so that there is no rise in noncompliance. If these conditions are met, optimality is reached if not, tax reforms become necessary to avoid resistances, and, at worse tax revolts.

in 2015	% GDP	Tax revenue as % of total tax revenue							% of total tax revenue	
		taxes on income		Social security contributions	taxes on property	taxes on consumption		all other taxes	Taxes on capital	taxes on consumption
		individual	corporate			VAT	Other			
Denmark	45.9	55.2	5.6	0.1	4.1	20	11.6	3.4	9.7	31.6
France	45.2	18.9	4.6	37.1	9	15.3	9.1	6.1	13.6	24.4
Belgium	44.8	28.3	7.4	31.9	7.8	15	8.8	0.8	15.2	23.8
Finland	43.9	30.2	4.9	28.9	3.3	20.6	11.8	0.3	8.2	32.4
Austria	43.7	24.1	5.2	33.6	1.3	17.7	9.6	8.4	6.5	27.3
Italy	43.3	26	4.7	30.1	6.5	14.2	13.1	5.4	11.2	27.3
Sweden	43.3	29.1	6.9	22.4	2.4	20.9	7.2	11.1	9.3	28.1
Netherlands	37.4	20.5	7.2	37.8	3.8	17.6	12	1.1	11	29.6
Germany	37.1	26.5	4.7	37.6	2.9	18.8	9	0.5	7.6	27.8
Luxembourg	36.8	24.5	11.9	29	8.9	17.6	7.9	0.3	20.8	25.5
Greece	36.4	15	5.9	29.4	8.5	20.1	19.2	1.8	14.4	39.3
Portugal	34.6	21.2	9	26.1	3.7	24.8	13.6	1.6	12.7	38.4
OECD average	34	24.4	8.9	25.8	5.8	20	12.4	2.7	14.7	32.4
Spain	33.8	21.3	7	33.8	7.7	19	10.7	0.5	14.7	29.7
UK	32.5	27.7	7.5	18.7	12.6	21.2	11.7	0.5	20.1	32.9
Japan	30.7	18.9	12.3	39.4	8.2	13.7	7.3	0.3	20.5	21
Australia	28.2	41.5	15.3	0	10.7	13	14.5	5	26	27.5
Switzerland	27.7	31.1	10.8	24.3	6.7	12.4	9.3	5	17.5	21.7
United States	26.2	40.5	8.5	23.7	10.3	0	17	0	18.8	17
Ireland	23.1	31.6	11.3	16.8	6.4	19.7	12.9	1.2	17.7	32.6

Source: OECD Revenue Statistics 2017.

Fiscal System, Fig. 1 OCDE tax system structures

Nevertheless, one difficulty is that efficiency and equity principles may oppose one to the other.

A good tax system is an efficient one. The efficiency principle states that the levy should provide the required level of resources to the State, under the constraint of the neutrality principle, that is, it should avoid distorting the initial economic decisions (Salanié 2011). Yet, as a matter of fact, taxes always impact economic decisions and resource allocation if they alter work incentives and generate a substitution of work with leisure. The same will hold for saving, investment, or innovation decisions. By doing so, taxes generate an excess burden measured through the deadweight loss, which is a loss of well-being due to these substitution effects. As a consequence, it is important to be careful about any effect that taxes have on economic efficiency. This is needed to decrease the deadweight loss associated with tax implementation but also because of its potential impact on growth and the macroeconomic equilibrium. Each type of tax instrument results in a specific excess burden, as detailed in the optimal taxation literature. Under the constraint of feasibility and enforcement, this may guide tax structure evolution. For instance, the lump-sum taxation, which is entirely disconnected from the economic decisions of an agent, benefits from a major advantage in terms of efficiency, but is rarely implemented because of equity concerns.

A good tax system is a fair one. The fiscal equity principle implies a fair distribution of the tax burden among the taxpaying population. It questions the initial distribution of the tax burden but also the final one resulting from taxpayers' *tax shifting* strategies (Prest 1955). *Tax shifting* corresponds to the shift of the tax burden from the legal taxpayer, as pointed out in the tax law, to a final and effective tax bearer. This analysis is the core of the incidence theory (Fullerton and Metcalf 2002). Thus, to go deeper into the analysis of fiscal system fairness, it is necessary, on the one hand, to estimate the tax burden any taxpayer should face and, on the other hand, to define principles of a fair distribution of the burden among a large set of heterogeneous taxpayers. Fiscal theory states that there are two ways to estimate the tax burden. The first one is to charge

each taxpayer with taxes equivalent to the benefits he derived from public services and public goods. This refers to the *benefit approach* that goes back to Smith (1776/1976) and that was refined in the twentieth century by economists like, among others, Wicksell (1896/1958), and according to which a social contract links citizens and the State. Nevertheless, implementation of such a criterion is complex due to the very nature of public goods, which are characterized by non-excludability, nonrivalry, and indivisibility. A second option is to distribute the tax burden according to taxpayers' tax capacity, that is, their ability to pay for public expenses.

The implementation of this second principle raises, in turn, several difficulties and questions about tax capacity estimates and the criteria for the fair repartition of the tax burden based on them. Ability to pay tax depends on the economic power of taxpayers and there are at least three relevant indicators to measure it: income, wealth, and consumption. In the heterogeneous population of taxpayers, one may well encounter, side by side, an employee and an annuitant, whose tax capacities may be estimated on the basis of income and wealth, respectively. For such reasons, and as each indicator would lead to a different distribution of the tax burden among the population, all fiscal systems combine several fiscal bases. This constitutes another factor justifying the evolution of fiscal systems' structure toward increasing – and to some degree, natural – complexity. At this stage, the question of the fair distribution of the burden among taxpayers still remains. Two rules deal with the thorny question of taxpayers' relative situations. The first one is known as the horizontal fairness principle and states that those taxpayers that are characterized by the same tax capacity should be treated identically. The second one, the vertical fairness principle, is more controversial and states that taxpayers who differ in their tax capacities should be treated differently on the basis of what is considered as the *acceptable level* of unfairness. For such reasons, fiscal systems are tributary of each country's culture and history.

Away from this subjective consideration, the chosen tax structure gives the big picture showing

which taxpayer is supposed to bear the tax burden while incidence theory details and concludes on the effective burden, once *tax shifting* of any kind has displaced it on to the final and effective tax bearer. One direct example of tax shifting could be any increase in selling price justified by an increase in corporate tax. But ways and means to transfer the burden are diverse and marked by a constant innovation, which is the fiscal dimension of the taxpayers' reactions. These tax transfers are nevertheless sensitive not only to economic determinants (labor market structure, demand, and supply elasticity) but also to fiscal technicity, that is, to the type of tax (direct, indirect) and the tax schedule (rates, brackets, exemptions, deductions). Planning, optimizing, avoiding, and evading taxes are also ways to decrease the burden up to, and sometimes beyond, legal boundaries, (Bazart 2002). It undermines efficiency and equity of the system by shifting the burden on somebody else and by constraining collection. In the end, it justifies reforms, such as: increase in some tax rates, increase in collection through specific taxes and tax bases, and introduction of new taxes.

Institutions, Laws and Tax Compliance

The traditional theoretical approach of optimal taxation has offered numerous contributions to the analysis of optimal taxes, relevant tax rates, and in the end desirable structures of tax systems (Mirrlees 1971; Atkinson and Stiglitz 1976; Atkinson 1991). Nevertheless, from the 1970s, the change in economic conditions under the constraint of a bidding budget constraint, and increasing public deficits, has given more importance to preoccupations about tax base erosion, such as tax evasion or avoidance. Reducing the tax gap became a policy option per se.

For such reasons, fiscal systems analysis has developed quickly on new grounds by adding administrative and enforcement dimensions to the debate to handle the diverse aspects of behavioral reactions to taxation. These were initially gathered under the wide definition of tax evasion

and soon under the even wider concept of non-compliance. From the initial work of Allingham and Sandmo (1972) on, decision to evade taxes was modelled as a risky decision. This decision depended on individual characteristics, among which the degree of risk aversion, and on environmental characteristics such as, at first, harshness of deterrent policies and quickly after a large set of factors such as: unfairness of the system, complexity of the system, uncertainty on rules, use of direct democracy. . . . On the one hand, a direct causal link was established between evasion and complexity, uncertainty and unfairness of the system. On the other hand, an immediate set of results supported the deterrent power of audit rates and penalties; but this has, afterward, been mitigated, in time as research provided new insights. Audit rates and penalties have to exist, but increasing deterrence harshness beyond some threshold is not possible and in fact unnecessary. Effectively, there are counter-productive effects to penalization. Observing first that deterrence's effects are nonlinear, decreasing above some threshold values of penalties and/or audit rates and second that procedural unfairness increases evasion (Rato and Gemmel 2012), it became clear that deterrence levels should be kept low. Kirchler (2007) offers a complete and synthetic analysis of the enforcement problem, known as the slippery slope framework. In this theory, enforcement is accepted under a high level of trust in the government, while in the opposite case, it generates resistances. As a consequence, besides deterrence, the concerns about complexity of tax law and fiscal systems as a whole has witnessed an increase in importance in the debates in the Western democracies.

This emphasized the need to focus on the terms of implementation and enforcement of the levy. Besides efficiency and equity considerations, the third pillar of administration is being given greater concern and importance. This refers to a twofold set of costs: administrative costs borne by the tax authorities and linked to its enforcement efforts (Martin (1944) and compliance costs supported by taxpayers while fulfilling their fiscal duties (Slemrod and Sorum 1984). All these costs result, for one part, from the level of transparency,

consistency, ease in understanding of fiscal laws and procedures. In sum, simplicity of the system favors taxpayers' compliance and acceptance of their liability while the reverse holds for complexity of the system, laws, and procedures. Empirical supports to these assertions are numerous, and it is interesting to note that Alm et al. (2010) obtains, experimentally, an increase in tax noncompliance when the tax law is complex; a decrease if in such a context of complexity, the tax administration acts as a facilitator and a provider of services to taxpayer-citizens. Country studies have also shown the importance of simplicity on tax compliance, for instance, Cuccia and Carnes (2001), Marcuss et al. (2013) for the USA, or Blaufus et al. (2017) for Germany. If the evolution of tax systems implies a natural level of complexity, to some point simplification becomes necessary. In western democracies, many reforms are initiated in line with this: for instance, in the UK, where a Simplification Office was experimented in the last decade or in France where withheld income tax is to be introduced. In all cases, reducing compliance and/or administrative costs is at stake.

Tax law thus has to be easily understandable to reinforce legitimacy of the tax, to avoid procedural unfairness, and provide, efficiently, a help by its expressive nature. That is to say that deterrent parameters also stand as a signal of the norm prevailing in the considered society. Their low level does not explain compliance as demonstrated in 1992 by Alm et al. Still they stigmatize that noncompliance is not acceptable and that the rule to follow is to pay due taxes. Norms activation impacts tax decisions (Myles and Naylor 1996; Bazart and Bonein 2014; Lefebvre et al. 2015), and it is crucial to give the good signal about prevailing social norm. Indirectly in sum, simplification decreases the tax enforcement efforts of the tax authority and improves relationships between taxpayers and the collector. As a matter of fact, it seems to be a win-win strategy as in the end fiscal systems are subject to the judgement of those taxpayers who bear the burden and fiscal history provides various testimonies about tax revolts and resistances to the levy when its legitimacy is weakened.

Conclusion

Tax collection is based on the consent of members of a community to fund the public sector. This consent is expressed through the vote of tax law and the act of fulfilling fiscal duties. Still, taxes are also marked from the very beginning by an underlying climate of resistance evident in the use of terms, such as tax liability, duty, and burden. Moreover, as taxpayers can react strategically to taxes on economic, political, and fiscal grounds, they distort efficiency, fairness, and the simple administration of the system. The extent of behavioral economic analysis of the taxation and fiscal system has shown the deep impact that these behavioral responses to taxation (Alm 2012) have on the system's stability and evolution. It underlines the limitations of reasoning only in terms of optimality of tax systems, as the classical optimal tax theory does. On the contrary, it states that there is room for manoeuvre to reconsider, beyond the analysis of the optimal tax system, the gains to be obtained from a correct appraisal of administrative and compliance costs, in line with the most recent tendency to go towards simplification, tighter enforcement efforts against evasion, and more collaboration with the tax authorities (Slemrod and Gillitzer 2013). The tax system in such a case will gain from a balanced consideration of its multidimensional nature.

Cross-References

► [Tax Evasion by Firms](#)

References

- Allingham M, Sandmo A (1972) Income tax evasion: a theoretical analysis. *J Public Econ* 1:323–338
- Alm J (2012) Measuring, explaining, and controlling tax evasion: lessons from theory, experiments and field studies. *Int Tax Public Financ* 19(1):54–77
- Alm J, Jackson B, McKee M (1992) Deterrence and beyond: toward a kinder gentler IRS. In: Slemrod J (ed) *Why people pay taxes: tax compliance and enforcement*. The University of Michigan Press, Ann Arbor, pp 311–329

- Alm J, Cherry T, Jones M, McKee M (2010) Taxpayer information assistance services and tax reporting behavior. *J Econ Psychol* 31(4):577–586
- Atkinson AB (1991) *Modern public finance*, vol I, II. Edward Elgar, Aldershot
- Atkinson A, Stiglitz J (1976) The design of tax structure: direct versus indirect taxation. *J Public Econ* 6:55–75
- Bazart C (2002) Les comportements de fraude fiscale. Le face à face contribuables administration fiscale. *Rev Fr Econ* 16:171–212
- Bazart C, Bonein A (2014) Reciprocal relationships in tax compliance decisions. *J Econ Psychol* 40(C):83–102. Special issue dynamics of tax evasion
- Bird RM (1971) Wagner's law of expanding state activity. *Public Financ* 26:1–26
- Blaufus K, Hechtner F, Mohlmann A (2017) The effect of tax preparation expenses for employees: evidence from Germany. *Contemp Account Res* 34(1):525–554
- Cuccia AD, Carnes G (2001) A closer look at the relation between tax complexity and tax equity perceptions. *J Econ Psychol* 22(2):113–140
- Fullerton D, Metcalf G (2002) Tax incidence. In: Auerbach A, Feldstein M (eds) *Handbook of public economics*. Elsevier, Amsterdam, pp 1787–1872
- Kirchler E (2007) *The economic psychology of tax behaviour*. Cambridge University Press, Cambridge
- Lefebvre M, Pestieau P, Riedl A, Villeval MC (2015) Tax evasion and social information: an experiment in Belgium, France, and the Netherlands. *Int Tax Public Financ* 22(3):401–425
- Marcuss R, Contos G, Guyton J, Langetieg P, Lerman A, Nelson S, Schäfer B, Vigil M (2013) Income taxes and compliance costs: how are they related? *Natl Tax J* 66(4):833–854
- Martin J (1944) Costs of tax administration: nature of public costs. *Bull Natl Tax Assoc* 29:104–112
- Mirrlees JA (1971) An exploration in the theory of optimum income taxation. *Rev Econ Stud* 38(2):175–208
- Musgrave R (1959) *A theory of public finance*. McGraw Hill, New York
- Myles GD, Naylor RA (1996) A model of tax evasion with group conformity and social customs. *Eur J Polit Econ* 12:49–66
- OECD (2017) Revenue statistics 2017: tax revenue trends in the OCDE. <https://www.oecd.org/tax/tax-policy/revenue-statistics-highlights-brochure.pdf>. Accessed 10 Jan 2018
- Peacock AK, Wiseman J (1961) *The growth of public expenditures in the UK*. Princeton University Press, Princeton
- Prest AR (1955) Statistical calculations of Tax burdens. *Economica* (New Ser) XXII:85–88, 234–245
- Rato M, Gemmel N (2012) Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *Natl Tax J* 65:33–57
- Salanié B (2011) *The economics of taxation*. MIT Press, Cambridge, MA/London
- Slemrod J, Gillitzer C (2013) *Tax systems*. MIT Press, Cambridge
- Slemrod J, Sorum N (1984) The compliance cost of the U.S. individual income tax system. *Natl Tax J* 37:461–474
- Smith A (1776/1976) *An inquiry into the nature and causes of the Wealth of Nations*. University of Chicago Press, Chicago
- Wicksell K (1896/1958) A new principle of just taxation. In: Musgrave RA, Peacock AT (eds) *Classics in the theory of public finance*. St. Martin's Press, New York, pp 72–118

Fixed Investment

Rafael Llorca-Vivero

University of Valencia, Valencia, Spain

Abstract

In a wide sense, the concept of “fixed investment” refers not only to investment in physical capital but also to expenses directed to other intangible assets such as high qualified labor, R&D, the acquisition of particular knowledge about markets or consumers' behavior, advertising, etc. Fixed investment is directly connected with the fixed cost of production of a firm, which is independent of the evolution of output by definition. The consequent emergence of economies of scale and the consideration of irreversible fixed costs (sunk costs) as a strategic variable for market competition (barriers to entry) have led to the development of new theories in the fields of Law and Economics, International Trade or Industrial Organization.

Definition

This expression refers to those expenditures (private or public) directed to purchase inputs that remain in the production process over relatively long periods of time (typically, until complete depreciation) and which are independent of the level of output.

The concept along the years

Traditionally, in economics the term “fixed investment” refers to outlays directed to the acquisition

of tangible assets (i.e., physical capital, which is normally denoted by “K” in formal models). This is typically one of the two generic inputs (the other is labor) considered in the production function. In this category, it is usually included structures and equipment if our analysis is exclusively focused on firms (micro level) and also infrastructures if we refer to the gross fixed capital formation of a territory (macro level). The difference between gross and net values is depreciation. The term “investment” refers to the addition of new assets during a period of time: It is a “flow.” The overall quantity of this factor of production over total employment (stock of capital to labor ratio) as well as its quality (technology) are essential determinants of labor productivity and, then, of economic growth and social welfare. Therefore, fixed investment has a double influence on the economy: a short-term impact through the generation of additional demand and a long-term effect via the increase in potential output.

The essential meaning of the word “fixed” in this context is that it represents an amount of money directed to those components that remain in the production process over time, normally until they are amortized. That is, these inputs are not easily adaptable to the particular circumstances of the evolution of output and, as such, are considered independent on its volume. This is the reason why the concept of “fixed investment” is usually amplified to include other more specialized expenses directed to intangible assets such as high qualified labor, investment on R&D, the acquisition of particular knowledge about markets or consumers’ behavior, and advertising. In fact, there is a subtle distinction in the literature between “fixed costs” and “sunk costs” when referring to this kind of investment. The latter generally refer to outlays that are focused on very specific economic activities and, therefore, that are not susceptible of alternative uses. That is, “sunk costs” are fixed costs that usually occur only once and which are irreversible.

In sum, the cost of production of a firm is composed by those expenditures that are directly related to the level of production and which are called *variable costs* (typically, low qualified labor, energy, raw materials, etc.) and those

related to fixed investment called *fixed costs*. In economic models, when modeling the *cost function*, the marginal cost of production is denoted by “c” (and it is multiplied by the variable reflecting the level of output) and the fixed cost of production is denoted by “F.” The relative weight of fixed costs compared to variable costs in the cost function of firms is generally considered as one of the key determinants of market structure. In fact, in the presence of fixed cost and constant marginal cost, the average (or unitary) cost of production is continuously decreasing as the level of output increases, a circumstance which provides advantage to larger firms over smaller ones thus leading to the appearance of imperfect competition. This occurs more likely in those sectors which require relatively higher amounts of fixed investment in the production process. These internal economies of scale are present in industries such as chemical, machinery, telecommunications, energy, vehicles, iron, and steel, etc.

The decision to invest in *permanent* assets (tangible or intangible) must be the result of an optimization problem, as it is the norm in economics. In this sense, there are two well-known analytical methods in order to choose the most appropriate project. The first one is the so-called net present value (NPV). The NPV of an investment is defined as the discounted value of the cash flows (profits plus amortizations) generated for the project over time. That is, the expected cash flow obtained from the investment in a given year in the future must be converted to present value using a discount factor. The summation of the corresponding amounts for the foreseeable number of years of life of the project has to be compared with the initial (fixed) investment. Those projects with positive NPV are susceptible to be accepted and the preferable will be the one with the highest NPV. The other method of evaluation of investment projects is the so-called internal rate of return (IRR). This is defined as the discount rate that makes the NPV of a project be equal to zero. Those projects with rates of return higher than the alternative in financial markets (typically, the interest rate of the riskless asset) are acceptable, and the best will be the one with the highest rate of return. This explains why when central banks

increase the interest rates of reference in the economy, the expected outcome is a reduction of capital formation by firms (private investment): Under these circumstances the number of profitable projects decreases. When we analyze public projects of investment (essentially, infrastructures) the approach is basically the same and it is called *cost-benefit analysis* (CBA). The main difference with private projects is that public managers have to take into account social costs and benefits. In this case, the issue is how to assign the corresponding money value. As before, those projects with the discounted present value of profits higher than the discounted present value of costs are potentially eligible.

The aforementioned economic tools of analysis can be useful in the ambit of law and economics. The existence of required initial fixed investment in business activities which are the result of a contractual relationship between parties can be studied under this perspective. One example is the franchise contract (see García-Herrera and Llorca-Vivero (2010)) in which the franchisor imposes to the franchisee not only the initial fixed investment (with the proper characteristic of *sunk cost*) in order to maintain the standard quality of the network but also other relevant variables such as resale prices. In this context, contract duration is going to be an essential element of adjustment to the equilibrium (for a general analysis about duration contracts, see also Guriev and Kvassov (2005)). Therefore, in order to reach the equilibrium, a positive correlation should be observed between the amount of the compulsory fixed investment and the duration of the franchise contract conditioned to other variables which are specific to the industry or to the business nature (i.e., price-cost margins, experience).

The consideration of the existence of fixed costs in production has led to the development of new theories in some branches of economics. For instance, the traditional trade theory was only capable to explain interindustry trade, that is, international trade that takes place in different industries among countries. This occurs because countries differ on technology (Ricardo's theory) or resources' endowment (Heckscher-Ohlin

theorem). Or, in other words, countries trade because they have comparative advantage in different types of goods (industries). However, the fact that the majority of trade among developed countries, which are similar in these two characteristics (technology and resources), is of an intraindustrial nature (trade of different varieties within the same industry) leads to the necessity of developing a new paradigm. The "new trade theory," formulated by Krugman (1979, 1980), centers the analysis in the existence of internal scale economies and product differentiation. In a monopolistic competition model, in which international trade is a way of expansion in market size, consumers have access to a higher number of varieties in a given industry at a lower price. This benefits obtained from world trade occur because firms are able to exploit economies of scale by means of output' expansion given the existence of fixed costs in production. That is to say, each country specializes in a reduced range of varieties. Obviously, the number of varieties in equilibrium is lower than the summation of existent varieties within countries in autarky but higher than the number of varieties consumers have at their disposal in the respective countries when no trade occurs.

More recently, the consideration of the existence of "sunk cost" as barrier to entry in international markets has led to further developments in modeling international trade. These "entry cost" are those required to profitable sales in foreign markets such as acquisition of knowledge about consumers' tastes or regulations (technical barriers), advertising, and the establishment of distribution channels. The theoretical and empirical models considering these sunk entry cost are known as the "new-new trade theory." This expression is normally used by reference to the influent Melitz (2003) article although some other authors previously emphasized the role of sunk cost in exporting as, for instance, Baldwin and Krugman (1989) or Roberts and Tybout (1997). Melitz (2003) notes, in a model with heterogeneous firms, than only the more productive ones are able to enter into export markets, giving an explanation to the observed pattern that just

a small fraction of domestic firms become exporters, a characteristic which is common across countries. The reason is that the fixed investment necessary to enter into foreign markets acts as a minimum threshold to be surpassed. Only those firms with expected net profits from exporting greater than this threshold will become exporters. In this sense, a theoretical and empirical distinction is made between the extensive margin of trade (which makes reference to the number of exporters) and the intensive margin of trade (which focuses in the volume of exports of these exporters). This line of research has demonstrated (see, for instance, Dixit (1989)) that the existence of sunk cost (for entry or exit) generates hysteresis in exporting, that is, prior experience is a determinant factor for current participation in exports markets. However, it seems that these fixed investments that generate such sunk costs rapidly depreciate after firm's entry and, therefore, are irrecoverable in the event of firm exit.

In a similar manner, irreversible fixed investment (sunk cost) has revealed as a strategic variable for firms' competition in the recent industrial organization theory. John Sutton (1991), in a work which encapsulates previous research (i.e., Shaked and Sutton (1983, 1987)) obtains relevant conclusions about market structure making a distinction between *exogenous* and *endogenous sunk cost*. In this context, the acquisition of a single plant of minimum efficient scale is considered as an *exogenous sunk cost* for firms given that the achievement of scale economies is conditioned by the existing technology. Decision variables for firms as advertising and R&D outlays (among other possibilities) are *endogenous sunk costs*, whose effect is the increase in consumers' willingness-to-pay (in Sutton's words) for the firm's product. The author demonstrates that, under the presence of *endogenous sunk costs*, the different industries will present a lower bound for market concentration, which is not affected by the increase in demand. This result is valid under very general assumptions about the nature of oligopoly models (number of products offered, type of price competition, etc.). This fact justifies why a high level of concentration persists in some industries

and it contradicts the previous general belief in the sense that market concentration indefinitely declines as the size of the markets increases, leaving this outcome as a special case in which *endogenous sunk costs* are nil. In a similar manner, an incumbent firm can preempt entry of potential competitors by investing in "sunk costs" such as advertising. This action will reduce the expected profits of these firms avoiding their entrance.

Cross-References

- [Cost-Benefit Analysis](#)
- [Franchise](#)

References

- Baldwin R, Krugman PR (1989) Persistent trade effects of large exchange rate shocks. *Q J Econ* 104:635–654
- Dixit A (1989) Hysteresis, import penetration, and exchange rate pass-through. *Q J Econ* 104:205–228
- García-Herrera A, Llorca-Vivero R (2010) How time influences franchise contracts: the Spanish case. *Eur J Law Econ* 30:1–16
- Guriev S, Kvassov D (2005) Contracting on time. *Am Econ Rev* 96:1369–1385
- Krugman PR (1979) Increasing returns, monopolistic competition and international trade. *J Int Econ* 9:469–479
- Krugman PR (1980) Scale economies, product differentiation, and the pattern of trade. *Am Econ Rev* 70:950–959
- Melitz MJ (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71:1695–1725
- Roberts MJ, Tybout JR (1997) The decision to export in Colombia: an empirical model of entry with sunk costs. *Am Econ Rev* 87:545–564
- Shaked A, Sutton J (1983) Natural oligopolies. *Econometrica* 51:1469–1484
- Shaked A, Sutton J (1987) Product differentiation and industrial structure. *J Ind Econ* 36:131–146
- Sutton J (1991) *Sunk Costs and market structure: price competition, advertising, and the evolution of concentration*. MIT Press, Cambridge, MA

Follow-Up Right

- [Droit de Suite](#)

Food Safety

Christophe Charlier

Department of Economics, Université Côte d'Azur, CNRS, GREDEG, Nice, France

Abstract

Food safety is a quality of food. Its originality is to be implemented through a mix of public regulations and private safety control schemes. These measures can interfere with free trade of food, especially when a country wishes to reach a food safety level higher than the level ensured by international food safety standards. Different important dispute rulings in this area have created jurisprudence and an important debate on the WTO SPS Agreement.

Definition

Food safety is a quality attribute of food supplied to consumers. Food safety covers the hygiene of foodstuffs, and the public regulations as well as the private measures implemented in supply chains to reach food innocuousness.

In the wake of food-borne disease crises, food safety has emerged as an important focus for public authorities and agri-food supply chains operators. Viewed by economist as a market failure, food safety requires public policies. Regulations address the food safety issue mainly with a combination of mandatory sanitary standards, liability regimes where the achievement of a sanitary goal is left to the discretion of food operators, and specific food operators' responsibilities. This situation has been considered as a determinant for the rise of food safety private standards. Food safety standards can be constraining for international trade, creating disputes ruled under the SPS Agreement. The different rulings of the WTO Panel and the Appellate Body have created jurisprudence and an important debate on this WTO Agreement.

Introduction

Food-borne diseases crises have made food safety a focus for public authorities and agri-food supply chains operators. Examples are numerous and include bovine spongiform encephalopathy in the 1990s, the 2011 European E coli outbreak, the 2008 Chinese melamine crisis, the numerous contaminations of food with dioxin, etc. Negligence, pollution, modification of production patterns, and diversification of supply sources for raw materials have been pointed out as different reasons for these episodes.

Economists conceive food safety disorder as a market failure (Antle 1996). Food safety is a quality characteristic that cannot be discovered by consumers before purchase. In such a situation, free market conditions cannot guarantee that consumers find the safety level they are paying for. Suppliers' reputation (especially in case of large firms) and private certification have been, respectively, considered as possible motive and way to alleviate this information shortcoming. However, the importance of food-borne diseases and the fact that food safety cannot be provided to consumers without cooperative efforts of many independent operators along the food chains (producers, processors, and distributors) clearly show the need for regulation on efficiency grounds. Two important characteristics follow. First, food safety is implemented through a mix of public regulations and private safety control schemes where the former can be considered as minimal regulatory standards. Second, even if capture of the regulator by economic interest groups should not be discarded, the necessity to regulate and the determination of the socially desirable level of food safety are based on health standards with little space for cost-benefit consideration.

All these measures taken in a risk management perspective can interfere with international trade of food. In some cases a food safety standard may be qualified as nontariff barriers to trade. Therefore, compatibility of food safety public regulations and private measures with international rules of the World Trade Organization (WTO) Sanitary and Phytosanitary (SPS) Agreement is to be sought. The SPS Agreement does not specify an

exclusive sanitary regulation mode and refers to the Codex Alimentarius (Codex) as the relevant standard-setting institution. However, observance of its principles and of the guidelines of the Codex has created a general framework for sanitary regulation organized around three main areas (see Jackson and Jansen 2010 for a discussion): risk assessment, risk management, and risk communication. The risk assessment stage requires that any sanitary regulation be justified by a risk evaluation. The risk management stage demands that food safety standards should be taken with reference to an acceptable level of risk and should be proportionate to this level of risk. Finally, the risk communication stage requires the delivery of information about the risk and the measures undertaken to manage it to the interested parties.

Food Safety Regulations

The regulation addresses the food safety issue mainly with a combination of “command and control” regulatory tools and liability regimes (Henson and Caswell 1999). With command and control regulations, authorities enact mandatory sanitary standards, conceived as minimal quality standards, and examine their good implementation. Banning products considered as dangerous is an extreme form of this kind of regulation. By contrast, liability is a more incentive approach: authorities only fix a sanitary goal, its achievement being left to the discretion of food operators.

Command and Control Regulations

Regulatory tools in the field of food safety are various. Some directly address the sanitary risks looking at the production processes or at the quality of the output. The Hazard Analysis Critical Control Points (HACCP) method is one of these. It requires controls at certain points in the production process previously defined as the most critical for sanitary risks. Food operators from the European Union (EU), for example, are asked to observe “relevant hygiene requirements” set by the Regulation (EC) No. 852/2004, among which procedures based on the HACCP principles are present. Maximum Residue Limits (MRLs)

for pesticides is another example, where a quality standard expressed in terms of a maximum presence of pesticide in products has to be met. In the EU, for example, the use of MRLs for pesticides is required for fresh products of plant and animal origin by the Regulation (EC) No. 396/2005.

Some other regulatory tools address information disclosure rather than intrinsic production processes’ or food’s sanitary risk. Two complementary requirements can be met here. The first one is product labeling. Product labeling is intended to fill the information gap on the sanitary quality of food. It focuses on the foodstuffs’ content, declaring the presence of an allowed component which is however suspected by consumers to be risky (GMO, meat geographical origin, etc.), or which is risky for certain kind of consumers (possible allergic reactions to a component, for example). The second regulatory tool is traceability. Traceability is defined as the ability to identify and trace the history and location of a product. It does not signal food safety but is considered as a risk management tool permitting accurate withdrawals of products from the food supply chain in case of a sanitary risk occurrence. Traceability systems can be different according to three dimensions (Golan et al. 2004): breadth (the amount of information delivered), depth (how far back information record is made), and precision (the tracking unit used). The Regulation (EC) No. 178/2002, for example, demands that all input to the food production process should be tracked, requiring operators to be able to identify “the business from which the food, feed, animal or substance that may be incorporated into food or feed has been supplied.” Breadth and depth of such a system are therefore maximal, but the precision is very weak since no requirement on batch formation appears. Charlier and Valceschini (2008) show that with these characteristics, the mandatory traceability alone can hardly reach the goal assigned by the regulation of a targeted withdrawal of products.

Liability and Operators’ Responsibility

Food operators may be held liable when a safety problem occurs and causes health damages to consumers. Liability is often qualified as an ex

post regulation since it does not involve *ex ante* mandatory standards but only the general goal of food safety to compel with. Liability is therefore intended to provide incentives to food operators to control food safety through private risk management. Two broad liability regimes can be distinguished. Under negligence rules, an operator is held liable only if a fault is proven (due care to safety issue has not been developed). With strict liability, negligence does not have to be demonstrated. An important part of law and economics literature initiated by Shavell (1987) investigates the compared incentive properties of these two regimes. In a food supply chain composed of many operators, distributors pay special attention to liability since they are in a direct relation with final consumers. This situation creates strong incentives to control food safety not only at the distribution stage but also at upstream stages of the food chain.

Together with liability, food safety regulation stipulate food operators' responsibilities that may create incentives to develop private safety schemes in order to be able to meet the expected obligations. For example, Regulation (EC) No. 178/2002 defines food operators' responsibilities with regard to safety procedures required in case of a sanitary breakdown: food withdrawal, information delivery, and cooperation with public authorities are covered. Charlier and Valceschini (2008) argue that these responsibilities ask for "proactive" risk management practices requiring a capacity of food operators to trace their products more precisely than with mandatory traceability. The latter should therefore be considered as a minimum standard eventually completed with more stringent private practices.

In this perspective, liability has been considered as a good lever for "co-regulation" of food safety issues (Garcia Martinez et al. 2007) that appears when public regulation is coordinated with private initiatives. The relation between public regulation and private initiatives is more complex than a one-way relation (Henson 2008). Private initiatives make easier compliance with regulatory aims, but they also infuse public relations (traceability and HACCP are good examples).

Private Standards

The development of *ex ante* food safety regulations requiring operators' "proactive" risk management gives rise to food safety private standards (among other reasons like improving supply chain management and product differentiation when coupled with labeling) especially in Europe. These standards have the particularity to be more demanding than the mandatory ones in order to ensure enhanced levels of food safety and manage liability more effectively proving due diligence. Very often, retailers provide leadership in this area (GLOBAL GAP is an example). Their position at the junction between the food supply chain and the consumers, their size, and their strategy to develop reputation are complementary explanations. The performance of private standards insuring food safety, in situations where public regulations can hardly operate (because of the complexity of the food chains, their internationalization, etc.), is recognized (see Fagotto 2014 for a discussion).

This form of self-regulation poses potential problems. Private standards are often prerequisites for doing business decided downstream (retail stage) to be imposed upstream (in farms and along the supply chain), increasing therefore costs of upstream operators. A parallel with "soft" law can be done. Without having any legal character, private standards have legal-like practical effects (Henson 2008). Furthermore, multiple private standards can be socially costly so that harmonization should be promoted (see the Global Food Safety Initiative). Finally, as seen in the next section, private standards for food safety can be puzzling for international trade.

Food Safety and International Trade

Globalization of the food supply chains in a context where food safety standards are not harmonized makes necessary the control of risks at the borders. When the national standards implemented are recognized by the Codex, they cannot be considered as barriers to trade. However, product rejection and withdrawal carried out in these circumstances are costly. The 2013 "Rapid Alert

System for Food and Feed Report” of the EU shows that fresh products from developing countries are mainly concerned. This observation gives rationale for technical and financial assistance of developing countries to ensure their participation to the international trade of food and feed products (Unnevehr 2007).

National food safety standards can be more demanding than the Codex’s ones. This situation occurs when a country wishes to reach a food safety level higher than the level ensured by international food safety standards. These national food safety standards can be constraining for international trade and considered by other countries as barriers to trade. The disagreement arising in such circumstances over the legality of the national standards is ruled under the SPS Agreement principles. These principles and their interpretation made by different WTO Panels and the Appellate Body allow a country to choose a food safety standard higher than the corresponding international one. But they also require a proper risk assessment in order to demonstrate the need for such a standard. The latter should also be proportionate to the level of risk, in order to avoid being more constraining for trade than necessary. The primacy given to scientific justification by the SPS Agreement is more demanding than the traditional nondiscrimination GATT principle. This raises an important debate on the SPS Agreement and the weight we should give to other criteria such as collective preferences, or cost-benefit analysis (Trebilcock and Soloway 2002).

The most difficult situation arises when the risk is not firmly asserted (i.e., where only a presumption of risk is furnished), presenting the food safety standards as a “precautionary” measure. The ruling of emblematic SPS cases like *EC – Hormones* and *EC – Approval and Marketing of Biotech Products* shows that precautionary SPS measures cannot override the risk assessment requirement. Even if incomplete, a risk assessment has to be done. The SPS measure decided in such a case must be provisional and reevaluated with a “more objective” risk assessment “within a reasonable period of time.”

An original situation arises with private standards. In 2005, St. Vincent and the Grenadines

complained at the WTO SPS Committee that EUREPGAP’s SPS exigencies were stricter than the relevant regulatory ones. This launches a debate on whether private standards are more trade-diverting and trade-reducing than public ones (Henson 2008, Hobbs 2010) and on whether the WTO has jurisdiction over private standards. Trade diversion is feared because of compliance costs that can be too high for developing countries, thus favoring suppliers from countries with higher public standards. Trade-reducing or trade-enhancing property of food safety private standards largely depends on the degree of specificity of the investments required (and therefore on the harmonization of the different private standards). The SPS Agreement addresses governments’ regulations. However, its Article 13 requires members to ensure that private entities implementing SPS measures comply with the Agreement. Private standards are therefore puzzling for international commercial trade and are one of the subjects currently debated in the SPS Committee. This debate might be long. In March 2014, countries participating to the SPS Committee still disagreed on the definition to give to private standards.

Cross-References

- [Labeling](#)
- [Market Failure: Analysis](#)
- [Market Failure: History](#)
- [Risk Management, Optimal](#)
- [Traceability](#)

References

- Antle JM (1996) Efficient food safety regulation in the food manufacturing sector. *Am J Agric Econ* 78: 1242–1247
- Charlier C, Valceschini E (2008) Coordination for traceability in the food chain. A critical appraisal of European regulation. *Eur J Law Econ* 25:1–15
- Fagotto E (2014) Private roles in food safety provision: the law and economics of private food safety. *Eur J Law Econ* 37:83–109
- Garcia Martinez M, Fearnle A, Caswell J, Henson S (2007) Co-regulation as a possible model for food safety governance: opportunities for public–private partnerships. *Food Policy* 32:299–314

- Golan E, Krissoff B, Kuchler F, Calvin L, Nelson K, Price G (2004) Traceability in the US food supply: economic theory and industry studies. Agricultural Economic Report 830, USDA, Economic Research Service
- Henson S (2008) The role of public and private standards in regulating international food markets. *J Int Agric Trade Dev* 4:63–81
- Henson S, Caswell J (1999) Food safety regulation: an overview of contemporary issues. *Food Policy* 24:589–603
- Hobbs JE (2010) Public and private standards for food safety and quality: international trade implications. *Estey Centre J Int Law Trade Policy* 11:136–152
- Jackson LA, Jansen M (2010) Risk assessment in the international food safety policy arena. Can the multi-lateral institutions encourage unbiased outcomes? *Food Policy* 35:538–547
- Shavell S (1987) Economic analysis of accident law. Harvard University Press, Cambridge, MA/London
- Trebilcock M, Soloway J (2002) International trade policy and domestic food safety regulation: the case for substantial deference by the WTO dispute settlement body under the SPS agreement. In: Kennedy DM, Southwick JD (eds) *The political economy of international trade law. Essays in honor of Robert E. Hudec*. Cambridge University Press, Cambridge, pp 537–574
- Unnevehr LJ (2007) Food safety as a global public good. *Agric Econ* 37:149–158

Forensic Science

Marie-Helen Maras¹ and Michelle D. Miranda²

¹John Jay College of Criminal Justice, City University of New York, New York, NY, USA

²Farmingdale State College State, University of New York, Farmingdale, NY, USA

Abstract

Forensic science applies natural, physical, and social sciences to resolve legal matters. The term forensics has been attached to many different fields: economics, anthropology, dentistry, pathology, toxicology, entomology, psychology, accounting, engineering, and computer forensics. Forensic evidence is gathered, examined, evaluated, interpreted, and presented to make sense of an event and provide investigatory leads. Various classification schemes exist for forensic evidence, with some forms of evidence falling under more than one scheme. Rules of evidence differ between

jurisdictions, even between countries that share similar legal traditions. This makes the sharing of evidence between countries particularly problematic, at times rendering this evidence inadmissible in national courts. Several measures have been proposed and organizations created to strengthen forensic science and promote best practices for practitioners, researchers, and academicians in the field.

Definition

Forensic science involves the application of the natural, physical, and social sciences to matters of law.

Introduction

Forensic science refers to the application of natural, physical, and social sciences to matters of the law. Most forensic scientists hold that investigation begins at the scene, regardless of their associated field. The proper investigation, collection, and preservation of evidence are essential for fact-finding and for ensuring proper evaluation and interpretation of the evidence, whether the evidence is bloodstains, human remains, hard drives, ledgers, and files or medical records.

Scene investigations are concerned with the documentation, preservation, and evaluation of a location in which a criminal act may have occurred and any associated evidence within the location for the purpose of reconstructing events using the scientific method. The proper documentation of a scene and the subsequent collection, packaging, and storage of evidence are paramount. Evidence must be collected in such a manner to maintain its integrity and prevent loss, contamination, or deleterious change. Maintenance of the chain of custody of the evidence from the scene to the laboratory or a storage facility is critical. A chain of custody refers to the process whereby investigators preserve evidence throughout the life of a case. It includes information about: who collected the evidence, the manner in which the evidence was collected, and all individuals who took possession

of the evidence after its collection and the date and time which such possession took place.

Significant attention has been brought to the joint scientific and investigative nature of scene investigations. Proper crime scene investigation requires more than experience; it mandates analytical and creative thinking as well as the correct application of science and the scientific method. There is a growing movement toward a shift from solely experiential-based investigations to investigations that include scientific methodology and thinking. One critic of the experience-based approach lists the following pitfalls of limiting scene investigations to lay individuals and law enforcement personnel: lack of scientific supervision and oversight, lack of understanding of the scientific tools employed and technologies being used at the scene, and an overall lack of understanding of the application of the scientific method to develop hypotheses supported by the evidence (Schaler 2012). Another criticism is that some investigators (as well as attorneys) will draw conclusions and then obtain (or present) evidence to support their version of events while ignoring other types of evidence that do not support their version or seem to contradict their version (i.e., confirmation bias). Many advocates of the scientific-based approach believe that having scientists at the scene will minimize bias and allow for more objective interpretations and reconstructions of the events under investigation.

A scene reconstruction is the process of putting the pieces of an investigation together with the objective of reaching an understanding of a sequence of past events based on the physical evidence that has resulted from the event. The scientific method approach is the basis for crime scene reconstructions, which includes a cycle of observation, conjecture, hypothesis, testing, and theory. The process of recognizing, identifying, individualizing, and evaluating physical evidence using forensic science methods to aid in reconstructions is known as criminalistics. Here, identification refers to a classification scheme in which items are assigned to categories containing similar features and given names. Objects are identified by comparing their class characteristics with those

of known standards or previously established criteria. Individualization is the demonstration that a particular sample is unique, even among members of the same class. Objects are individualized by their individual characteristics that are unique to that particular sample (De Forest et al. 1983). Other important concepts in criminalistics include the comparison of objects to establish common origin using either a direct physical fit method or by measuring a number of physical, optical, and chemical properties using chemistry, microscopy, spectroscopy, chromatography, as well as a variety of other analytical methods. Furthermore, in forensic science, exclusion can be as critical as inclusion. Being able to compare materials to determine origin may rule out potential suspects or scenarios.

Forensic Evidence

Forensic scientists examine firearms, toolmarks, controlled substances, deoxyribonucleic acid (DNA), fire debris, fingerprint and footwear patterns, and bloodstain patterns (to name a few). Forensic evidence is collected, processed, analyzed, interpreted, and presented to: provide information concerning the corpus delicti; reveal information about the modus operandi; link or rule out the connection of a suspect to a crime, crime scene, or victim; corroborate the statements of suspects, victims, and witnesses; identify the perpetrators and victims of crimes; and provide investigatory leads.

Evidence classification schemes include: physical evidence, transfer evidence, trace evidence, and pattern evidence. Physical evidence includes objects that meaningfully contribute to the understanding of a case (e.g., weapons, ammunition, and controlled substances). Transfer evidence refers to evidence which is exchanged between two objects as a result of contact. Edmond Locard had formulated this exchange principle, stating that objects and surfaces that come into contact will transfer material from one to another. Trace evidence is evidence that exists in sizes so small (i.e., dust, soil, hair, and fibers) that it can be transferred or exchanged between two surfaces

without being noticed. Pattern evidence refers to evidence in which its distribution can be interpreted to ascertain its method of deposition as compared to evidence undergoing similar phenomena. This type of evidence can include imprints, indentations, striations, and distribution patterns. Criminalistics is concerned with the analysis of trace and transfer evidence and can include, but is not limited to, pattern evidence (fingerprints, footwear, gunshot residue), physiological fluids (blood, semen), arson and explosive residues, drug identification, and questioned documents examination. Questioned documents examination includes the evaluation and comparison of handwriting, inks, paper, and mechanically produced documents, such as those from printers.

Alternate classification schemes for evidence include: direct evidence, circumstantial evidence, hearsay evidence, and testimonial evidence. Many of these terms can be used interchangeably for a given type or sample. Direct evidence refers to evidence that proves or establishes a fact. Circumstantial evidence is evidence that establishes a fact through inference. Hearsay evidence refers to an out-of-court statement that is introduced in court to prove or establish a fact. Depending on a country's rules of evidence, this type of evidence may or may not be admissible in court. Countries, such as the United States, have stipulated in what circumstances hearsay evidence may be admissible (U.S. Federal Rules of Evidence). Testimonial evidence refers to evidence given by a lay or expert witness under oath in a court of law.

Forensic scientists, specifically laboratory analysts and individuals that have testified as expert witnesses, have come under much scrutiny and have been the subject of criticism for a variety of reasons. Some of the criticisms of these laboratory analysts include: the lack of understanding of the science and technology behind their tests and instruments employed (making them more akin to technicians rather than scientists), pro-law enforcement and pro-prosecution tendencies (especially for those individuals working for labs directly affiliated with state, government, or law enforcement agencies), the tendency to testify beyond their knowledge or expertise, the

falsification of credentials, the lack of laboratory quality assurance policies or the misunderstanding or misapplication of the quality assurance practices in place, data falsification, overstating the value or weight of the evidence, and the misuse of statistics (Moenssens 1993).

Domestic rules of evidence stipulate the criteria used to determine the competency of eyewitnesses and experts to testify. Limits on the admissibility of purportedly scientific evidence also exist by requiring a judge to ensure that an expert's testimony is both valid and reliable (e.g., U.S. Federal Rules of Evidence).

Rules of Evidence

Domestic rules of evidence vary between countries. Rules of evidence dictate the type of information that can be collected from computers and related technologies. These rules also proscribe the ways in which evidence should be collected in order to ensure its admissibility in a court of law. In order for evidence to be admissible in court, it must first be authenticated. This evidence must also be collected in a manner that preserves it and ensures that it is not altered in any way. The key to authenticating evidence is the maintenance of the chain of custody.

Formal and informal information sharing mechanisms are used to facilitate cooperation between countries in criminal investigations. Formal information sharing mechanisms include multilateral agreements, bilateral agreements, and mutual legal assistance treaties between countries. The latter requires each party to the treaty to provide the other party with information and evidence about crimes included in the treaty. By contrast, informal sharing mechanisms involve direct cooperation between police agencies. However, the evidence retrieved through this mechanism may be rendered inadmissible in a court of law because the rules of evidence differ between jurisdictions (even those jurisdictions with similar legal traditions). As such, forensic evidence obtained from another country might not be accepted in another national court.

Branches of Forensic Science

There are several branches of forensic science including (but not limited to): forensic economics, forensic anthropology, forensic odontology, forensic pathology, forensic toxicology, forensic entomology, forensic psychology, forensic accounting, forensic engineering, and computer forensics. The field of forensic economics emerged when courts began allowing expert testimony by specialists in a variety of different fields. Forensic economics is a branch of forensic science that applies economic theories and methods to matters of law. Forensic economists do not investigate illicit activity; instead, they apply economic theories to understand incentives which underlie criminal acts. Originally, forensic economics applies the discipline of economics to the detection and quantification of harm caused by a particular behavior that is the subject of litigation (Zitzewitz 2012). Forensic economics has also been used in the detection of behavior that is essential to the functioning of the economy or that may harm the economy (Zitzewitz 2012).

Forensic anthropology is a branch of science that applies physical or biological anthropology to legal matters. Particularly, it is concerned with the identification of individuals based on skeletal remains. Experts in this field examine human remains in order to determine the cause of death and to ascertain the characteristics of the person's remains they are examining (e.g., gender, age, and height) by evaluating the bones and any antemortem, perimortem, and postmortem bone trauma. Forensic odontology, sometimes referred to as forensic dentistry, is a branch of science that applies dental knowledge to legal matters. It is concerned with the identification of individuals based on dental remains and individual dentition. Forensic odontologists may also evaluate bite mark evidence in the course of their forensic endeavors. Forensic pathology, also referred to as forensic medicine, is concerned with the investigation of sudden, unnatural, unexplained, or violent deaths. Forensic pathologists conduct autopsies to determine the cause, mechanism, and manner of an individual's death. Forensic toxicology is concerned with the recognition, analysis, and

evaluation of poisons and drugs in human tissues, organs, and bodily fluids. Forensic entomology is a branch of science that applies the study of insects to matters of law. Experts in this field are primarily used in death investigations, for example, to shed light on the time and cause of death. Specifically, the life cycle of insects is studied to provide investigatory leads and information about a crime.

Forensic psychology involves the study of law and psychology and the interrelationship between these two disciplines. The American Board of Forensic Psychology defines forensic psychology as "the application of the science and profession of psychology to questions and issues relating to law and the legal system." Bartol and Bartol (1987) "view forensic psychology broadly, as both (1) the research endeavor that examines aspects of human behavior directly related to the legal process; and (2) the professional practice of psychology within, or in consultation with, a legal system that embraces both civil and criminal law" (3). There is considerable disagreement about the nature and extent of activities and roles that fall under the domain of forensic psychology (DeMatteo et al. 2009).

Forensic accounting is a branch of forensic science that applies accounting principles and techniques to the investigation of illicit activities and analysis of financial data in legal proceedings. Forensic engineering is concerned with the investigation of mechanical and structural failures using the science of engineering to evaluate safety and liability. Lastly, computer (or digital) forensics "is a branch of forensic science that focuses on criminal procedure law and evidence as applied to computers and related devices" such as mobile phones, smartphones, portable media players (e.g., iPads, tablets, and iPods), and gaming consoles (Maras 2014, p. 29). Computer forensics involves the acquisition, identification, evaluation, and presentation of electronic evidence (i.e., information extracted from computers or other digital devices that can prove or disprove an illicit act or policy violation) for use in criminal, civil, or administrative proceedings. Electronic evidence is volatile and can easily be lost and manipulated. Maintaining a chain of custody is essential in the preservation and admissibility of electronic evidence.

Strengthening Forensic Science

Increased application of DNA evidence and improved practices and methodologies of crime scene investigations, evidence analysis, and quality assurance measures have resulted in many convictions being reviewed and subsequently overturned. The trend toward wrongful convictions has exposed limitations in forensic science methodologies as well as some forensic analysts. Increased scientific scrutiny, increased quality control within the laboratory, as well as increased professional standards for employment of scientists in the field of forensic science have contributed to improvements in the field, but have not completely ameliorated the lack of oversight apparent in forensic science. The Innocence Project (n.d.) lists unvalidated or improper forensic science (further subdivided into the absence of scientific standards, improper forensic testimony, forensic misconduct) as one of the most common contributing factors to wrongful convictions. According to the Innocence Project (n.d.), other contributing factors to wrongful convictions are eyewitness misidentification, false confessions/admissions, government misconduct (misconduct by law enforcement officials, prosecutorial misconduct), informants, and bad lawyering (inadequate or incompetent counsel).

Driven in part by the status of wrongful conviction in the United States, the National Academy of Sciences Report, *Strengthening Forensic Science in the United States: A Path Forward*, was drafted to address many of the problems plaguing forensic science (Committee on Identifying the Needs of the Forensic Sciences Community, 2009). This document addressed several challenges facing the forensic community, including: disparities in the forensic science community (standard operating procedures, resources, oversight); lack of mandatory standardization, certification, and accreditation; the scope and diversity of forensic science disciplines; problems relating to the interpretation of forensic evidence (such as the degree of scientific research and validity for the various disciplines); the need for research to establish limits; and measures of performance and the admission of forensic science evidence in litigation (concerning the scientific rigor of the

discipline and resultant interpretations as well as the qualification of the expert providing testimony). The report proposed thirteen (13) recommendations ranging from establishing best practices and a scientific foundation within all forensic science disciplines, accreditation of all forensic laboratories, certification of all forensic scientists, increased research to address the reliability and validity of the various forensic science disciplines (e.g., uncertainty measurements, effects of observer bias and human error, development of standardized and scientific techniques, technologies, and procedures) to the development of a code of ethics. Furthermore, organizations, such as the American Academy of Forensic Sciences, European Association of Forensic Sciences, European Network of Forensic Science Institutes, and the International Association of Forensic Sciences, have been created to improve the exchange of forensic science knowledge and best practices between practitioners, researchers, and academicians in the field of forensic science.

References

- Bartol CR, Bartol AM (1987) History of forensic psychology. In: Weiner LB, Hess AK (eds) *Handbook of forensic psychology*. Wiley, New York, pp 3–21
- De Forest PD, Gaensslen RE, Lee HC (1983) *Forensic science. An introduction to criminalistics*. McGraw Hill, New York
- DeMatteo D, Marczyk G, Krauss DA, Burl J (2009) Educational and training models in forensic psychology. *Train Educ Prof Psychol* 3(3):184–191
- Innocence Project (n.d.) The causes of wrongful. Retrieved from <http://www.innocenceproject.org/understand/>
- Maras M-H (2014) *Computer forensics: cybercriminals, laws and evidence*, 2nd edn. Jones and Bartlett, Sudbury
- Moenssens A (1993) Novel scientific evidence in criminal cases: some words of caution. *J Crim Law Criminol* 84(1):1–21
- Schaler RC (2012) *Crime scene forensics: a scientific method approach*. CRC Press, Boca Raton
- Zitewitz E (2012) Forensic economics. *J Econ Lit* 50(3):731–769

Forgery

► Counterfeiting Models: Mathematical/
Economic

Franchise

Alicia García-Herrera
Ilustre Colegio de Abogados de Valencia,
Valencia, Spain

Abstract

Since the second half of the twentieth century, the variations in the exchange system of goods and services have allowed the expansion of integrated distribution networks, based on franchise or other distribution agreements. Franchise offers specific benefits to the firms, reducing transaction costs. Manufacturers and suppliers may have access to new markets, raising capital, sharing risks and saving costs, but maintaining the control of the franchisees' behaviour through the terms of the contract. Franchising is also attractive to franchisees, because of the backing of a successful and recognized system and the ongoing support of the brand to the entrepreneur. However, certain practices, like resale maintenance or price discrimination, territorial restrictions or refusal to supply (among others) may restrict competition among firms by establishing barriers to entry, and consequently they have been controlled by antitrust law. In the internal relationship, the unequal allocation of rights, with the attribution of broad powers to franchisors, including termination at *will*, favours the opportunistic behaviour of both parties. The legal approach has considered the vulnerability of the franchisee and the unequal bargaining power between the parties by imposing pre-contractual disclosures rules and regulating the termination. In front of these views, economic analysis has criticised the state interventionism in franchising. Both perspectives, opposite, should be considered in order to have an adequate comprehension of the reality inherent to franchise contracts.

Synonyms

Franchising

Definition

Franchise is a successful system of business organization to market goods and services based in a contractual relationship between two legally independent parties, according to which one well-established business, the franchisor, in exchange for the pay of an initial sum, fees or periodic royalties, transfer or license to the other part, the franchisee, the right to use in a given area and during a period of time, their trademark, trade name, or another industrial and intellectual property rights, with the provision of commercial support and technical assistance.

Franchise as a Business Model

From a broad perspective, franchise can be defined as a contractual relationship involving two legally independent entrepreneurs with the goal of market goods or services in a given location during a specified or indefinite period of time. Using a narrow approach, franchise commonly refers to those agreements known as *business format franchise*. In this sense, the purpose of contract is the transmission by the franchisor of a successful and recognized business system, giving continuous provision of commercial and technical assistance to the franchisee. Subsequently, the franchisor must make available to the franchisee – through appropriate license agreements – all *intangible assets* that had led the company success (*good will*), like the brand name or another industrial and intellectual property rights as well as the *know-how*. The franchisee has obligations of confidentiality and no competition about franchisor's business methods, even after termination of the relationship. In exchange to join the network, the franchisee assumes the payment of an *up-front lump sum* or fixed initial *fee* and ulterior royalties, commissions, or percentages of retail sales. Given that his business opportunity is the distribution of the product or service under franchisor's techniques and knowledge, the franchisee may make highly specific investments to maintain the value and quality of the brand – thus, franchisors must guarantee a reasonable length of the contract. The contract clauses, standardized, attribute to the franchisor

wide powers of monitoring. The franchisor controls the supply, distribution, and the resale of goods or services, although he is obligated to provide equal treatment to all franchisees. He imposes conditions about characteristics of the franchised outlet and staff training because the network has to be homogeneous and uniform. By recommending resale prices and coordinating franchise advertising and promotional activities, the franchisor is involved in the distribution process.

Business format franchise must be differentiated from *product distribution franchise*, which is basically focused in the reselling of franchisor's brand products (often soft drinks, automobiles, and gasoline), frequently with the allocation to the franchisee of exclusive agreements about supplies and area (not necessarily present in business format franchise). In this case, the franchisor habitually does not provide to franchisees an entire system of business. The degree of integration is apparently minor than in the business format franchise. However, both types of distribution agreements can be combined in practice. The *business format franchise* must also be differentiated from *selective distribution*, which is characterized by the selection of resellers considering criteria directly related with the specialties of luxury markets or complex technology products.

The existing literature refers the origin of modern franchise to the USA, even though etymologically the term comes from the Old French word *Franc* (*free*) and its derivation *francher* or *affranchir*. The medieval franchise consisted on privileges granted by kings and lords to their vassals to obtain licenses about trade, fishing or forestry rights, exemptions about taxes or customs, although it was also associated to the statutes of the *villas francas*, released of manor or vassalage. In the sense of concession or privilege, the term franchise still remains in government grants and in the field of sports. Despite this, the franchise, as we know it today, arises in the late nineteenth century as a distribution method used by manufactures to avoid the application of the *Sherman Act*, which impeded to manufactures the resale to final costumers. The great boom of the franchise occurs after Second World War, with the expansion of the *business format franchise*

(s. Martinek 1992, Martinek, Semler & Flohr 2015).

Nowadays, franchise networks represent almost a third of retailing in the USA. According to the *Franchise Business Economic Outlook for 2014*, a report prepared by the *International Franchise Association (IFA) Educational Foundation*, around 3.5 % of the US GDP corresponds to franchise business (a total of \$ 472 billions). This method of distribution is especially relevant to small and medium sized firms. In Europe, the statistics of the *European Franchise Federation* show that franchise is also growing, although the impact of franchise is still minimal comparing to USA or other countries, as Japan, Canada, and Australia. New technologies as the *Internet* – via e-commerce – may be a likely way of development of franchise.

Franchise contracts have interested both lawyers and economists. Initial economic works showed that franchise was an efficient method for reducing monitoring and agency costs (Caves and Murphy 1976; s. also model of Rubin, 1978). The agency theory explains that franchising is an optimal business decision when the cost of monitoring is high, as in the case of dispersed units located far from the franchisor (Mathewson and Winter 1985; Brickley and Dark 1987; Shane 1996, 1998, among others).

But, in the business relationship incentive and interest conflicts may arise, reducing the efficiency of the franchise contract. A franchise is a long term and *relational contract*, in which certain conditions are implicit, due to the inability to predict at the outset all contingencies. This fact implies that both franchisors and franchisees must consider common interest, being reciprocally obligated to act in *good faith* (this being understood as a fair behavior and the prohibition of reciprocal opportunism). But, both franchisors and franchisees have also incentives to take advantage of the loopholes and uncertainties of the written contract in their own benefit. From the agency theory perspective, franchisors must protect the value of their trademarks from the problems of *adverse selection* (the franchisor cannot ensure that the franchisee is able to reach the purpose of the contract) and *moral hazard*. Due to vertical and horizontal externalities, the franchisee

has a tendency to practices like double marginalization, underinvestment, and internal *free riding*. The franchisee may also be worried about the franchisor *holdup*. In the relationship, he acts as a passive investor. Through *encroachment* (invasion of the area of protection or the franchisee's territory) or simply by exercising their right to cancel or not renovate *at will*, franchisors may appropriate profits of efficient franchisees, such as *sunk cost* and the local goodwill built by their efforts.

The legal approach has considered the vulnerability of franchisees, in general less sophisticated and with a lower bargaining power than franchisors – thus, the franchise contract may be subject both to laws regarding unfair terms and to those prohibiting discrimination (antitrust law and contract law). The imbalance between the parties has justified the global tendency to the regulation of franchise agreements or, in the absence of law, designing mechanisms of court enforcement (based on *Common Law*). Complementary, different associations have developed *Codes of Ethics* for franchising.

The first laws about franchise arise in the USA. At the federal level, the Federal Trade Commission (FTC) Rule 1979, amended in 2007, governs presale disclosure obligation and registration. The government also regulates the termination of the relationship in specific areas, as automobile (Federal Automotive Dealer Franchise Act, FADFA 1956) and petroleum sectors (Petroleum Marketing Practices Act, PMPA 1978). From 1971 to 1992, a third part of states enacted relationship laws about franchise and automobile dealers. Franchise relationship laws impose the performance of the contract in *good faith*, encouraging the stability of the relationship, prohibiting *encroachment*, and establishing restrictions about termination. The termination or not renovation of the contract is based on various formal requirements (notification, notice, cure period) and on allegation of *good cause*. This legislation has influenced other legal systems and international law (s. the draft UNIDROIT *Model Franchise Disclosure Law*). Many countries in Asia, Latin America, and Canada regulate franchise to a greater or lesser extent. In general terms, franchise

has been not “*encoded*” in Europe, but in the last decade the tendency is changing. The Italian law is a good example. France, Belgium and Spain have enacted laws about disclosure obligations of franchisor and registration. The *Spanish Draft Commercial Code* establishes also parameters about the duration of franchise and distribution agreements (*reasonability*), the termination of the relationship, and the compensation of the franchisee (articles 543–18 to 543–249).

Over the last decades, starting from agency theory, the economic literature has tried to explain the economics of franchise contracts. A great number of works focus especially in the study of monetary clauses (*royalty rate* and *initial franchise fee*). The monetary clauses ensure a continuous flow of rents that benefits both franchisors and franchisees and offer incentives in order to control *double-sided moral hazard* (starting by Rubin, 1978, s. for a model Bhattacharya and Lafontaine 1995 and later empirical works that support this model, Lafontaine and Shaw 1999). A complementary approach has showed that franchise contracts themselves are designed to solve the problem of postcontractual opportunistic behavior as well as incentive conflicts through mechanisms of *self-enforcement* (Brickley et al. 1991; Mathewson and Winter 1994; Klein 1995; Dnes 1993, 1996). The disciplinary powers of the franchisor and, if required, the termination of the relationship can constitute a very efficient *hostage* to assure the optimal performance. The end of the relationship supposes to the franchisee the loss of the ongoing rents and future profits and additionally, the recovery of specific investments can be difficult. The economic literature considers that the risk of opportunistic behavior by the franchisee is higher than the risk of the franchisor *holdup*. Theory suggests that franchisors would tend to throw out of the network less efficient franchisees (Blair and Lafontaine 2010) and have no incentives to the appropriation of resources – they are interested in maintaining the brand prestige, which could be damaged by litigations derived of termination, thus generating the image of “*hard network*.” This approach explains the structure of franchise contracts and justifies the asymmetrical allocation of rights in favor of

franchisors, with the power of termination *at will* (s. empirical work of Arruñada et al. 2001, 2005, 2009, about automobile dealing agreements).

Most recent economic works have considered that, in long-term contracts, optimal contract duration can be a mechanism to reduce incentive conflicts and to avoid the problem of underinvestment (Guriev and Kvasov 2005). Given that the literature traditionally has assumed that initial investments are a key factor for the expected length of the franchise agreements (Joskow 1987; Brickley et al. 2006), a theoretical model to determine optimal duration of franchise is developed in García-Herrera and Llorca Vivero 2010 and empirically tested.

The economic analysis of the franchise contract, supported by empirical works, suggests that in general both *Franchise termination laws* – and even, those that protect franchisees from unfair treatment and discrimination – have no justification, inducing to a reduction in the use of franchise (s. Smith 1982; Beales and Muris 1995; Blass and Carlton 2001, about gasoline retailing; most recently, about automobile industry Arruñada et al. 2009; Lafontaine and Scott Morton 2010; Zanarone 2009; Zanerone 2012, about rigidity to adapt the contract). But the impact of franchise relationship laws is not homogenous because of its differences (Klick et al. 2008). Moreover, given that the opportunistic behavior of franchisors is not entirely compensated by the contractual engineering of the franchise, complementary legal, court, and extralegal enforcement mechanisms can be justified under certain conditions (s. Dnes 2009).

Vertical restraints associated to franchise contracts have been also analyzed from the *antitrust law* perspective. Some practices like the resale price maintenance or price discrimination, territorial restriction, tied-in sales, or the refusal to supply may breach antitrust laws. The economic analysis has justified vertical restraints associated to franchise contracts because of their efficiencies and benefits (Klein 1995; Rey and Stiglitz 1995; Mathewson and Winter 1994; Lafontaine and Slade 2007 among others). Since the eighties, the European Commission has enacted block exemption regulations (BER) about distribution and dealership agreements,

complemented by guidelines, and after the case *Pronuptia*, also about franchise. These rules, unified nowadays – excluding automobile sector – have been criticized and successively amended (the last modification was in 2010 (Commission Regulation (EU) No. 330/2010 of 20 April 2010 on the Application of Article 101(3) of the Treaty on the Functioning of the European Union to Categories of Vertical Agreements and Concerted Practices [2010] OJ L102/1-7; Commission Regulation (EU) No. 461/2010 of 27 May 2010 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to categories of vertical agreements and concerted practices in the motor vehicle sector [2010] OJ L129/52-5; Commission Notice, Supplementary Guidelines on Vertical Restraints in Agreements for the Sale and Repair of Motor Vehicles and for the Distribution of Spare Parts for Motor Vehicles [2010] OJ C138/16-27.7.), reflecting the changes in distribution, as *e-commerce* and relaxing policies about passive sales and resale price maintenance).

Summary and Future Directions

Economic analysis has explained the business decision about franchise and the economics of franchise contracts, justifying the asymmetrical allocation of rights and vertical restraints associated. The empirical works have also demonstrated in general the negative effect of franchise relationship laws over the business decision about market goods or services through franchise. However, these generic results require qualifications and more empirical work should be necessary. The current tendency in Europe, as before in USA, is to ensure the stability of the contracts and their equilibrium by mechanism of legal and court enforcement, providing certainty and incentives to entrepreneurs to invest in franchise. In the reduction of litigations derived from under-performance, renegotiation or termination of the contract, it should be studied the complementary role of mediation, negotiation or arbitration clauses. EU competition rules for franchising agreements should probably be amended in the future to reflect the changes in distribution caused by the Internet phenomenon.

Cross-References

- [Economic Analysis of Law](#)
- [Fixed Investment](#)
- [Information Disclosure](#)
- [Transaction Costs](#)
- [Vertical Integration](#)

References

- Arruñada B, Garicano L, Vázquez L (2001) Contractual allocation of decision rights and incentives: the case of automobile distribution. *J Law Econ Org* 17:257–284
- Arruñada B, Garicano L, Vázquez L (2005) Completing contracts ex-post: how car manufacturers manage car dealers. *Rev Law Econ* 1:149–173
- Arruñada B, Vázquez L, Zananone G (2009) Institutional constraints on organizations: the case of Spanish car dealerships. *Manag Decis Econ* 30:15–26
- Beales H, Muris TJ (1995) The foundations of franchise regulation: issues and evidence. *J Corp Financ Contract Organ Gov* 2:157–197
- Bhattacharya S, Lafontaine F (1995) Double-sided moral hazard and the nature of the share contracts. *Rand J Econ* 26:761–781
- Blass A, Carlton D (2001) The choice of organizational form in gasoline retailing and the lost of laws that limit that choice. *J Law Econ* 44:511–524
- Brickley JS, Dark FH (1987) The choice of organizational form: the case of franchising. *J Financ Econ* 18(2), 401–420.
- Brickley JS, Dark FH, Weisbach M (1991) The economic effects on franchise termination laws. *J Law Econ* 34:101–130
- Brickley JA, Misra S, VAN Horn L (2006) Contract duration: evidence from franchising. *J Law Econ* 49:173–196
- Caves RE, Murphy WF II (1976) Franchising: firms, markets and intangible assets. *South Econ J* 42:572–586
- Dnes AW (1993) A case-study analysis of franchise contracts. *J leg Stud* 22:367–393
- Dnes AW (1996) The economic analysis of franchise contracts. *J Inst Theor Econ* 152:297–324
- Dnes AW (2009) Franchise contracts, opportunism and the quality of law. *Entrep Bus Law J* 3:258–274
- García-Herrera A, Llorca Vivero R (2010) How time influences franchise contracts: the Spanish case. *Eur J Law Econ* 30:1–16
- Guriev S, Kvassov D (2005) Contracting on time. *Am Econ Rev* 95:1369–1385
- Joskow P (1987) Contract duration and relationship-specific investments: empirical evidence from coal markets. *Am Econ Rev* 77:168–185
- Klein B (1995) The economics of franchise contracts. *J Corp Fin* 2:9–37
- Klick J, Kobayashi B, Ribstein L (2008–2009) Federalism, variation and state regulation on franchise termination. *Entrep Bus Law J* 3:355–380
- Lafontaine F, Scott Morton F (2010) State franchise laws, dealer terminations, and the auto crisis. *J Econ Perspect* 24:233–250
- Lafontaine F, Shaw KL (1999) The dynamics of franchise contracts. *J Polit Econ* 107:1041–1080
- Lafontaine F, Slade ME (2007) Vertical integration and firm boundaries: the evidence. *J Econ Lit* 45:631–687
- Mathewson GF, Winter RA (1985) The economics of franchise contracts. *J Law Econ* 28:503–526
- Mathewson GF, Winter RA (1994) Territorial rights in franchise contracts. *Econ Inq* 32:181–192
- Rey P, Stiglitz J (1995) The role of exclusive territories in producer's competition. *Rand J Econ* 26:431–451
- Shane SC (1996) Hybrid organizational arrangements and their implications for firm growth and survival: a study of new franchisors. *Acad Manage J* 39:216–234
- Shane SC (1998) Making new franchise systems work. *Strateg Manag J* 19:697–707
- Smith RL (1982) Franchise regulation: an economic analysis of state restrictions on automobile distribution. *J Law Econ* 26:431–451
- Zananone G (2009) Vertical restraints and the law: evidence from automobile franchising. *J Law Econ* 52:691–700
- Zanerone (2012) Contract adaptation under legal constraints. *J Law Econ Org* 29:799–834

Further Reading

- Blair RG, Lafontaine F (2010) *The economics of franchising*. Cambridge: Cambridge University Press
- Board S (2011) Relational contracts and the value of loyalty. *Am Econ Rev* 101:3349–3367
- Killion W (2008) The modern myth of the vulnerable franchisee: the case for a more balanced view of the franchisor-franchisee relationship. *Fr Law J* 28:23–33
- Lafontaine F, Blair R (2008–2009) The evolution of franchising and franchise contracts: evidence from the United States. *Entrep Bus Law J* 3:381–434
- Martinek M (1992) *Moderne Vertragstypen, vol II*. Verlag CH Beck, München
- Martinek M (1987) *Franchising*. v. Decker, Heidelberg
- Martinek M, Flohr E (2011) *European distribution law: A commentary*, Hart Publishing, Oxford (Hart, R & Parker, J, publishers)
- Martinek M, Semler FJ, Flohr E (2015) *Handbuch des Vertriebsrechts*, 4th edn. CH Beck, München
- Peters L (2000) The draft Unidroit Model Franchise Disclosure Law and the move towards national legislation. *Uniform Law Rev* 5(4):717–735
- Sertsios G (2013) Bonding through investments: evidence from franchising. *J Law Econ Org* 31(1):187–212
- Vertical Restraints in the Internet Economy (2013) Meeting of working group of competition law. Available at www.bundeskartellamt.de

Franchising

- [Franchise](#)

Freedom

► Liberty

Frisking

Matt E. Ryan
Department of Economics, Duquesne University,
Pittsburgh, PA, USA

Definition

Established by the Supreme Court in *Terry v. Ohio* (1968), police officers may perform a frisk – a limited search of a civilian’s outer clothing in pursuit of weapons or nonthreatening contraband – when they determine suspicious activity is afoot or otherwise feel threatened.

Legal Framework

The legal framework surrounding frisking stems from *Terry v. Ohio* (1968):

We merely hold today that where a police officer observes unusual conduct which leads him reasonably to conclude in light of his experience that criminal activity may be afoot and that the persons with whom he is dealing may be armed and presently dangerous, where in the course of investigating this behavior he identifies himself as a policeman and makes reasonable inquiries, and where nothing in the initial stages of the encounter serves to dispel his reasonable fear for his own or others’ safety, he is entitled for the protection of himself and others in the area to conduct a carefully limited search of the outer clothing of such persons in an attempt to discover weapons which might be used to assault him.

Terry concerns only weapons; procedures for stops involving contraband come from *Minnesota v. Dickerson* (1993):

The question presented today is whether police officers may seize nonthreatening contraband detected during a protective patdown search of the sort permitted by *Terry*. We think the answer is

clearly that they may, so long as the officer’s search stays within the bounds marked by *Terry*.

Arizona v. Johnson (2009) extends *Terry* to traffic stops by citing *Berkemer v. McCarty* (1984), which notes that “[m]ost traffic stops resemble, in duration and atmosphere, the kind of brief detention authorized in *Terry*.” As such, these United States Supreme Court cases form the foundation of the role of the police officers’ permissible behavior concerning frisks in the two arenas generally examined empirically.

Stop-and-Frisk

“Stop-and-frisk” programs constitute stopping, questioning, and frisking pedestrians. In theory, stop-and-frisk programs are discriminatory if implemented unevenly across groups – be it race, ethnicity, gender, age, or any other margin. However, finding the appropriate null hypothesis, or benchmark, to statistically test these claims can be difficult. For instance, consider a hypothetical police department that implements a stop-and-frisk program; in doing so, they perform an unequal number of frisks across races. This outcome alone is not evidence of discrimination as the proper comparison must be made in order to determine any deviation from a no-bias level. City population figures could provide a benchmark but requires the assumption that criminal activity within the city is in proportion to population across races. Proportion of crimes committed across races could be a benchmark yet requires an assumption of criminal activity across races being proportional to crime committed. Moreover, discrepancies in frisking across races must control for discrepancies in exposure to policing across races as well. For a healthy discussion of benchmarking issues, see Ridgeway (2007).

Nevertheless, many studies investigate stop-and-frisk programs. Whether race-based discrimination exists in the implementation of New York City’s stop-and-frisk program is a particularly popular subject, with most studies finding some sort of inequality along race lines. Ridgeway (2007) finds that nonwhites received slightly more frequent

frisks, though differences in raw statistics overstate racial disparities. Gelman et al. (2007) find that black and Hispanics are stopped twice as frequently as whites, though arrested less frequently. Significant racial disparities exist in New York City pertaining to implementation of marijuana enforcement across both stops and arrests (Geller and Fagan 2010). Friedman (2015) finds African-American suspects less likely to be in possession of contraband as compared to white suspects. Goel et al. (2016) show that blacks and Hispanics are disproportionately stopped in scenarios with a low probability of a successful frisk – i.e., discovering weapons or contraband. Fryer (2016) notes that black and Hispanics are more likely to have an interaction with police involving force. Conversely, Coviello and Persico (2015) generate a unique measure of racial bias and find that race disparities in police pressure across precincts are not correlated with this measure. New York's stop-and-frisk program likely received considerable scholarly attention – at least in part – due to the sheer volume of stops performed; the database utilized by Fryer (2016) contains nearly five million stops, with annual stops between 2003 and 2013 numbering well into the hundreds of thousands. A US district court ruling in 2013, however, determined the implementation of the New York stop-and-frisk program to be unconstitutional (the program itself was not deemed to be unconstitutional), and consequently the number of stops performed annually by the New York City Police Department has dropped by over 90%.

Pertaining to the broad practice of frisking, Persico and Todd (2006) show that officers administer searches so as to maximize the number of successful searches. Antonovics and Knight (2009) note that searches are more likely to occur when the races of the officer and the searched differ; several additional studies consider officer race in relation to suspect race; see Skogan and Frydl (2004), Brown and Frank (2006), Sklansky (2006), and Gilliard-Matthews et al. (2008). Further, Durlauf (2005) notes that there exists an equity/efficiency trade-off in racial profiling – namely, the inequity in racial profiling must be weighed against the reduction in crime rates should underlying cross-race

differences in illicit activity dictate such a relationship to exist.

Frisking and Traffic Stops

While stop-and-frisk programs concern officers' interaction with citizens "on the street," a number of studies have investigated frisking in the context of a traffic stop. Most find a similar racial component to frisking during traffic stops as with the above-discussed "on the street" stops. In an early study, however, Knowles et al. (2001) find no evidence of racially prejudiced search behavior against black Maryland motorists, instead finding a modicum of bias against white and Hispanic motorists. Examining a wide swath of Missouri municipalities, Rojek et al. (2004) show that black and Hispanic drivers are approximately twice as likely as white drivers to be searched once stopped. Schafer et al. (2006) find that black and Hispanic drivers are more likely to be searched in an anonymized Midwestern city. Rosenfeld et al. (2012) find that young black males are searched more frequently than young white males in St. Louis and that this gap disappears for drivers over the age of 30. Novak and Chamlin (2012) show that the search rate for white motorists in Kansas City increased as the percentage of blacks in a particular neighborhood increased; black motorists were not searched more frequently. Ritter (2013) finds evidence of implicit race discrimination in traffic stops performed in Minneapolis. In Rhode Island, Carroll and Gonzalez (2014) show that black drivers are more likely to be frisked than white drivers, conditional on the racial composition of the community in which the traffic stop takes place. In Pittsburgh, Ryan (2016) finds a black male to be up to 8% more likely to receive a frisk when compared to an equivalent white driver.

Cross-References

- [Criminal Sanctions and Deterrence](#)
- [Economic Analysis of Law](#)
- [Government Failure](#)

- [Public Enforcement](#)
- [Traffic Lights Violations](#)

References

- Antonovics K, Knight B (2009) A new look at racial profiling: evidence from the Boston police department. *Rev Econ Stat* 91(1):163–177
- Arizona v. Johnson (2009) 129 S. Ct. 781, 555 U.S. 323, 172 L. Ed. 2d 694
- Berkemer v. McCarty (1984) 468 U.S. 420, 104 S. Ct. 3138, 82 L. Ed. 2d 317
- Brown R, Frank J (2006) Race and officer decision making: examining differences in arrest outcomes between black and white officers. *Justice Q* 23(1):96–126
- Carroll L, Gonzalez M (2014) Out of place: racial stereotypes and the ecology of frisks and searches following traffic stops. *J Res Crime Delinq* 51(5):559–584
- Coviello D, Persico N (2015) An Economic analysis of black-white disparities in the New York police department's stop-and-frisk program. *J Leg Stud* 44(2):315–360
- Durlauf SN (2005) Racial profiling as a public policy question: efficiency, equity, and ambiguity. *Am Econ Rev* 95(2):132–136
- Friedman M (2015) The role of race in police interdictions: evidence from the New York police department's use of stop, question and frisk policing. Available at: <https://ssrn.com/abstract=2689766>
- Fryer R (2016) An empirical analysis of racial differences in police use of force. No. w22399, National Bureau of Economic Research
- Geller A, Fagan J (2010) Pot as pretext: marijuana, race, and the new disorder in New York City street policing. *J Empir Leg Stud* 7(4):591–633
- Gelman A, Fagan J, Kiss A (2007) An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *J Am Stat Assoc* 102(479):813–823
- Gilliard-Matthews S, Kowalski B, Lundman R (2008) Officer race and citizen-reported traffic ticket decisions by police in 1999 and 2002. *Police Q* 11(2):202–219
- Goel S, Rao JM, Shroff R (2016) Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Ann Appl Stat* 10(1):365–394
- Knowles J, Persico N, Todd P (2001) Racial bias in motor vehicle searches: theory and evidence. *J Polit Econ* 109(1):203–229
- Minnesota v. Dickerson (1993) 508 U.S. 366, 113 S. Ct. 2130, 124 L. Ed. 2d 334
- Novak K, Chamlin M (2012) Racial threat, suspicion, and police behavior: the impact of race and place in traffic enforcement. *Crime Delinq* 58(2):275–300
- Persico N, Todd P (2006) Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in Wichita. *Econ J* 116(515):F351–F367
- Ridgeway G (2007) Analysis of racial disparities in the New York police department's stop, question and frisk practices. RAND Corporation, Santa Monica
- Ritter J (2013) Racial bias in traffic stops: tests of a unified model of stops and searches. No. 152496, University of Minnesota, Department of Applied Economics
- Rojek J, Rosenfeld R, Decker S (2004) The influence of driver's race on traffic stops in Missouri. *Police Q* 7(1):126–147
- Rosenfeld R, Rojek J, Decker S (2012) Age matters: race differences in police searches of young and older male drivers. *J Res Crime Delinq* 49(1):31–55
- Ryan M (2016) Frisky business: race, gender and police activity during traffic stops. *Eur J Law Econ* 41(1):65–83
- Schafer J, Carter D, Katz-Bannister A, Wells W (2006) Decision making in traffic stop encounters: a multivariate analysis of police behavior. *Police Q* 9(2):184–209
- Sklansky D (2006) Not your father's police department: making sense of the new demographics of law enforcement. *J Crim Law Criminol* 96(3):1209–1243
- Skogan W, Frydl K (2004) Fairness and effectiveness in policing: the evidence. National Academies Press, Washington, DC
- Terry v. Ohio (1968) 392 U.S. 1, 88 S. Ct. 1868, 20 L. Ed. 2d 889

Frivolous Suits

- Yannick Gabuthy^{1,2} and Eve-Angéline Lambert^{1,3}
- ¹BETA, CNRS, University of Strasbourg, Strasbourg, France
- ²University of Lorraine, Nancy, France
- ³BETA UMR 7522, Université de Lorraine, Nancy, France

Synonyms

[Nuisance Lawsuits](#)

Definition

Frivolous lawsuits refer to cases that are brought by plaintiffs with the only objective to extract settlement offers from defendants. Whereas a wide literature questions the credibility of frivolous litigation, this phenomenon had significant policy implications by inspiring several legal reform acts designed to deter meritless claims. Indeed, the issue of frivolous suits may be important from a welfare perspective since such claims may consume substantial resources (due to litigation costs, judicial

congestion, etc.) and have negative distributive consequences (since a payment is made to a party who has no legal entitlement to recovery).

Introduction

Many scholars and policy makers have expressed concerns about frivolous suits. However, despite this broad concern, there is no consensus about how a frivolous case should be defined (Bone 1997; Spier 2007). A first approach defines frivolous suits as negative expected value suits, i.e., claims in which the expected award at trial (probability of winning times award in case of victory) is lower than the plaintiff's litigation costs. But as Bone notes, the problem with such definition is that cases with very high merits and high probability of winning would be considered as frivolous suits if the costs are very high too. According to a second definition, a frivolous suit is defined by its very small probability of success. But this definition implies that if a jury is biased toward a plaintiff, then even though the probability that the defendant be liable is very low, that suit would not be frivolous according to that definition, although most people would consider it is. A third definition is not based on the probability of success (which can be subjective) but on the plaintiff's belief that there is little or no chance that the defendant be liable.

Departing from these definitions, it appears that a frivolous plaintiff ("she") would rather drop her case rather than going to trial. However, the defendant ("he") may fear the prospect of incurring high litigation costs in case of trial or face uncertainty by ignoring whether the claim is frivolous or not (Katz 1990). In such situations, he could be prone to make a settlement offer to the plaintiff.

Empirical Evidence

Given the negative welfare implications of frivolous litigation (increase in the overall number of cases in courts, judicial congestion, litigation costs, unjustified transfers of wealth), frivolous

suits are frequently cited as a major cause of the civil judicial system's most serious ills, despite the very little reliable empirical data confirming the importance of this phenomenon. This is notably the case in the USA where the number of nuisance suits has been an often-voiced concern for decades. For example, in a reported survey of American jurors in cases in which firms and corporations were defendants, more than 80% of the jurors indicated that they agree/strongly agree with the statement according to which "there are far too many frivolous lawsuits today" (Polinsky and Rubinfeld 1993). In the same way, the society's perceptions of the tort system are highly influenced by anecdotes of specious claims, one of the most emblematic being the Mc Donald's coffee case (see *Liebeck v. Mc Donald's Restaurants*, Docket No D-202 CV-93-02419, 1995 WL360309, Bernalillo County, N.M. Dist. Ct. August 18, 1994). The concerns about the existence of frivolous litigation gave rise in the 1980s and the 1990s in the USA to several litigation reform acts designed to deter meritless suits. For instance, Rule 11 of the Federal Rules of Civil Procedure provides for the imposition of sanctions on individuals who present claims which are deemed to be frivolous.

Threat to Litigate and Credibility

P'ng (1983) highlights that even if a defendant is aware that the claim is frivolous, the fear to see the case going to trial anyway might lead him to make a settlement offer that would be lower than his trial costs. However, following the argument by Bebchuk (1988), the plaintiff's threat to litigate is not credible in such a situation and a settlement should not occur, lessening the plaintiff's incentives to file a frivolous case: Knowing that the plaintiff will be not incited to go to trial *ex post*, the defendant is not prone to settle *ex ante*. In this paper, Bebchuk highlights that a negative expected value suit can be filed only in the presence of uncertainty. Indeed, if the plaintiff has private information about the level of damage she suffered or about her expected litigation costs, then the defendant might not know that

the expected value of litigation to the plaintiff is negative, making the plaintiff's threat to litigate the claim credible (see also Katz 1990, who considers an asymmetric information framework).

Bebchuk (1996) considers litigation with two stages and litigation costs at each stage. A party can propose a settlement at each stage, which can be accepted or refused by the other one. In this model, even if the case has a global negative expected value, the claim may, depending on the last stage's litigation costs, become positive expected valued. The plaintiff would thus have a credible threat to go to trial at the stage of litigation and bring the defendant to make an offer. On the opposite side, Schwartz and Wickelgren (2009) challenge this result and argue that nuisance suits are not credible. Their result relies on the assumption that during litigation, both parties can make offers at any time and at no cost. This implies that the plaintiff would not be able to extract a settlement that is large enough to make the initial threat of filing credible. Other arguments against the credibility of frivolous suits rely on a possible strategy for defendants that would be to develop a reputation of refusing to settle (Bone 1997). Nevertheless, as noted by Hubbard (2014), the same argument can apply to the opposite reasoning since plaintiffs might also commit to a policy of refusing to be deterred (Farmer and Pecorino 1998; Chen 2006).

Public Regulation

Some policy instruments are considered in literature to deter frivolous suits by undermining the ability of plaintiffs to extract settlements. Polinsky and Rubinfeld (1993) analyze the sanctions that can be enforced to this end and highlight the conditions that these sanctions should fulfill. In particular, the sanctions would be more (respectively less) desirable when litigation costs due to their use are low (respectively high). Moreover, when sanctions are enforced, they should be set so as to deter frivolous plaintiffs and not so as to compensate non-frivolous defendants. Finally, they underline the fact that in order to avoid discouraging legitimate plaintiffs to sue because of mistakenly

imposed sanctions, the judgment awarded to a legitimate plaintiff should be raised, all by keeping the expected costs borne by the defendant constant.

The English fee-shifting rule, which requires the losing litigant to pay the litigation costs of the winning party, might be a second instrument to deter frivolous suits by discouraging low-probability-of-prevailing cases. Nevertheless, the effect of the English rule on the incentives to file frivolous claims is ambiguous. Bebchuk and Chang (1996) show that if the plaintiff could predict the trial outcome with certainty, then a frivolous suit should not occur under the English rule. However, such a claim may occur if the plaintiff cannot predict the trial outcome without error. Indeed, with small litigation costs, a frivolous plaintiff could file anyway because of a small but positive chance of victory. In the same way, consider a negative expected value case where the probability of winning is quite high but the damages are very small. As argued by Spier (2007), the English rule will encourage such a case by enhancing its value given that the plaintiff's litigation costs will be shifted to the defendant at trial. Farmer and Pecorino (1998) analyze frivolous suits in a repeated play setting in which a lawyer may develop a reputation for proceeding to trial with a frivolous suit when facing a refusal by the defendant of the plaintiff's pretrial offer. According to them, such a reputation is necessary to maintain a credible threat of trial in future periods, and the value of this future reputation generates the credibility to pursue a case to trial in the current period. In this context, the English rule would discourage frivolous suits since the defendant's willingness to pay to settle decreases in the percentage of litigation costs that can be shifted to the losing party in case of trial.

Finally, some scholars and commentators have argued that the American system of contingent fees, under which the attorney gets a share of the judgment if his client wins and nothing if he loses, may encourage frivolous claims and should therefore be prohibited. These arguments are based on the idea that a plaintiff's threat to litigate is higher with contingent fees because the lawyer is not paid in case of losing. Dana and Spier (1993) challenge this argument: When the plaintiff's

attorney has better information about the merits of the case than his client, then this latter is constrained to rely upon the lawyer's recommendation, given that the contingent fee arrangement will encourage this lawyer to pursue only cases with sufficiently high expected returns. In this context, the contingent fee regime should decrease the extent of frivolous litigation. Instead, under a fixed or an hourly fee, the lawyer would be incited to lead the plaintiff blindly into litigation regardless of the claim's merit.

Cross-References

- [Legal Disputes](#)
- [Litigation Decision](#)

References

- Bebchuk LA (1988) Suing solely to extract a settlement offer. *J Leg Stud* 17(2):437–450
- Bebchuk LA (1996) A new theory concerning the credibility and success of threats to sue. *J Leg Stud* 25(1):1–25
- Bebchuk LA, Chang HF (1996) An analysis of fee shifting based on the margin of victory: on frivolous suits, meritorious suits, and the role of rule 11. *J Leg Stud* 25(2):371–403
- Bone RG (1997) Modeling Frivolous Suits. *Univ Pa Law Rev* 145:519–605
- Chen Z (2006) Nuisance suits and contingent attorney fees. *Rev Law Econ* 2(3):363–371
- Dana JD Jr, Spier KE (1993) Expertise and contingent fees: the role of asymmetric information in attorney compensation. *J Law Econ Org* 9(2):349–367
- Farmer A, Pecorino P (1998) A reputation for being a nuisance: frivolous lawsuits and fee shifting in a repeated play game. *Int Rev Law Econ* 18(2):147–157
- Katz A (1990) The effect of frivolous lawsuits on the settlement of litigation. *Int Rev Law Econ* 10(1):3–27
- P'ng IPL (1983) Strategic behavior in suit, settlement, and trial. *Bell J Econ* 14(2):539–550
- Polinsky AM, Rubinfeld DL (1993) Sanctioning frivolous suits: an economic analysis. *Georgetown Law J* 82: 397–436
- Schwartz WF, Wickelgren AL (2009) Advantage defendant: why sinking litigation costs makes negative-expected-value defenses but not negative-expected-value suits credible. *J Leg Stud* 38(1):235–253
- Spier KE (2007) Litigation. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*, vol 1, pp 259–342
- William H. J. Hubbard (2014) “Nuisance Suits”, University of Chicago Public Law & Legal Theory, Working Paper No. 479

Galbraith, John Kenneth

Alexandre Chirat
Triangle, Université Lumière Lyon II, Lyon,
France

Abstract

John Kenneth Galbraith was a post-Keynesian and an institutionalist economist. He is famous for having built an unconventional vision of American Capitalism in his postwar trilogy (1952b, 1958, 1967). His socioeconomic analysis deals with subjects such as theory of the firm, managerial revolution, public spending, theory of consumption, price control (1952a), financial crises (1955), power relationship (1983), and more generally with the dynamic of capitalist institutions.

Biography

“Ken” Galbraith (October 15, 1908–April 29, 2006), born in Ontario, was trained after the 1929 Great Crash and passed a Ph.D. at Berkeley in Agricultural Economics (Galbraith 1981). He moved to Harvard during the 1930s and spent a year in Cambridge among Keynes’ fellows (UK). During World War II, Galbraith led the Office for Price Administration. He also oversaw a committee for the evaluation of the effects of American bombings on the German War Economy.

Galbraith then made a journalist career at *Fortune Magazine*. At the political level, he was a leading figure, with his friend Arthur Schlesinger, of the *American for Democratic Action (ADA)* and a special adviser of Democratic candidates running for the Presidency. In 1958, he lamented the “private opulence and public squalor” in his best-seller *The Affluent Society* (1958). J.F. Kennedy named him Ambassador to India in 1961. This experience delayed the publication of his masterwork, *The New Industrial State*, which gave rise to several debates among economists. He spent the rest of his academic life struggling against neoclassical economics and its “conventional wisdom,” against the fact that economists tended to overlook real social issues in favor of purely analytical constructions and intellectual puzzle. For these reasons, John Kenneth Galbraith was one of the most renowned public intellectuals of postwar Anglo-Saxons societies. Indeed he always relied on public enlightenment and developed his ideas for large audiences. For instance, he took part in a TV series entitled *The Age of Uncertainty*, broadcast on the BBC during the 1970s, which aimed to get people acquainted with economic knowledge (Parker 2005).

Innovative and Original Aspects

In his work Galbraith has always insisted on the idea that power *de facto* tends to overthrow *de jure* relationship. For instance, in line with Berle and

Means (1991), he explained that corporations are ruled by those who possess and control crucial factors of production. In the “planning system,” this factor is “organized intelligence” (1967). That is why he thought that a “technostructure” rather than its owners controls modern corporations.

His contribution to Law and Economics is linked to his analysis of the structure of the American economy. In the first opus of his trilogy, *American Capitalism*, he explained that monopoly and oligopoly are now the norm rather than the exception. Thus, the Antitrust Legislation, which is embodied by the Sherman and Clayton Acts, is irrelevant. Whereas some economists pleaded for an enforcement of the antitrust laws against increasing monopoly and monopolistic power, Galbraith explained that competition among giants (corporations, unions, workers’ and buyers’ cooperatives) is susceptible to lead to some equilibrium (1952b). This is at the heart of his idea that “countervailing power” is the new regulative mechanism of the American economy. “In Galbraith’s theory the offsetting power of groups on opposite sides of the market fills the breach, and market power is thereby kept from upsetting the allocative and distributive mechanism” (Peterson 1957). In this regard, he dismissed the Clayton Act for serving as a means to dismantle the bargaining power of Unions rather than the original market power of giant modern corporations.

In *The Affluent Society*, Galbraith explains that antitrust laws were defended by economists since they aim to protect consumers from abuses and to struggle against inequalities. However, in the affluent society, “increased production” is becoming a substitute to “redistribution” (1958). Thus, antitrust laws are anachronistic in American modern capitalism. They are a “charade” according to Galbraith since he considers that big corporations are more efficient to achieve increasing outputs. He notes that “it is at least amusing that during both World Wars the enforcement of the antitrust laws, the traditional design for ensuring greater resource-use efficiency, was suspended” (1958). With the second opus of his trilogy, Galbraith belongs to “a minority of economists,” who consider that “the antitrust laws have outlived their

usefulness” and are unable “to deal with the phenomenon of oligopoly” (Breit and Elzinga 1968).

Galbraith pursues such a line of reasoning in *The New Industrial State*. American conception of antitrust laws is a paradox. On the one hand, there is a deep ideal of competition requiring the absence of market power, thanks to the enforcement of the antitrust laws. On the other, there is an ideal American way of life, which requires big corporations in order to achieve increasing outputs. “No one can ask [a firm] to be an oligopolist for the purposes of capital investment, organization and technology and to be small and competitive for the purposes of prices and allocative efficiency. There is a unity in social phenomena which must be respected” (1967). There are some other contradictions. Regarding antitrust legislation, monopoly is illegal. But oligopoly without overt collusion, even if it has the same economic consequences, is not. Moreover, antitrust laws are sometimes used to condemn two small merging competitors rather than a single firm that controls a greater extent of market shares. His judgment is without appeal: “present enforcement [of antitrust laws] attacks the symbol of market power and leaves the substance” (1967).

Instead of relying on antitrust laws, Galbraith thinks that direct social control of corporations is more efficient. This preference is first understandable because he thinks as every institutionalist that concentration is “the inevitable product of economic imperatives” (Muller 1975). Then, his experience at the head of the Office for Price Administration convinced him that direct negotiation between corporations and public administrations is workable. He argues that price control is a more effective means for avoiding oligopolies’ abuses (1952a, 1967). Muller, considering the relevance of Galbraith’s analysis, adds that direct public control and indirect ones, through antitrust laws, could be complementary so as to reduce the undesirable effects of market power.

This succinct summary of Galbraith’s conception of antitrust laws is one of his numerous innovative analyses. Nowadays such a reflexion appears even more original since Competition Policy, especially in the European Union, has

become a substitute rather than a complement to Industrial Policy.

Impact and Legacy

John Kenneth Galbraith did not give birth to a school of thought, since on the one hand he was on the fringe of academia and on the other hand he did not like to teach. James Galbraith sums up that “the Galbraith paradox is that the great theorist of organizations worked alone – he was an intellectual entrepreneur” (2007). Some of his works are largely known, others deserve to be. The impact of Galbraith’s works on postwar economics was however crucial, since he always refused to abandon his critical glance towards economics.

Cross-References

► [Institutional Economics](#)

References

- Berle A, Means G (1991) Modern corporation and private property. Transaction Publisher, New Brunswick
- Breit W, Elzinga K (1968) Oligopoly, the Sherman Act and “the new industrial state”. *Soc Sci Q* 49:49–57
- Galbraith JK (1952a) A theory of price control. Harvard University Press, Cambridge, MA
- Galbraith JK (1952b) American capitalism. Houghton Mifflin Company, Boston
- Galbraith JK (1955) The great crash. Houghton Mifflin Company, Boston
- Galbraith JK (1958) The affluent society. Houghton Mifflin Company, Boston
- Galbraith JK (1967) The new industrial state. Houghton Mifflin Company, Boston
- Galbraith JK (1981) A life in our time. Houghton Mifflin Company, Boston
- Galbraith JK (1983) Anatomy of power. Houghton Mifflin Company, Boston
- Galbraith JK (2007) Foreword to the new industrial state, 4th edn. Princeton University, Princeton
- Muller WF (1975) Antitrust in a planned economy: an anachronism or an essential complement? *J Econ Issues* 9:159–179
- Parker R (2005) John Kenneth Galbraith, His Life, His Politics, His Economics. University of Chicago Press
- Peterson S (1957) Antitrust and the classic model. *Am Econ Rev* 47:60–78

Games

Ulrich Blum

Faculty of Law and Economics, Martin-Luther-University of Halle Wittenberg, Halle, Germany
University of International Business and Economics, Beijing, China

Abstract

This entry begins by setting games within the general framework to demonstrate their conceptual importance for describing this activity of humankind. Then the concept of merging economics and mathematics into what is called game theory is developed. Descriptions of important properties, definitions, and typical examples of games are provided.

Definition

A game is a structured interaction of at least two players. It consists of rules that set the institutional framework of possible moves. These lead to outcomes that are subject to individual preferences, which generate the incentives to play. If individual ends are incompatible, dilemmata develop, which make the underlying strategic interaction particularly visible.

Games and Science

Even before John von Neumann and Oskar Morgenstern (1944) merged mathematical and economic theory in their groundbreaking work *Theory of Games and Economic Behavior* games were attracting interest in other fields. For instance, Johan Huizinga (1933) looked at games from a psychological perspective showing that they are a fundamental force and a formative element of culture as defined by philosophical anthropology. Games date back to a time before the development of humankind: Many creatures, especially primates, but also other mammals as

well, show game-related behavioral patterns as an expression of rivalry. From a phylogenetic perspective, animals play, and from an ontogenetic perspective, games are incorporated into the existence and the development of the personal individuality of humankind. Games represent a field of tension between conflict – mankind’s aggression – and cooperation, the attempt to bind interests. If we say “a game is captivating,” then we metaphorically combine these antagonistic aspects in one sentence. In history and social sciences, managing conflict and mapping it in parlor games have always accompanied human development. In fact, John von Neumann (1928) looked at this issue in his first book on game theory. The greatest works on strategy, which are often compared today to game-theoretic models, were developed by scholars like Sun Zi (544–466 BC) from China or Carl von Clausewitz (1780–1831) from Germany. The *Art of War* (~500 BC, 2003, 2007) and *On War (Vom Kriege, 1832)* are two classic books about leadership under conditions that involve strategic, operational, or tactical interaction – an aspect of humankind that is also found in important religious works such as the Talmud, the Bible, and the Koran.

The work of John von Neumann (1928) and Oskar Morgenstern triggered not only a broad scientific development of game theory, such as the work by John Nash (1950) on equilibria published half a decade later, it also triggered a very strict observation of reality and its dilemma situations. On the level of political thinking, the concept of zero-sum games is perhaps the best-known outflow and potentially the least understood. In terms of the military, the concept of a credible threat during the Cold War was the direct result of game-theoretic thinking.

Economic and Political Aspects of Game Theory

If human rivalry is to be oriented toward a common good, and if the social contract written by societies is to be productive, the distinction between institutional rules and the moves allowed within this framework is of utmost importance.

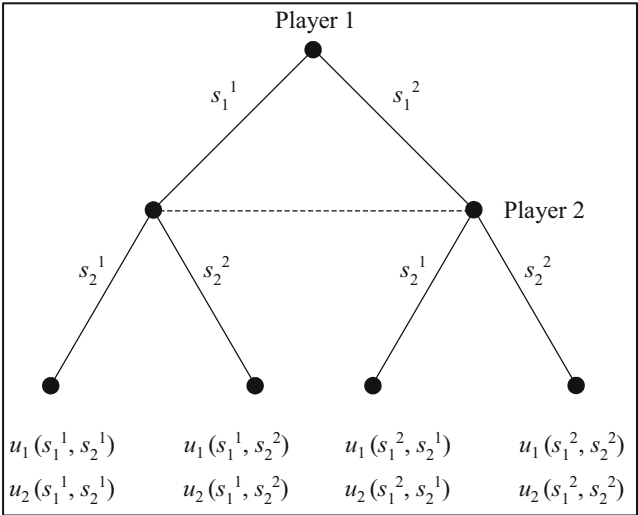
This relates to extending competition in markets to the idea of competition among institutions, especially public institutions. According to the philosophy of Immanuel Kant (1787) and, in particular, his concept of the “categorical imperative,” economic institutions should be designed in such a way that the maxim of their implementation can be accepted as a general rule or regulation. Thus, dilemmata reach beyond individual rivalry, extending to the interaction between institution setters, institutions, and their efficiency in relation to other institutions, thus binding together economics and law. This lies at the core of the very successful German social market economy in the tradition of Walter Eucken (1952) and Ludwig Erhard (1957). Rules are to be set so that they are self-enforcing and follow mankind’s open incentive concept. In such a world, games have two properties: rivalry over better arrangements at the higher institutional level and rivalry within the economic or social world below this roof. This addresses two aspects accentuated by game theory: hierarchy and dynamics. Who and how are the rules set for games, how open are they to long-term change, and who is in charge? Furthermore, with regard to moves (the complete set of which is called strategies), how can players learn to distinguish between “good” moves and “bad” moves? The final chapter will take up this subject using institutional competition as an example.

The rest of the entry is structured as follows: First we will look at the information side of game theory and the prerequisites for efficient activities. Then we will look at how one type of game may switch to other type when circumstances change. Finally, we will look at dynamic contexts.

The General Structure of Games

A game is a model of strategic (or tactical) interaction between players who have sets of possible actions. Since the model captures this interaction between the players, outcomes are interdependent. Dixit and Nalebuff (2010) is an introductory text that requires little formal experience. Standard textbooks are Rasmusen (2006) or Osborne and Rubinstein (1994). Binmore (1991) adds humor

Games, Fig. 1 Extensive form of a two-player game



Games, Fig. 2 Prisoner's dilemma game

		Player 2	
		silence	fink
Player 1	silence	- 1 / - 1 A C	- 9 / <u>0</u>
	fink	<u>0</u> / - 9 B D	<u>- 6</u> / <u>- 6</u>

to theory. Players come up with strategies s and have preferences that are modeled by ordinal pay-offs or cardinal functions, for instance, production functions $p(.)$ or utility functions $u(.)$. In its extensive form, the game starts with two possible moves from the first player's strategy set (s_1^1 and s_1^2). This is followed by the moves of the second player. Figure 1 illustrates the sequential structure.

One of the best-known games is the prisoner's dilemma. The fundamental idea is straightforward: A murder has been committed and the police apprehend two suspects tramping in the area. They lock them up separately and have them interrogated by an attorney. The vagrants would be best off if they both remained silent; then they would be sentenced to only 1 year each for vagrancy. However, if one of them finks (cheats, defects) and accuses the other, he will be set free and the other will be jailed for 9 years. If both accuse the other person, they will both be jailed for 6 years. To all extent and purposes, it

would be socially efficient to remain silent. However, under the absence of coordination, the best possible strategy is to accuse the other, which, however, produces an inferior outcome (Fig. 2).

If both defendants remain silent, the outcome is better than if they accuse each other (both payoffs are strictly higher values). The best strategies for one player against the other are underlined. However, this advantage does not exist with respect to the other two pairs of payoffs. In addition, the payoff resulting from mutual accusation remains unchanged in the sense that neither prisoner can deviate without disadvantaging himself. This leads to the following definition:

A set of player strategies is a Nash equilibrium if no deviation of any individual player's strategies (i.e. given the strategies of the other players) results in a better outcome for that player.

Results in field D are thus underlined twice as they are the Nash equilibrium.

Games, Fig. 3 Tragedy of the commons game

		all other players				
		cooperate	defect	act at 10% in consort	act at 30% in consort	act at 60% in consort
player A	cooperate	10/10	2/12	2.8/11.8	4.4/11.4	6.8/10.8
	defect	12/2	3/3	11.1/2.1	9.3/2.3	6.6/2.6

One of the most interesting games in the genre of prisoner’s dilemmas is the tragedy of the commons (Hardin 1968). In this game, a joint resource is distributed. An overexploitation and destruction of the resource can only be prevented if all parties agree to the terms of exploitation. However, under the absence of coordination, each individual will seek to achieve his/her own optimal outcome, and the resource will break down. The resource-efficient outcome will only be achieved if there is a high enough expectancy to cooperate. Figure 3 shows that this is the case if 60% act in consort. If player A then decides to defect, he will be worse off (6.6 instead of 6.8 points). This is not yet the case when only 30% act in consort.

There are many interesting dilemma situations: The chicken game models a situation where two cars approach each other on a narrow road with the question being, which car gives way to which? The battle of the sexes models a situation where a couple wants to spend their weekend away from home; he wants to go to a soccer game, while she wants to go to the theater. Both strategies on their own are incompatible with one another, and there is no joint utility-maximizing outcome. The problem in this game is that mixed strategies are not possible. A compromise can only be found, and socially efficient results can only emerge, if a mix between the two is possible, for instance, if two weekends are taken into consideration.

From these discussions, two other notions can easily be derived:

- In zero-sum games, the extent of the loss on the one side is identical to the gain on the other side.
- Zero-sum games are a special variant of constant sum games.
- In policy discussions, people think situations are zero-sum situations even when they truly are not. For instance, following Ricardo’s classic theory of comparative costs, trade is a redistribution that makes both parties better off.

Games with Functions as Payoffs

Often, games are set within a dynamic context as they develop in an evolutionary way or are played repetitively. Players are then able to gain experience, which may encourage them not to fall into inefficient Nash equilibria, e.g., into the prisoner’s dilemma trap. Another aspect is the changing of payoffs in dynamic utility or production functions, as in the following case which combines the economics of competition with the legal side of good regulations and governance. The example follows (Blum et al. 2005, section 7.2.2). In a competitive environment, the antitrust organization sets penalties, p , if competition rules are broken. The standard reward from competition is

Games, Fig. 4
Competition game

		Player 2	
		obeys rules	cheats on rules
Player 1	obeys rules	ordo-economy	one-sided power
		$s + f - g(e)$	$s - p$
	cheats on rules	$s + f - g(e)$	$s - p - g(e)$
		$s - p - g(e)$	$-p$
		One-sided power	destruction of market

Games, Table 1 Basic game structures

Basic game	Line player (player 1)
Prisoner's dilemma	$B_1 > A_1 > D_1 > C_1$
Chicken game	$B_1 > A_1 > C_1 > D_1$
Assurance game	$A_1 > D_1 > C_1$ and $A_1 > B_1$
Social optimum	$A_1 > C_1 > D_1$ and $A_1 > B_1$

$s \geq 0$; it becomes zero if all players violate competition laws. An extra externality or fringe profit, $f \geq 0$, emerges if all players comply with the rules. The effort of players is given as $e \geq 0$ and the related disutility is $g(e) \geq 0$ (Fig. 4).

Table 1 lists three types of games and the structure of their payoffs, to which we have added a social-optimum outcome.

If we evaluate the outcomes, we obtain:

- **Prisoner's dilemma:** (player 1, $B_1 > A_1 > D_1 > C_1$; player 2, $C_2 > A_2 > D_2 > B_2$)

$$B_1 > A_1 (\text{or } C_2 > A_2 \text{ resp.}) \Leftrightarrow s - p > s + f - g(e) \Leftrightarrow f < g(e) - p. \quad (\text{P1})$$

$$A_1 > D_1 (\text{or } A_2 > D_2 \text{ resp.}) \Leftrightarrow s + f - g(e) > -p \Leftrightarrow f > g(e) - (p + s). \quad (\text{P2})$$

$$D_1 > C_1 (\text{or } D_2 > B_2 \text{ resp.}) \Leftrightarrow -p > s - g(e) - p \Leftrightarrow g(e) > s. \quad (\text{P3})$$

- **Chicken game:** (player 1, $B_1 > A_1 > C_1 > D_1$; player 2, $C_2 > A_2 > B_2 > D_2$)

$$B_1 > A_1 (\text{or } C_2 > A_2 \text{ resp.}) \Leftrightarrow s - p > s + f - g(e) \Leftrightarrow f < g(e) - p. \quad (\text{C1})$$

$$A_1 > C_1 (\text{or } A_2 > B_2 \text{ resp.}) \Leftrightarrow s + f - g(e) > s - g(e) - p \Leftrightarrow f > -p. \quad (\text{C2})$$

$$C_1 > D_1 (\text{or } B_2 > D_2 \text{ resp.}) \Leftrightarrow s - g(e) - p > -p \Leftrightarrow g(e) < s. \quad (\text{C3})$$

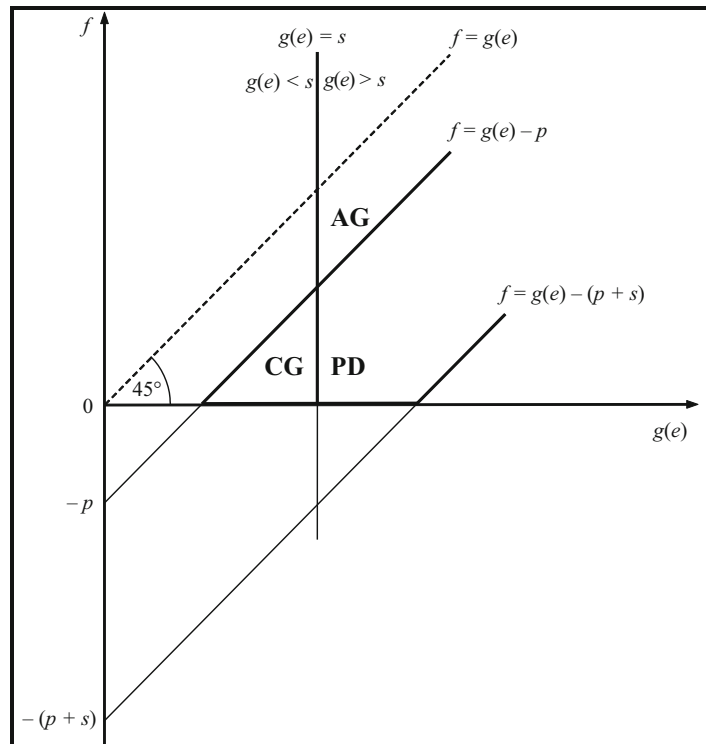
- **Assurance game:** (player 1, $A_1 > D_1 > C_1$ and $A_1 > B_1$; player 2, $A_2 > D_2 > B_2$ and $A_2 > C_2$)

$$A_1 > D_1 (\text{or } A_2 > D_2 \text{ resp.}) \Leftrightarrow s + f - g(e) > -p \Leftrightarrow f > g(e) - (p + s). \quad (\text{A1})$$

$$D_1 > C_1 (\text{or } D_2 > B_2 \text{ resp.}) \Leftrightarrow -p > s - g(e) - p \Leftrightarrow g(e) > s. \quad (\text{A2})$$

$$A_1 > B_1 (\text{or } A_2 > C_2 \text{ resp.}) \Leftrightarrow s + f - g(e) > s - p \Leftrightarrow f > g(e) - p. \quad (\text{A3})$$

Games, Fig. 5 Efficient areas of the competition game. (Source: Blum et al. 2005, p. 191)



If we plot these results on a graph, we obtain the following result, which is illustrated in Fig. 5. It seems that good competition regimes depend on credible penalties; sufficient externalities for good competitive behavior that can be internalized, i.e., through long-term growth; and a limited disutility of effort, i.e., limited levels of transaction costs for compliance.

true abstraction of reality – otherwise, failure can be bitter. In summer 2015, Greek finance minister Varoufakis made this experience when trying in vain to provoke the lender Troika in a poker-type manner and through impulsive behavior in order to improve his negotiation position, a strategy already proposed 50 years ago in the Strategy of Conflict by (Schelling 1960).

Cross-References

What Can Be Learned from Game Theory

Game theory is a formal tool that shows the extent to which the outcomes of different decisions interact, especially if players have opposing goals. It provides a rigorous analytical environment for evaluating efficiency in static or dynamic environments. Although the ability to model reality in a very concrete way is limited, the overall structural messages are extremely important in understanding how to manage dilemma situations. In addition, players should check whether the game is a

- ▶ Abuse of Dominance
- ▶ Blackmail
- ▶ Bounded Rationality
- ▶ Cartels and Collusion
- ▶ Coase and Property Rights
- ▶ Corruption
- ▶ Credibility
- ▶ Equilibrium Theory
- ▶ Governance
- ▶ Information Deficiencies in Contract Enforcement
- ▶ Institutional Economics
- ▶ Rationality

References

- Blum U, Dudley L, Leibbrand F, Weiske A (2005) *Angewandte Institutionenökonomik: Theorien, Modelle, Evidenz*. Gabler, Wiesbaden
- Binmore, K., 1992, *Fun and Games*, D.C. Health Company, Lexington (MA)
- von Clausewitz C (1832) *Vom Kriege*. Dümmlers Verlag, Berlin; (2008) *On war*. Oxford World's Classics, Oxford
- Dixit AI, Nalebuff BJ (2010) *The art of strategy, a game theorists guide to success in business and life*. Borton, New York
- Erhard L (1957) *Wohlstand für Alle*. Econ, Düsseldorf; (1958) *Prosperity through competition*. Frederick A. Praeger, New York
- Eucken W (1952/1962) *Grundsätze der Wirtschaftspolitik*. J.C.B. Mohr, Tübingen/Zürich
- Hardin G (1968) The tragedy of the commons. *Science* 162:1243–1248
- Huizinga J (1933) *Over de grenzen van spel en ernst in de kultur*. Rector's speech, Leiden
- Kant I (1787) *Kritik der reinen Vernunft*. Königsberg; (2015) *Critique of practical reason*, Cambridge texts in the history of philosophy. Cambridge
- Nash G (1950) *Non-cooperative games*, PhD thesis, Mimeo, Princeton
- Neumann J (1928) Zur Theorie der Gesellschaftsspiele. *Math Ann* 100:295–300
- Neumann JV, Morgenstern O (1944) *Theory of games and economic behavior*. Princeton University Press, Princeton
- Osborne M, Rubinstein A (1994) *A course in game theory*. The MIT-Press, Cambridge
- Rasmusen E (2006) *Games and information: an introduction to game theory*. Blackwell, Oxford
- Sun Zi (2003) commented by CAO Cao et al. (【春秋】孙武撰,【三国】曹操 等注,宋本十一家注孙子,上海:上海古籍出版社). Shanghai Ancient Books Publishing House, Shanghai
- Sun Zi (~500 BC, 2007) *The art of war*. Columbia University Press, New York
- Schelling, Th., 1960 (1981), *The Strategy of Conflict*, Harvard University Press, Boston (Mass).

GE/Honeywell

Philipp Schumacher
Wildau Institute of Technology, Technical
University of Applied Science, Wildau,
Brandenburg, Germany

Definition

The GE/Honeywell case, one of the most controversial merger cases in the history of European

merger control, was subject to a long and heated transatlantic debate on the vertical and conglomerate aspects of the European Commission's 2001 prohibition decision. However, horizontal overlaps in the specific markets for aircraft and marine engines were, ultimately, the only reason given by the Court of First Instance for upholding the controversial decision.

The GE/Honeywell Case

In 2001, the attempted US\$ 45 billion merger between General Electric Company (GE) and Honeywell International, Inc. could have become the largest deal of its kind in industrial history. However, while the U.S. Department of Justice Antitrust Division (DoJ) allowed the merger on the condition that minor structural remedies be carried out, the European Commission prohibited the merger (European Commission 2001a, b). Although the Court of First Instance (CFI) criticized the Commission's analysis of the vertical and conglomerate effects of the merger, it dismissed GE's appeal in 2005, upholding the prohibition decision based on alleged horizontal effects.

GE is a diversified conglomerate and participates in a large variety of markets worldwide. The main industries it served in 2001 were: aircraft engines, transportation systems, industrial systems, power systems, plastics, lighting, medical systems, domestic appliances, broadcasting (via NBC), financial services, and information services. GE's only independent competitors in the supply of civil aircraft engines were Pratt & Whitney, part of United Technologies (USA), and Rolls-Royce (UK). GE also owned one of the world's largest lease aircraft fleets through GE Capital Aviation Services (GECAS).

At time of merger proposal, Honeywell International was a diversified technology and manufacturing company, serving customers worldwide with aerospace products and services; control technologies for buildings, homes, and industry; automotive products turbochargers; and speciality materials. Its aerospace division had a strong position in flight control systems, the

combination of cockpit displays, computers, and software that control the aircraft.

There are a large number of publications available on the GE/Honeywell case which may be categorized by five groups. Firstly, there are official publications from the European Commission, the CFI, and the US DoJ. A second group of literature mainly investigates the divergent outcomes and discusses which merger control system is superior, comprising Varian (2001), Jin (2002), Kolasky (2002a, b, c, d), Burnside (2002), Hochstadt (2003), Emch (2004), Evans and Salinger (2002), Pflanz and Caffara (2002), Ruffner (2003), Morgan and McGuire (2004), Schnell (2004), Fox (2002a, b), Girardet (2006). Thirdly, there is a large body of economic and legal literature focusing on parts of the EC decision. Here, the focus is particularly on *conglomerate effects*, i.e., bundling and leveraging, which received most of the scientific community's attention, (e.g., Briggs and Rosenblatt 2001, 2002; Drauz 2002; Emch 2003; Reynolds and Ordovery 2002; Choi 2001; Nalebuff 2002; Grant and Neven 2005; Patterson and Shapiro 2001). The fourth group of literature analyses the ruling of the Court of First Instance comprising Howarth (2006), Montag (2006), Pflanz (2006), Kolasky (2006), Weidenbach and Leupold (2006), Schwaderer (2006, 2007), Fox (2006), von Bonin (2006). Representing the fifth group, Schumacher (2010, 2013) shows that in contrast to the decision of the European Commission and the CFI, the merger would not have led to anti-competitive *horizontal effects*. According to the two-level approach of the European Commission, firstly the competitive situation of the engine manufacturer in relation to its direct customer, the aircraft or marine vessel manufacturer, and secondly the competitive effects in the respective market of end-use applications must be considered. Schumacher (2010) shows that GE was not in the position to exert market power prior to the merger and would not have been *ex post*. Applying the new EC Merger Regulation of 2004 and its Horizontal and Non-horizontal Merger Guidelines or its 2001 predecessor, an overall competitive appraisal has to go beyond the definition of the relevant market and the calculation of structural indicators such as market shares.

Assessment of the competitive closeness, firstly, of the parties' aircraft engines and, secondly, of aircraft powered by these engines requires the analysis of bidding data and technical characteristics. The engine markets affected are world-wide bidding markets characterized by volatile market shares, the high importance of after-sales revenue and profitable outside options. Based on publicly available data, information from aerospace professionals and industry knowledge, the conclusion of an effects-based analysis of GE/Honeywell is diametrically opposed to the Commission's decision regarding both horizontal and nonhorizontal aspects. As shown in similar complex aerospace mergers GE/Smiths Aerospace (2007) and UTC/Goodrich (2012), the "more economic approach" to European merger control, including empirical analysis firmly grounded in industry-specific expertise, can help to avoid a pro-competitive merger being prohibited again based on speculative theories of harm inconsistent with the realities of the case.

Cross-References

- Efficiency
- Horizontal Effects
- Merger Control
- Merger Remedies

References

- Briggs JD, Rosenblatt HT (2001) A bundle of trouble: the aftermath of GE/Honeywell. *Antitrust mag* 16(1): 26–31
- Briggs JD, Rosenblatt HT (2002) GE/Honeywell – live and let die: a response to Kolasky and greenfield. *George Mason Law Rev* 10(3):459–470
- Burnside A (2002) GE, honey, I sunk the merger. *Euro Compet Law Rev* 23(2):107–110
- Choi JP (2001) A theory of mixed bundling applied to the GE/Honeywell merger. *Antitrust Mag* 16(1):32–33
- Court of First Instance (2005) Case T-210/01, General Electric Company vs. Commission of the European Communities. Available from: <http://eur-lex.europa.eu/legal-content/EN/TEXT/PDF/?uri=CELEX:62001TJ0210&rid=5>. Accessed 10 June 2017
- Drauz G (2002) Unbundling GE/Honeywell: the assessment of conglomerate mergers under EC

- competition law. *Fordham Int Law J* 25(4): 885–908
- Emch E (2003) GECAS and the GE/Honeywell merger: a response to Reynolds and Ordoover, *Economic Analysis Group Discussion Paper*, EAG 03-13. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=451961. Accessed 10 June 2017
- Emch E (2004) “Portfolio effects” in merger analysis: differences between EU and U.S. *Pract Recomm Future, Antitrust Bull* 49(1–2):55–100
- European Commission (2001a) Case No. COMP/M.2220 General Electric/Honeywell, OJ (2004) L48/1
- European Commission (2001b) The commission prohibits GE’s acquisition of Honeywell, Rapid – Press Releases – Europa, July 3. Available from: http://europa.eu/rapid/press-release_IP-01-939_en.htm. Accessed 10 June 2017
- European Commission (2007) Case No. COMP/M.4561 GE/Smiths Aerospace, OJ (2007) C133/02
- European Commission (2012) Case No. COMP/M.6410 – UTC/Goodrich, OJ (2012) C388/04
- Evans DS, Salinger MS (2002) Competition thinking at the European Commission: lessons from the aborted GE/Honeywell merger. *George Mason Law Rev* 10:489–532
- Fox EM (2002a) Mergers in global markets: GE/Honeywell and the future of merger control. *Univ Pennsylvania J Int Econ Law* 23:457–468
- Fox EM (2002b) U.S. and European merger policy – fault lines and bridges. *George Mason Law Rev* 10:471–488
- Fox EM (2006) The European court’s judgement in GE/Honeywell – not a poster child for comity or convergence. *Antitrust mag* 20(2):77–81
- Girardet F (2006) Internationales Fusionskontrollrecht – Konflikt und Konvergenz – Eine Untersuchung mit Schwerpunkt auf dem Europäischen und US-amerikanischen Fusionskontrollrecht anhand der Zusammenschlussvorhaben Boeing/MDD und GE/Honeywell. Peter Lang, Frankfurt am Main
- Grant J, Neven DJ (2005) The attempted merger between general electric and Honeywell – a case study of transatlantic conflict. *J Compet Law and Econ* 1(3): 595–633
- Hochstadt ES (2003) The Brown Shoe of European Union competition law. *Cardozo Law Rev* 24:287–395
- Howarth D (2006) The court of first instance in GE/Honeywell. *Eur Compet Law Rev* 27(9):485–493
- Jin P (2002) Turning competition on its head: economic analysis of the EC’s decision to bar the GE-Honeywell merger. *Northwest J Int Law Bus* 23:187–211
- Kolasky WJ (2002a) Conglomerate mergers and range effects: it’s a long way from Chicago to Brussels. *George Mason Law Rev* 10:533–550
- Kolasky WJ (2002b) United States and European competition policy: are there more differences than we care to admit? Speech at the European Policy Center, Brussels, April 10 Available from: <https://www.justice.gov/atr/file/519811/download>. Accessed 10 June 2017
- Kolasky WJ (2002c) North Atlantic competition policy: converging toward what?, Speech at the BIICL second annual international and comparative law conference, London, May 17. Available from: <https://www.justice.gov/atr/file/519801/download>. Accessed 10 June 2017
- Kolasky WJ (2002d) GE/Honeywell: continuing the transatlantic dialog. *Univ Pennsylvania J Int Econ Law* 23:513–537
- Kolasky WJ (2006) GE/Honeywell: narrowing, but not closing, the gap. *Antitrust mag* 20(2):69–76
- Montag F (2006) Das GE-Honeywell-Urteil des EuG: Absage an konglomerate und vertikale Effekte in der europäischen Fusionskontrolle? In: Brinker I, Scheuing DH, Stockmann K (eds) *Recht und Wettbewerb – Festschrift für Rainer Bechtold zum 65. Geburtstag*. Verlag C.H.Beck, München, pp 339–355
- Morgan EJ, McGuire S (2004) Transatlantic divergence: GE-Honeywell and the EU’s merger policy. *J Eur Publ Policy* 11(1):39–56
- Nalebuff B (2002) Bundling and the GE-Honeywell merger. Working paper No. 22, Working paper series ES – Economics/Strategy. Yale School of Management. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=327380. Accessed 10 June 2017
- Patterson DE, Shapiro C (2001) Transatlantic divergence in GE/Honeywell: causes and lessons. *Antitrust Mag* 16(1):18–25
- Pfanz M (2006) Economic analysis and judicial review: the CFI on GE/Honeywell. *Zeitschrift Wettbewerbsrecht* 4(2):139–178
- Pfanz M, Caffara C (2002) The economics of G.E./Honeywell. *Eur Compet Law J* 23:115–121
- Reynolds RJ, Ordoover JA (2002) Archimedean leveraging and the GE/Honeywell transaction. *Antitrust Law J* 70:171–198
- Ruffner T (2003) The failed GE/Honeywell merger: the return of portfolio-effects theory. *DePaul Law Rev* 52:1285–1231
- Schnell DK (2004) All bundled up: bringing the failed GE/Honeywell merger in from the cold. *Cornell Int Law J* 37:217–262
- Schumacher P (2010) Revisiting the General Electric/Honeywell case applying the European Commission’s more economic approach. Shaker Verlag, Aachen
- Schumacher P (2013) The EU’s flawed assessment of horizontal aspects in GE/Honeywell: re-visiting the last pillar of the European prohibition decision. *Eur J Law Econ*, 2013 35(2):211–240
- Schwaderer MA (2006) Anything new after GE/Honeywell. In: Bösch KV, Füller JT, Wolf M (eds) *Variationen im Recht – Festbeigabe für Franz Jürgen Säcker*. Berliner Wissenschafts-Verlag, Berlin, pp 151–161
- Schwaderer MA (2007) Conglomerate merger analysis – the legal context: how the European courts’ standard of proof put an end to the ex ante assessment of leveraging. *Zeitschrift für Wettbewerbsrecht* 5(4):482–514
- Varian HR (2001) In Europe, G.E. and Honeywell Ran Afoul of 19th-century thinking, *New York Times* June 28, C2
- Von Bonin A (2006) Vertikale und konglomerate Zusammenschlüsse nach dem Urteil GE/Kommission. *Wirtschaft und Wettbewerb* 56(5):466–476
- Weidenbach G, Leupold H (2006) Das GE/Honeywell-Urteil des EuG – Spannende Rechtsfragen im Überfluss. *Europäisches Wirtsch Steuerrecht* 17(4):154–163

Gender Diversity

Isabel-María García-Sánchez

Universidad de Salamanca, Salamanca, Spain

From a business perspective, gender diversity represents the heterogeneity of a workforce in terms of gender, including that of top management and the board of directors. Gender occupational segregation is a phenomenon which is present in most countries, regardless of economic conditions and the existence of anti-discrimination laws that ensure full citizenship for women. However, the presence of women in positions of power and decision-making in organizations and in public and private institutions is virtually nonexistent. Most analysts agree that organizational dynamics are the main obstacle to the career advancement of women and that these are still dominated by androcentric values that exclude them from decision-making. This still does not solve the problem of reconciling work with privacy.

The inclusion criterion for gender diversity in organizations corresponds to a strategy of promoting the participation of women in positions of responsibility. The central idea of the concept of diversity is the maximum utilization of the potential of heterogeneous groups, i.e., how they are different in terms of gender. Variability is emphasized, so that each person is valued for what they are and what they can contribute with regard to their personal characteristics. A diversity strategy gives organizations the ability to attract and retain diverse talents which are representative of both sexes. In this sense, gender diversity recognizes that there is no single way to work but that labor plurality may be advantageous for an organization, as it encourages innovation and complementary perspectives.

Generalized differences were identified in the gender management practices due to women and men having several approaches to overall style, decision-making, and interpersonal relationships. Several authors, such as Robinson and Dechant (1997), suggest that the incorporation of women into the main management bodies entails

a significant increase in business due to the economic result set of skills, knowledge, and personal experiences of women. Women have a higher ability to achieve a better understanding of the demands of customers, increasing their chances of entering new markets (Carter et al. 2003), in order to promote an effective resolution of problems and evaluate alternatives (Rose 2007) and to hire the best executives and directors to limit bias against minorities (Campbell and Minguez-Vera 2008). All of this will result in positive signals being sent out to the labor market and the generation of capital for products that will reduce the costs borne by companies (Rose 2007).

Additionally, the proponents of gender diversity on the board feel that their presence is meant to increase business organizational values and business results through the creation of new knowledge and perspectives (Carter et al. 2003). These are the key factors in the process of adopting sustainable forms of behavior in the economic sphere. Furthermore, according to Calas and Smircich (2004), with regard to the traditional feminine qualities, rhetoric is a rationale that must include the value being held for radical changes and for configuring a new business model that creates value not only for shareholders but also for different stakeholders.

The above characteristics are specified in the globally proven fact that women are more sensitive to other perspectives of corporate behavior, have a greater ability to identify the needs of different stakeholders, and can provide more opportunities to satisfy them (Brennan and McCafferty 1997). Thus, counselors need to be more participative and democratic and show more group spirit than their colleagues in promoting participation and dialogue within the board and, by extension, in encouraging greater communication and paying more attention to the interests of the various stakeholders (Eagly et al. 2003). They also need to have had a greater number of experiences outside the business world (Hillman et al. 2002). All of the above creates, ultimately, an increased focus on CSR (Singh and Vinnicombe 2004).

However, from the point of view of value creation, the presence of more women may also involve a number of disadvantages associated

mainly with the diversity of opinions and the appearance of discrepancies that may cause delays in decision-making (Smith et al. 2006).

References

- Brennan N, McCafferty J (1997) Corporate governance practices in Irish companies. *Ir J Manag* 18:116–135
- Calas MB, Smircich L (2004) Revisiting ‘dangerous liaisons’ or does the ‘feminine-in-management’ still meet ‘globalization’? In: Frost PJ, Nord WR, Krefting LA (eds) *Managerial and organizational reality: stories of life and work*. Prentice-Hall, Upper Saddle River, pp 467–481
- Campbell K, Mínguez-Vera A (2008) Gender diversity in the boardroom and firm financial performance. *J Bus Ethics* 83:435–451
- Carter DA, Simkins BJ, Simpson WG (2003) Corporate governance, board diversity and firm value. *Financ Rev* 38:33–53
- Eagly AH, Johannesen-Schmidt MC, van Engen M (2003) Transformational, transactional and laissez-faire leadership styles: a meta-analysis comparing women and men. *Psychol Bull* 129:569–591
- Hillman AJ, Cannella AA Jr, Harris IC (2002) Women and racial minorities in the boardroom: how do directors differ? *J Manag* 28(6):747–763
- Robinson G, Dechant K (1997) Building a business case for diversity. *Acad Manag Exec* 11(3):21–31
- Rose C (2007) Does female board representation influence firm performance? The Danish evidence. *Corp Govern An Int Rev* 15(2):404–413
- Singh V, Vinnicombe S (2004) Why so few women directors in top U.K. Boardrooms? Evidence and theoretical explanations. *Corp Govern* 12(4):479–488
- Smith N, Smith V, Verner M (2006) Do women in top management affect firm performance? A panel study of 2,500 Danish firms. *Int J Perform Manag* 55(7):569–593

Genocide

Charles H. Anderton¹ and Jurgen Brauer²

¹College of the Holy Cross, Worcester, MA, USA

²Hull College of Business, Georgia Regents University, Augusta, GA, USA

Abstract

The entry defines crimes of mass atrocity, such as genocide, provides data on their intensity, and discusses domestic and international

institutions formed to address the crimes. In addition, the entry briefly surveys the economic literature on mass atrocity crimes from theoretical and empirical perspectives. It pays particular attention to the economics of international law.

Definition

By combining the Greek *genos* (a people, tribe, race) and the Latin *cide* (to kill), Raphael Lemkin (1944, p. 79) invented the word genocide. Article 2 of the 1948 United Nations (UN) Convention on the Prevention and Punishment of the Crime of Genocide defines *genocide* as “any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such: (a) Killing members of the group; (b) Causing serious bodily or mental harm to members of the group; (c) Deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part; (d) Imposing measures intended to prevent births within the group; (e) Forcibly transferring children of the group to another group” (United Nations 1948). As of 31 October 2014, only 146 UN members are Party to the Convention.

Data, Mass Atrocity Crimes, and Institutions

Data

In a summary of large-sample datasets on atrocities involving civilians, Anderton (in Anderton and Brauer forthcoming; henceforth “in A/B forthcoming”) identifies 201 distinct cases of state-sponsored genocides and mass atrocities (GMAs) from 1900 to 2013, 43 state-perpetrated genocides from 1955 to 2013, and 34 GMAs perpetrated by non-state groups from 1989 to 2013. Some well-known genocides include the Armenian genocide (1915–1918; estimated fatalities ~1.5 million), the Holocaust (1933–1945; ~10 million), Cambodia (1975–1979; ~1.9 million), Rwanda (1994; ~0.8 million), and Sudan-Darfur (2003–2011; ~0.4 million). A cautious estimate of

intentional civilian fatalities associated with the 202 state-perpetrated GMAs since 1900 is 84 million. Less well-known are non-state perpetrated atrocities such as conducted by the so-called Islamic State, with estimated fatalities of 8,198 from 2005 to 2013 (see Uppsala Conflict Data Program at <http://www.pcr.uu.se>).

Mass Atrocity Crimes

As defined in the UN Convention, *genocide* is the intentional destruction, in whole or in part, of a specific group of people. In non-genocidal *mass killing*, perpetrators do not seek to destroy a group as such (Waller 2007, p. 14). *Crimes against humanity* encompass widespread or systematic attacks against civilians involving inhumane means such as extermination, forcible population transfer, torture, rape, and disappearances. *War crimes* are grave breaches of the Geneva Conventions including willful killing, torture, willfully causing great suffering or serious injury, and extensive destruction and appropriation of property. *Ethnic cleansing* is the removal of people of a particular group from a state or region using means such as forced migration or mass killing (Pergorier 2013). *Violence against civilians* (VAC) can incorporate mass atrocities but it also includes incidents that are relatively small, specifically, less than 1,000 per case or per year. Along with genocide, crimes against humanity, war crimes, and ethnic cleansing are both legal and scholarly terms.

At the Nuremberg trials of 1945–1946, the International Military Tribunal found none of the accused guilty of crimes committed prior to the outbreak of war on 1 September 1939; litigation was limited to atrocities during wartime (Schabas 2010, pp. 126–127). Lemkin's (1944) and the UN's conceptions of genocide were novel precisely because they spoke to criminal acts committed in wartime or in peacetime. Nevertheless, the UN definition of genocide has been subject to critical scrutiny by scholars, for instance in regard to groups left out (e.g., political), how to identify "intent," the inability of the Convention to prevent genocide, the relationship of genocide to other atrocities, and misuse of the term (e.g., Curthoys and Docker 2008). The UN definition of genocide

has not expanded since 1948 to include other groups, but international criminal law has evolved. Schabas (2010, p. 141) maintains that the expanded concept of crimes against humanity has "emerged as the best legal tool to address atrocities" and "genocide as a legal concept remains essentially reserved for the clearest cases of physical destruction of national, ethnic, racial, or religious groups."

International and Domestic Institutions

The twentieth and twenty-first centuries display the emergence and growth of international and domestic laws designed to prevent, punish, and/or foster restitution for atrocity crimes. Table 1 shows a selection of such institutions as well as sources that provide further information. Adjudication of mass atrocity crimes began in earnest following World War I with the establishment of the Turkish Military Tribunal (TMT) (1919–1920), which prosecuted organizers of the Armenian genocide. The trials, characterized as "a milestone in Turkish legal history" (Dadrian 1997, p. 30), revealed the systematic planning behind the genocide, enrichment of perpetrators through looting of victims' assets, and the lack of military necessity for the forced relocation of Armenians. However, the TMT convicted only 15 men among the hundreds who orchestrated the genocide (Dadrian 1997).

Following World War II, the International Military Tribunal (IMT) at Nuremberg was established in which leading officials were tried for war crimes and crimes against humanity (1945–1946). Twelve Nazi leaders received the death sentence and many others were given long jail terms. The trials had an important influence on the growth of international criminal law including the 1948 Genocide Convention, the International Criminal Tribunal for the Former Yugoslavia (ICTY) established in 1993, the International Criminal Tribunal for Rwanda (ICTR) established in 1994, and the International Criminal Court (ICC) ratified in 2002. As of December 2014, the ICTY had indicted 161 people for atrocity crimes associated with the wars in the former Yugoslavia in the 1990s. As of December 2014, the ICTR had indicted 95 people for atrocity

Genocide, Table 1 Selection of legal institutions, jurisprudence, and international norms related to genocide prevention and post-genocide justice

Selection of legal institutions (or norms)	Selection of sources for further information
International	
International Military Tribunal at Nuremberg (IMT) (1945–1946)	US Holocaust Memorial Museum (http://www.ushmm.org)
Convention on the prevention and punishment of the crime of genocide (1948, 1951)	United Nations (https://treaties.un.org), Schabas (2010), US Holocaust Memorial Museum (http://www.ushmm.org)
International Criminal Tribunal for the former Yugoslavia (ICTY) (1993)	United Nations (http://www.icty.org)
International Criminal Tribunal for Rwanda (ICTR) (1994)	United Nations (http://www.unictt.org)
International Criminal Court (ICC) (2002)	International Criminal Court (http://www.icc-cpi.int)
Norms on the responsibilities of transnational corporations and other business enterprises with regard to human rights (2003)	Hillemanns (2003)
Extraordinary Chambers in the Courts of Cambodia (ECCC) (2003)	Extraordinary Chambers in the Courts of Cambodia (http://www.eccc.gov.kh/en)
Responsibility to Protect (R2P) (2005)	United Nations (UN A/RES 60/1 www.un.org/Docs ; http://www.un.org/en/preventgenocide)
Domestic	
Turkish military tribunal (1919–1920)	Dadrian (1997)
US Alien Tort Claims Act (1789, 1980)	Michalowski (2013)
Prosecution of civilian atrocities (not necessarily genocide) in domestic courts (includes Nuremberg and others)	Schabas (2003), Prevent genocide international (http://www.preventgenocide.org)

crimes associated with the 1994 civil war and genocide and it established the legal precedent that mass rape during wartime is genocidal. Following the huge backlog of cases awaiting trial in Rwanda, the government turned to the Gacaca court system, based on traditional law developed within communities (Bornkamm 2012). As of October 2014, the ICC has indicted 36 individuals for atrocity crimes including three current or former heads of state: Omar al-Bashir (Sudan), Uhuru Kenyatta (Kenya), and Laurent Gbagbo (Côte d'Ivoire). According to its Statutes, the ICC has jurisdiction with respect to genocide, crimes against humanity, and war crimes.

Another important international genocide development occurred at the 2005 UN World Summit, in which member states unanimously adopted a norm known as the *Responsibility to Protect* (R2P). R2P was part of the impetus for UN Security Council Resolution 1973, passed on 17 March 2011, which authorized member states to take actions, including enforcement of a no-fly zone, to protect civilians from attacks by the

Libyan military. Nevertheless, the UN's R2P resolution has no legal force (UN Doc. A/RES/60/1, paras 138, 139).

Following the Nuremberg trials and the UN Convention, several dozen nations have developed domestic laws to put on trial suspected Nazi war criminals and/or perpetrators of more recent atrocities (Schabas 2003). For example, in 2000 the Chilean Court of Appeals lifted former President Augusto Pinochet's immunity from prosecution, paving the way for trial for his role in civilian atrocities that occurred during his leadership (Pinochet died prior to any conviction). The case is notable not only because it involved a state's prosecution of its former leader, but also because Pinochet's initial arrest occurred in London based on an application of "universal jurisdiction" by European judges. Universal jurisdiction is a principle by which a state (or states, in the Pinochet case) asserts its right to prosecute a person for an alleged crime regardless of the crime's location and the accused's residence or nationality (Lunga 1992).

Not shown in Table 1 are formalized norms within for-profit and non-profit organizations designed to inhibit complicity in atrocities. The ICC followed the ICTY and ICTR in having jurisdiction only over “natural persons” and not “legal persons” (Cernic 2010, p. 141), which ruled out prosecution of corporations complicit in genocide (individual agents within corporations can be tried). Multinational corporations have been complicit in genocide in many cases, but have not usually faced prosecution (Kelly 2012). Nevertheless, there have been legal efforts, including use of the US Alien Tort Claims Act, to bring litigation against corporations for alleged complicity in atrocities and other human rights abuses. Such litigation is leading companies to develop norms to avoid such complicity (Michalowski 2013).

Theoretical and Empirical Aspects

This section asks: What are (some of) the risk factors that economic theory and associated empirical work point to, i.e., what makes the risk nonzero ($r > 0$)? The section thereafter asks: How can genocide risk be reduced below 1 ($r < 1$)?

Theoretical Perspectives

Formal economic models of genocide are relatively new in the literature on conflict, peace, and security between and within states. Verwimp (2003), Ferrero (2013), Anderton and Carter (2015), and Anderton and Brauer (in A/B forthcoming) present nonstrategic constrained optimization models to highlight conditions under which a political authority would choose genocide as part of its goal of controlling territory or government (or both). The models reveal conditions in which genocide has a low opportunity cost for the authority, specifically, when genocide enhances the authority’s control in the context of crisis or war, is not too disruptive to economic activities (e.g., trade), is conducive to looting victims’ wealth, and is not likely to generate third-party intervention. Under such conditions, genocide is “cheap,” and a positive amount demanded can exist. Genocide prevention requires that the opportunity cost of genocide is made high through sanctions, credible threats of third-party

intervention to help victims and/or oppose authorities, threats of prosecution, and surveillance of atrocities which can lead to “naming and shaming” of perpetrators. In addition to modeling genocide risk factors, Anderton and Brauer (in A/B forthcoming) use a Lancaster household production model to study the “optimal” choice of genocidal techniques (e.g., mass killing, starvation, forced relocation, etc.) by a regime that has already chosen genocide. Among the results is a “bleakness theorem” in which protection policies along one or just a few dimensions have relatively little overall effect, and sometimes no effect, in protecting victims.

Game theory models of genocide consider strategic interactions between warring groups and/or between an oppressive in-group and an out-group in which intentional destruction of civilian groups is part of war tactics or strategy. For example, Azam and Hoeffler (2002) identify conditions in which warring sides use violence against civilians to strengthen themselves in their strategic interaction. Focusing on the years preceding the 1994 Rwandan genocide, Verwimp (2004) develops a four-player game to model the strategic interactions among the regime, the domestic opposition, a violent rebel group, and the international community. Within the game, eliminating the moderate Hutu opposition and exterminating the Tutsi can be “optimal” strategies. Anderton (2010) draws upon the bargaining theory of war to show how severe threat against an authority group or an incentive to eliminate a persistent rival can lead to genocide as an “optimal” choice. Anderton (2010) and Gangopadhyay (in A/B forthcoming) use evolutionary game theory to model how genocide can become socially contagious (acceptable) among “ordinary people.” Vargas’ (in A/B forthcoming) model of contestation between a government and a rebel group reveals the incentives of each to side to kill the civilians who are supporting the enemy. Within the model, Vargas finds that the strengthening of either side can have ambiguous effects on the total number of civilians killed, thus showing that third-party support for one side or the other can potentially increase civilian killing. Esteban et al.’s (forthcoming; in A/B forthcoming) inter-temporal models of

contestation between a government and a rebel group reveal several important and sometimes counterintuitive results. In particular they find that new discoveries of resources, democratization of the polity, and third-party intervention to defend vulnerable civilians can enhance incentives for mass killing if they materialize under the “wrong” conditions.

The lessons of constrained rational choice and game theory models for thinking about the emergence of laws designed to punish and prevent genocide are critical. Laws that come into being will be evaluated by potential perpetrators of genocide as part of the constraint set being faced. Such agents, if determined to carry out genocide, have an extensive menu of inputs for working around such laws to achieve objectives. Laws to punish or prevent genocide must consider the multiple options available to potential perpetrators and the potential for laws to lead to unintended consequences. This concern is especially significant in the context of strategic interplay between a government, rebel organization, and possible third-party intervener. If not carefully designed, law to prevent and punish genocide can serve to increase incentives for atrocities. (On the design of law, see the next section.)

Perspectives from behavioral economics also help to study genocide (e.g., Anderton and Brauer; Slovic, Västfjäll, and Gregory, both in A/B forthcoming). Especially important is the reference-dependent objective function of one or a few leaders who have become accustomed to control of political, economic, and/or territorial goods. Experiments in behavioral economics often find evidence of *loss aversion*, in which, relative to a reference point such as current hold on power, subjects believe they are worse off from a loss than a similar gain leads them to feel better off. The notion of loss of power as a form of extreme crisis or existential threat in the minds of leaders is palpable in many genocide case studies (Totten and Parsons 2013). Such losses, coupled with the behavioral phenomenon of loss aversion, suggest that leaders could make extreme choices including repressive violence or genocide to avoid loss (Midlarsky 2005, pp. 64–74).

Empirical Perspectives

About 30 published large-sample cross-country empirical studies of genocide or other forms of VAC risk or seriousness exist (see Anderton 2014; Anderton and Carter 2015; and Hoeffler in A/B forthcoming). Most of these focus on genocide risk or severity from the perspective of *countries*, and thus they focus on the problem of genocide from the “macro” or top-down perspective. Another branch of empirical genocide literature focuses on particular countries, regions, or locales in which genocide took hold and spread, thus emphasizing a “micro” or bottom-up perspective. While almost all of the empirical studies of genocide in the literature focus on risk or seriousness based on historical data, studies are emerging with an emphasis on forecasting (e.g., Rost 2013; Butcher and Goldsmith in A/B forthcoming).

The most prominent macro-empirical study of genocide risk in the literature is by Harff (2003), who focused on a sample of states that experienced “state failure” (e.g., civil war, regime collapse) from 1955 to 1997. Of 126 state failures in the sample, 35 led to genocide. Conditioned on state failure, Harff used logit analysis to identify six significant risk factors for genocide onset: magnitude of political upheaval; history of prior genocide; exclusionary ideology held by the ruling elite; autocratic regime; ethnic minority elite; and low trade openness. Failing to make the list of significant risk factors was economic development, which Harff proxied by infant mortality. Another important macro-empirical study of mass atrocity risk is Easterly et al. (2006), who assemble a dataset for many countries for the period 1820–1998. Among their key results, they find that mass atrocity is significantly less likely at high levels of democracy and economic development, in which the latter was proxied by real income per capita.

Regarding potential *economic* risk factors for genocide, subsequent empirical research suggests that Harff’s result for trade is not robust with most studies reporting no significant impact of trade on atrocity risk. In addition, Harff’s result on economic development is open to question because an inverse relationship between real income per capita and atrocity risk or seriousness is one of the few modest empirical regularities in the literature.

Other economic risk factors considered in the empirical literature are income inequality and resource dependence, in which no empirical regularities have yet emerged, and *economic* discrimination, which is only beginning to be considered but in which two studies report a significant positive effect on genocide risk (Rost 2013; Anderton and Carter 2015). In the emerging empirical forecasting literature on genocide, no clear results as yet stand out in regard to the roles of economic variables. In their survey of such literature, Butcher and Goldsmith (in A/B forthcoming) suggest that “while economic factors might have an underlying causal effect on the likelihood of genocidal violence in a society, as predictors in forecasting models they might be overshadowed by political or demographic factors that are more proximate to genocide onset.”

In addition to large-sample, macro-empirical studies of genocide risk or severity are country-specific, micro-empirical studies that focus on particular characteristics of a nation, region, or individuals that led to the onset or spread of atrocity (e.g., Ibañez and Moya in A/B forthcoming; Justino in A/B forthcoming). Country-specific studies typically identify historical, social, and economic conditions particular to the country and tactical and strategic aspects of war that are critical for understanding atrocity. Such finer-grained elements can be glossed over in large-sample cross-section studies. For example, Ibañez and Moya’s (in A/B forthcoming) study of Colombia reveals many dynamic, nuanced, and interrelated aspects of community and household incentives for civilians to flee violence, why some do not flee, and why contesting military forces (government, militias, rebels) tactically and strategically kill and/or force the relocation of civilians. Such complexities are obviously critical in considering laws to reduce risks of future civilian atrocities, but also in prosecuting perpetrators and designing reparations in post-genocide settings.

Economics of International Law

Despite some cases of GMA having been brought to trial in national and international courts or

tribunals – the Armenian trials in Turkey, the Nuremburg trials, the Pinochet case, and more recent tribunals regarding Cambodia, Rwanda, and the Balkan wars of the 1990s – the overall record of reducing the risk of GMA to below certainty ($r < 1$) is only mildly encouraging. There are several reasons for this. First, even assuming away issues of ignorance and apathy, as a matter of *economics*, unilateral action runs into the problem of sufficient scale and multilateral, collective action into issues related to strategic behavior, free-riding, coordination, agency, benefit appropriation, and cost shifting. Even assuming that none of these pose a problem, all options rely on the existence of well-codified and well-functioning regimes of national and international law and their enforcement. Second, as a matter of *law*, then, reducing GMA risk is difficult because (a) state sovereigns are cautious to accede to any international treaty that may later expose them to legal liability in the first place and (b) because state sovereigns generally do not cede jurisdiction over nonstate GMA actors to international bodies (e.g., Nigeria maintains jurisdictional prerogative over Boko Haram; and if a nonstate actor prevails in an internal conflict it may not be brought to justice at all). And third, as a matter of *institutional design*, these topics bring up questions, to echo Oliver Williamson (1999), as to what kind of bad GMAs are in the first place and, correspondingly, what kind of good GMA-related laws are, and how to best supply them.

On the demand (or usage) side, are GMA and GMA-related law private (excludable and rivalrous), public (nonexcludable, nonrivalrous), club (excludable, nonrivalrous), or common-resource pool (nonexcludable, rivalrous) bads or goods, or some changing mixture thereof? And on the supply side, are they best provided by private or public actors, or some changing combination of the two, and what is the technology of their production (e.g., best-shot, weakest-link, aggregate effort, or variants thereof)? What sort of issues in agency, transaction costs, and institutional design arise? While a considerable global public goods (GPG) literature has sprung up in economics (e.g., Kaul and Conceição 2006 and literature cited

therein), application to the design of international law as an instance of GPGs is thin in general and almost entirely absent in regard to law and GMA (see, e.g., a recent symposium of papers in the *European Journal of International Law*, 23(3), 2012).

As regards GMAs, we suggest that indiscriminate chemical weapons gassing may be conceptualized as a *public bad* for the affected population if it is neither feasible to exclude oneself from the gassing nor feasible to seek effective shelter (there can be no rivalry for shelter if there is none to be had). Those who do manage to crowd into a shelter, however, partake in the benefit it offers, the shelter being a common-resource pool good, while those left behind on the street suffer a *common-resource pool bad* (nonexclusionary but rivalrous). In contrast, genocide would be a *club bad* precisely because its architects differentiate and select victims. Finally, examples of a *private bad* suffered in violent conflict include un-orchestrated rape in war or the death of a soldier in the performance of his or her duties (the “expected” bad in war, but not a war crime). Similarly, in regard to the good that GMA-related law may provide, international law of war is intended as a GPG in that all soldiers share in the benefits the law provides and none of them are excluded. In contrast, national law is private to the state whose legislative body passes it: It excludes nationals of other states and reserves benefits to its own nationals. But all international law is effectively a club good, benefitting those who accede, and becomes a pure GPG if and only if *all* states become Party to the treaty in question. In practice, however, it is conceivable that benefits can be withheld so that the benefits law offers become rival to those with the means to access its provisions when needed. Thus, while the Genocide Convention is (not quite) a global public good in principle, the evident practice of “too little, too late” suggests that its enforcement is rivalrous and therefore constitutes a common-resource pool good.

Even this cursory “walk around goods space” (Brauer and van Tuyl 2008, Chap. 8) suggests that the good (or bad) in question can take various forms and that each may change across

geographic space and time. Neither the goods nor the bads are necessarily unitary (of a single form), and to conceive of GMA simply as a global public bad requiring a global public good response may be inadequate. Moreover, as Shaffer (2012) points out, international laws can be rivalrous to each other and their construction is designed, in part, to trade off against multiple national laws (legal pluralism).

In addition, economically efficient (no under- or overprovision) GMA-law in response to GMAs may depend on the summation technology of GMA production. Applying Hirshleifer’s (1983) insight, that some GPGs are best provided as best-shot products (the single-best effort suffices; no need for anyone else to contribute to its provision), weakest-link products (the weakest provider limits the good’s effectiveness), or aggregate effort products (the more is provided by all, the better for all), Shaffer (2012; esp. Table 2, p. 690), argues that best-shot GPGs are best dealt with in global administrative law, weakest-link GPGs by fostering legal pluralism, and that only aggregate effort GPGs may require a global constitutionalist approach. To illustrate, when a single country has effectively become the world’s only superpower to intervene in other states’ (GMA or GMA-alleged) affairs, it may be tempted to overreach or under-reach according to its own cost-benefit calculated perception of its Responsibility to Protect, regardless of the wishes of all other UN members. But superpower intervention or nonintervention solely at its own discretion challenges global legitimacy (the US is often accused in this regard; France, in regional interventions, less so). Such situations, Shaffer (2012) argues, are best dealt with by global administrative law which might hold the “incumbent” of the superpower “office” responsible for its actions. We imagine (since Shaffer does not address GMA), that instead of a Genocide Convention, there might exist an UN-approved automatic trigger obligating the superpower to intervene in cases of GMA, subject to global administrative law. As of this writing, little has been theoretized in this regard.

An additional issue pertains to trans-generational global public goods. Again, this is

insufficiently theorized but probably of great importance in cases of GMAs since each event carries significant generational implications (for a review see, e.g., Ibañez and Moya in A/B forthcoming). For public goods provision, Sandler (1999) speaks for four levels of awareness rules: First, the myopic view considers making a marginal cost (MC) contribution to the provision of a GPG only up to the sum of the marginal benefits (MB) a state estimates for its own current generation, $MC = \Sigma MB$. Second, although still selfish, a forward-looking view is to include one's own offspring generations, i , such that $MC = \Sigma MB_i$. Since the expected benefits are larger, this translates into greater willingness to make a larger MC contribution. Third, a more generous view of the benefits summation includes other states' populations, j , but only for the current generation ($MC = \Sigma MB_j$). The most enlightened view of all – we call this the “Buddha rule” – sums the expected benefits across all generations across all populations, $MC = \Sigma MB_{ij}$. Since such benefit is likely to be large, it justifies correspondingly large outlays.

Finally, design criteria for GPG that would take account of goods (or bads)-space, summation technologies, transboundary, and trans-generational aspects have been discussed in the literature (Sandler 1997) but rarely in regard to GMA-related national and international law (Myerson, in A/B forthcoming, is an exception). It would appear that a fruitful field of inquiry is ready for exploration in this regard.

References

- Anderton CH (2010) Choosing genocide: economic perspectives on the disturbing rationality of race murder. *Def Peace Econ* 21(5–6):459–486
- Anderton CH (2014) A research agenda for the economic study of genocide: signposts from the field of conflict economics. *J Genocide Res* 16(1):113–138
- Anderton CH, Brauer J (eds) (forthcoming) Economic aspects of genocide, mass atrocities, and their prevention. Oxford University Press, New York
- Anderton CH, Carter J (2015) A new look at weak state conditions and genocide risk. *Peace Econ Peace Sci Public Policy* 21(1):1–36
- Azam J, Hoeffler A (2002) Violence against civilians in civil wars: looting or terror? *J Peace Res* 39(4):461–485
- Bornkamm PC (2012) Rwanda's Gacaca courts: between retribution and reparation. Oxford University Press, New York
- Brauer J, van Tuyll H (2008) Castles, battles, and bombs. The University of Chicago Press, Chicago
- Cernic JL (2010) Human rights law and business: corporate responsibility for fundamental human rights. Europa Law Publishing, Groningen
- Curthoys A, Docker J (2008) Defining genocide. In: Stone C (ed) The historiography of genocide. Palgrave Macmillan, New York, pp 9–41
- Dadrian VN (1997) The Turkish Military Tribunal's prosecution of the authors of the Armenian genocide: four major court-martial series. *Holocaust Genocide Stud* 11(1):28–59
- Easterly W, Gatti R, Kurlat S (2006) Development, democracy, and mass killing. *J Econ Growth* 11(2):129–156
- Esteban J, Morelli M, Rohner D (forthcoming) Strategic mass killings. *J Polit Econ*
- Ferrero M (2013) You shall not overkill: substitution between means of group removal. *Peace Econ Peace Sci Public Policy* 19(3):333–342
- Harff B (2003) No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. *Am Polit Sci Rev* 97(1):57–73
- Hillemanns CF (2003) UN norms on the responsibilities of transnational corporations and other business enterprises with regard to human rights. *Ger Law J* 4(10):1065–1080
- Hirshleifer J (1983) From weakest-link to best-shot: the voluntary provision of public goods. *Public Choice* 41(3):371–386
- Kaul I, Conceição P (eds) (2006) The new public finance: responding to global challenges. Oxford University Press, New York
- Kelly MJ (2012) Prosecuting corporations for genocide under international law. *Harvard Law Policy Rev* 6(2):339–367
- Lemkin R (1944) Axis rule in occupied Europe: laws of occupation, analysis of government, proposals for redress. Carnegie Endowment, Washington, DC
- Lunga L (1992) Individual responsibility in international law for serious human rights violations. Springer, New York
- Michalowski S (ed) (2013) Corporate accountability in the context of transitional justice. Routledge, New York
- Midlarsky MI (2005) The killing trap: genocide in the twentieth century. Cambridge University Press, New York
- Pergorier C (2013) Ethnic cleansing: a legal qualification. Routledge, New York
- Rost N (2013) Will it happen again? On the possibility of forecasting the risk of genocide. *J Genocide Res* 15(1):41–67
- Sandler T (1997) Global challenges. Cambridge University Press, New York
- Sandler T (1999) Intergenerational public goods: strategies, efficiency and institutions. In: Kaul I, Grunberg I, Stern MA (eds) Global public goods:

- international cooperation in the 21st century. Oxford University Press, New York, pp 20–50
- Schabas WA (2003) National courts finally begin to prosecute genocide, the ‘crime of crimes’. *J Int Crim Justice* 1(1):39–63
- Schabas WA (2010) The law and genocide. In: Bloxham D, Moses AD (eds) *The Oxford handbook of genocide studies*. Oxford University Press, New York, pp 123–141
- Shaffer G (2012) International law and global public goods in a legal pluralist world. *Eur J Int Law* 23(3):669–693
- Totten S, Parsons WS (eds) (2013) *Centuries of genocide: essays and eyewitness accounts*, 4th edn. Routledge, New York
- United Nations (1948) Convention on the prevention and punishment of the crime of genocide. <https://treaties.un.org>
- Verwimp P (2003) The political economy of coffee, dictatorship, and genocide. *Eur J Polit Econ* 19(2):161–181
- Verwimp P (2004) Games in multiple arenas and institutional design on the eve of the Rwandan genocide. *Peace Econ Peace Sci Public Policy* 10(1):1–47
- Waller J (2007) *Becoming evil*. Oxford University Press, New York
- Williamson O (1999) Public and private bureaucracies: a transaction cost economics perspective. *J Law Econ Organ* 15(1):306–342

Geodesy

Otmar Schuster

Haus der Geoinformation, Mülheim an der Ruhr,
Germany

Abstract

Geodesy, starting as a scientific discipline exploring the size of the earth in the eighteenth century, soon produced the basis of modern land tax systems and enabled the institution of individual property rights in real estate, a backbone of the constitutional order. Geodesy stimulated the development of the modern state by data collection and land management methods. The satellite techniques and the advancement and the expansion of the sensor technology based upon the theoretical mastery of the space in methods and incoming data made the geodesy and its professionals effective in the public management and the economy. Geodesy as an economic sector

producing and managing geo-information is deeply woven in different disciplines of science, public administration, and economy but recognizable by its methods and results.

Definition

Geodesy is a discipline in science, government, land management, and industry. It addresses the shape and the shape near earth, the national government of land, the land management, and real estate property.

Geodesy: Reference Frame for the Earth

Geodesy explains the knowledge about the shape and shape near earth. It comprises specific methods which are useful to analyze tectonic movements but also to manage countries, land, and plots.

It needed more than two millennia to go back to the anthropocentric worldview, but then in the seventeenth and more in the eighteenth century, the technical progress blessed Europe with telescope and a growth of mathematics. The conquest of the continents brought new knowledge about the earth. The question for the right dimension and the shape of the earth moved into the center of the scientific research. Since the angular measurement succeeded to feature a considerable accuracy, the measurement of length was the biggest problem for 200 years (Torge 2009).

The geodetic science succeeded to improve the determination of the dimension of the earth by long triangulation chains along the several meridians on the northern and southern hemisphere. The determination of the International Meter was more or less a by-product. On the basis of this and the expansive surveying methods, the European governments had new tools to manage the countries in a just way. The national economies were able to flower out in an unimagined way.

This deconvolution was a precondition for the preeminence of the European nations in the world. The implemented techniques gave distinction to the scientific specialization under the name of

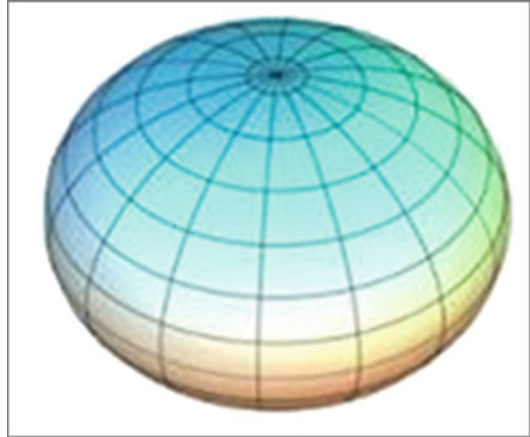
“geodesy.” The application of the methods in small-scale areas of about less than 1 km^2 , in which the spatial discrepancies are negligible to small for a sufficient accuracy (Kahmen 2005), coined the profession of the “geometer,” later referred to as “surveying engineer”.

Whereas the systems of base mapping – big or small – got a national character, because the military security asked for that, geodetic technology always had to be international. Only with international cooperation and in addition with the cooperation with the neighbored geosciences, the possibility evolved for tackling the phenomenon “earth” (Ledersteger 1969) (Fig. 1). And all the more so when the scientific progress bestowed methods on us which could determine the gravity field, the rotation of the earth, the movements of the continents, and the mass movements of the earth’s interior (Fig. 2).

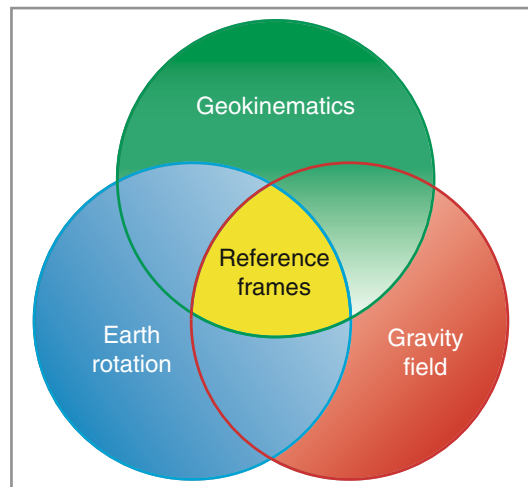
To determine the geodetic frame for national control networks was always one of the most important tasks of geodesy. But international cooperation on this became more and more successful (Schuster 2005). The International Earth Rotation and Reference Systems Service (IERS) maintains the International Terrestrial Reference System (ITRS), which describes the procedures for creating reference frames suitable for use with measurements on or near the earth’s surface. This International Terrestrial Reference Frame consists of the high-precision coordinates and velocities of about 400 globally distributed stations (Fig. 3). These three-dimensional Cartesian coordinates and velocities define the positions and the plate tectonic movement, the ITRS.

The ITRF points are part of the European ETRF and the regional networks like the German control net, called DREF, or even networks of lower order under national control; they form together the international frame of high precision. It makes us measure the dislocations of the 19 tectonic plates (Fig. 4), (Sella et al. 2011).

From the beginning astronomical geodesy was necessary to locate the datum of every triangulation net on the ellipsoid. Soon satellite orbits were calculated, and the theoretical earth tides were compared with those measured by gravimeters and horizontal pendulums. Insofar geodesy was



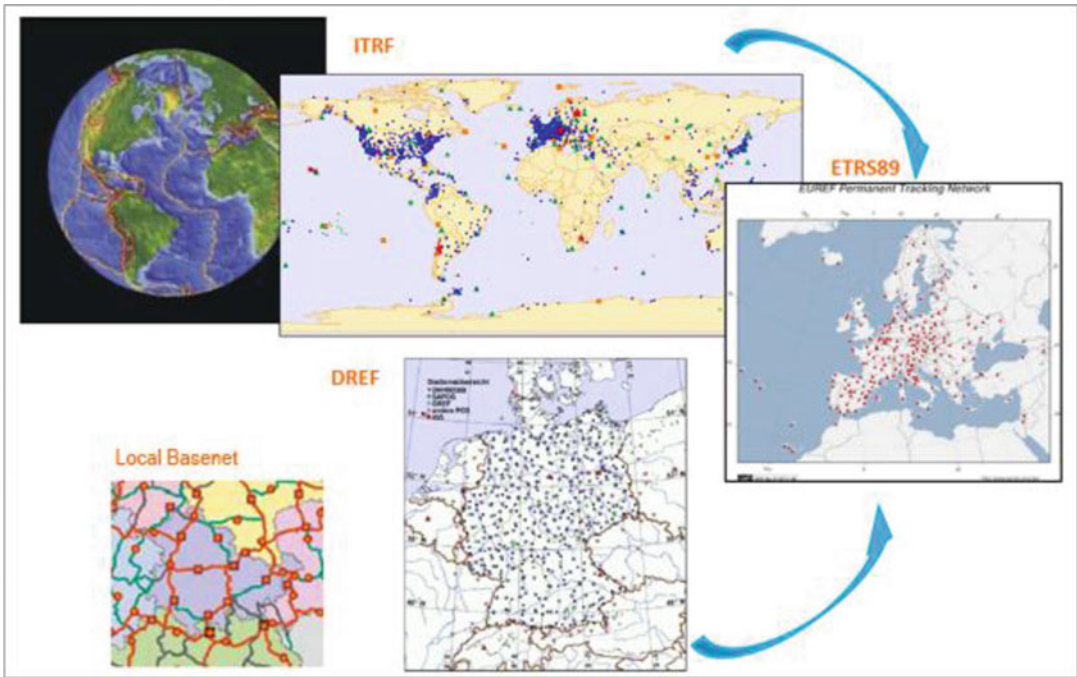
Geodesy, Fig. 1 The earth as a mathematical abstraction



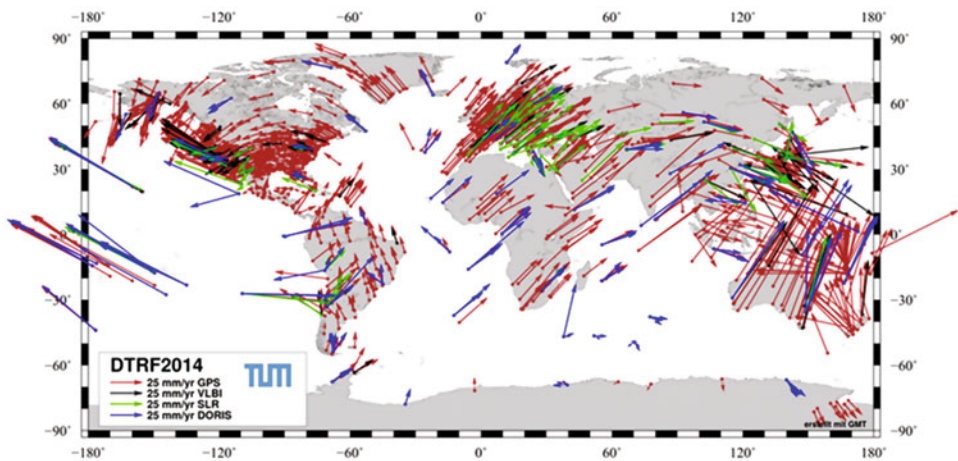
Geodesy, Fig. 2 Study object earth

prepared for the huge next step, the use of the earth near satellites. Up to that time, the meteorological conditions only were disruptive elements for the optical determination of lengths and heights, but then the meteorological surroundings more and more came into view with the higher accuracy of all results.

Geodetic measures are deeply influenced by the shift of masses (Fig. 5) between tectonic plates and the respective interdependence with solid earth and oceans, atmosphere, hydrosphere, cryosphere, biosphere, and asthenosphere (universe). The scientific picture of the earth changed from the spherical shape of the spheroid to the ellipsoid, further to the



Geodesy, Fig. 3 The shape of the earth – under control (Courtesy to Prof. Harald Schuh, GFZ Potsdam)



Geodesy, Fig. 4 Tectonic plate dislocations (ITRS Combination Centre at DGFI-TUM 2014) <https://www.dgfi.tum.de/international-services/itrs-combination-centre/>

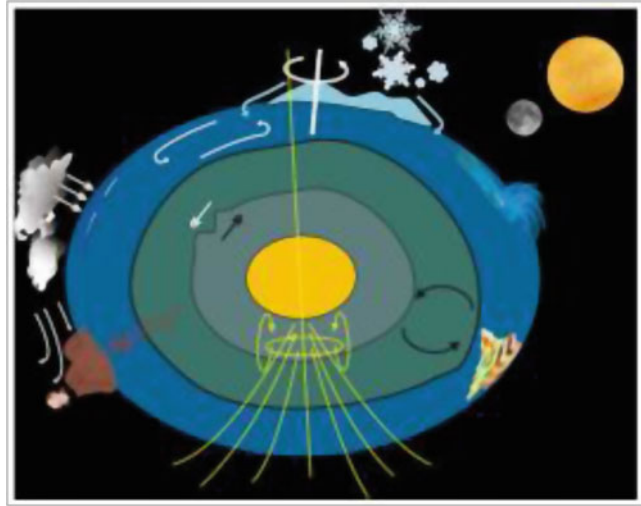
geoid. In addition the geodetic science taught us that our earth is in an ongoing process of change.

Today, the highly accurate geodetic frames ITRF and ETRF and the national frames like DREF are combined, even merged. Everybody

can locate himself or move in that system. His measurements – mounted in the net – are now worldwide identifiable for about better than 1 [cm] (Seitz et al. 2016). Unlike earlier times the surveying engineer can detect, if the sunspot

Geodesy, Fig. 5

Opposing forcing affecting the earth (Vienna University of Technology)



activity is strong and the magnetic field is jammed. This he has to take care of when performing his measurements.

Individual Property Rights

During the colonization of earth, people became increasingly aware that fruitful land is a scarce resource. It is worth to settle there and to assert the ground. On those places, where people met for trade and change, regulations were accepted for the peaceful way of dealing with each other. Not mentioning the tremendous achievements of the Romans in surveying of their empire, the geodesy steps into the limelight of history, when the harvest tax systematically turned to ground tax. At the end of the eighteenth century, the science provided the technological possibility to expanse scalable maps over the country and to document securely the property borders (Fig. 6). Since that time a fair, area-based tax regime became possible. Tax cadastres became fashionable for modern states. Soon large-scale soil fertility evaluation systems for farmland were set up on these cadastres.

The systematization of the taxation (Fig. 7) and the spread over the countries took technically and politically nearly a century in Europe. But this installation provided the nations with safe funds (Fig. 8). The other side of the coin was that the citizen's rights in their property were

strengthened. Geodetic technique was weaved in the deploying civil system of justice. The civil code, the land register ordinance, and the official regulations for the property cadastre came to their heyday at the beginning of the twentieth century. The taxable value is a legal definition for private plots, operational premises, and farming or forestry property. It should be equivalent to the market value, but all over Europe, there is a great variety of tax procedures and level of taxation. In countries like the Netherlands and Sweden, the method of mass evaluation for the ground taxation is discussed. A lot of member states of the European Union are seeking for an easy and secure way for this taxation. The qualitative classification of the nations concerning just ground tax, secure property, and successful mortgaging is an unmistakable sign for the development of land management.

The development did not come to halt (beside the communistic part of the world). Legal and management principles changed with the development of the economy and its regulations. The improved cartography and copying techniques lead to further cadastres showing other attributes of parcels and their buildings (i.e., cadastre on public land charges or pollution suspicion cadastre, geothermal cadastre, ao). They influence the rights in land substantially.

The geodetic professionals working in private or public businesses managed the property

Geodesy, Fig. 6 Modern cadastral map, 1:1000



cadastre and taxation in form of cadastre measurements and cadastre amendments and the evaluation of the ground (Schuster 1981).

In the nineteenth and twentieth century, the rural and urban land readjustment and the management of settlement became an important economic factor driven by geodetic professionals. With the measures of public authorities, the farmers succeeded to raise their crop yields substantially; the drift of the population to the cities was stopped. They succeeded to overcome the big streams of people into the industrial areas.

At least after the Second World War, they managed to integrate 12 million people into the West German population. Geodetic reallocation and evaluation measures were a significant part of the success.

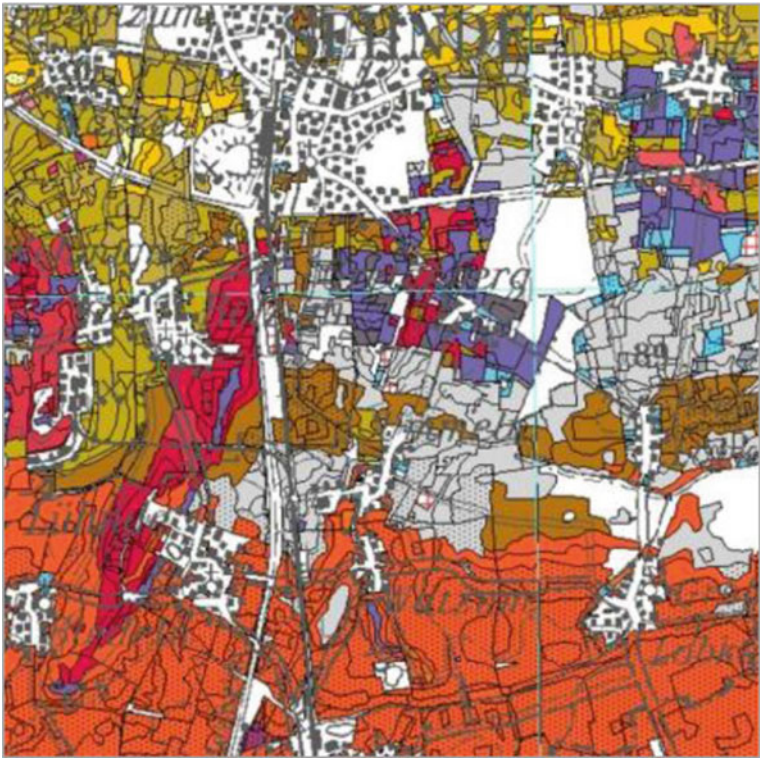
Real Estate Economy

Geodesy is a basis for all real estate economy.

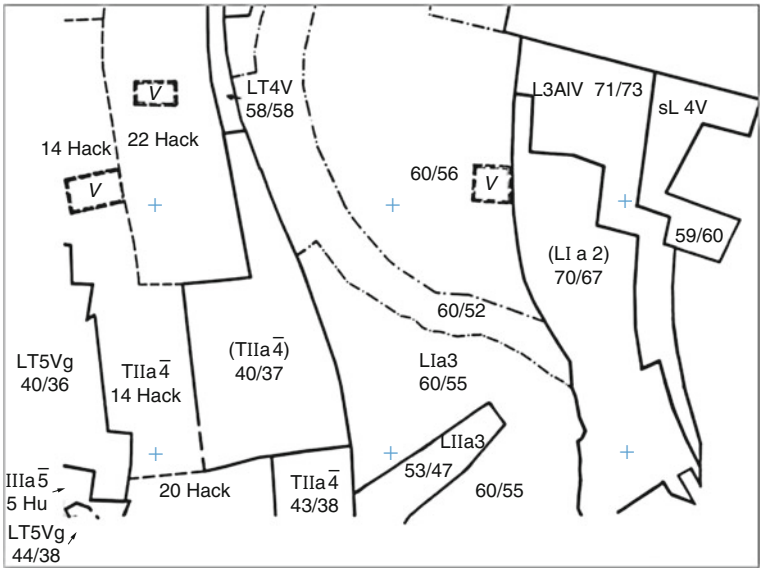
The upcoming time of geo-information opens new economic prospects and effective tools for the protection of our environment. The safe governmental documentation of properties and the loan capabilities (Kerl and Schuster 2003) provided the growing economy by capital generation. The results are reflected in a growing built-up space for industry and residence. Diverse forms of right or property were developed on the basis of the property register and cadastre. They developed their own business cycles. Geodetic services were part of this process of development, but managed mostly only the geometric relocation of the property borders. Slowly they supplied the growing demand of their clients in consulting concerning the more and more complex subject.

In the socialist countries, this development failed even if the geodetic basics and also services for mapping and the building regulations were still or even newly existing. Without the individual property in land, the economic shaping forces were lame. The real estate economy was no area

Geodesy, Fig. 7 Soil fertility valuation, 1:5000



Geodesy, Fig. 8 Soil fertility map, 1:2000



of added value. After the political transformation in 1989, this form of economy on the basis of land quickly induced a takeoff. The authorities quickly transformed their character and got reformed. But

with these events, the societal questions around land are not yet answered. The more developed a nation and its civil class is, the more sophisticated and dense is the wood of regulations based on

land. This business life pushes the gross national product of a country immensely. But in that sphere, there is easily coming up the danger of non-genuine growth: economic blows and the growth of bureaucracy can jeopardize the national economy. Both emerge because of inappropriate regulations and misguided public work.

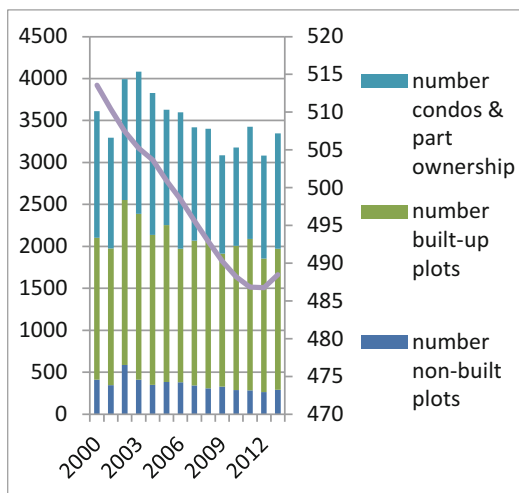
Geodesy is part of this economic process by its land management component, which expanded its methods and performance strongly in the second half of the twentieth century. The German federal building law (BauGB) and the federal land use order (BauNVO) together with the rapidly growing public law considerably influenced the methods of property evaluation and readjustment, which were used in the public and private part of the business life. Other developed countries have similar regulations.

The development of standard ground values and the installation of committees of expert valuers became an internationally acclaimed model for subduing the blows or deflation elements.

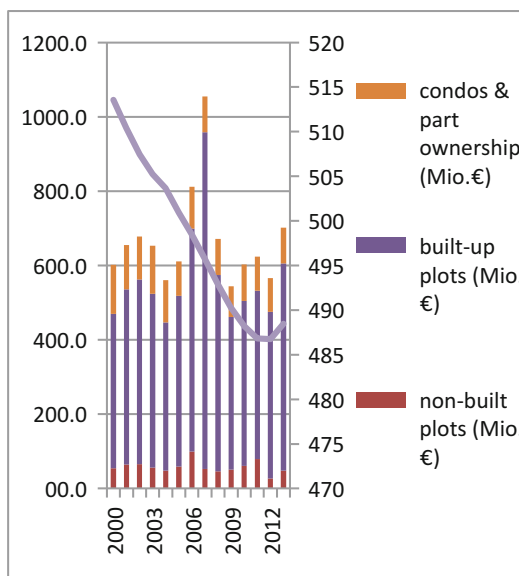
Since the Millennium not only the migration from country to town was noticeable but also the diminishing of an aging European population becomes apparent. The economic output of the house construction industry became much smaller and this happened as well in that part of the geodetic economy. On the other hand, the investments for the maintenance and renewal of the buildings pushed ahead by certain tax reductions and ecological regulations grew strongly. The industry sold their houses for workers by privatization, so the number of transactions swelled (Fig. 9). The people as economic actors wish to safeguard their property, and by that the economic transactions remained on a high level.

Foreign capital searched from 2005 to 2008 for investment opportunities. This pushed the price level especially in the European agglomeration zones. Geodetic services are involved in that part of the economy with evaluation of real estate, property consulting, and evaluation for mortgages.

These different types of propulsions let the transaction traffic remain on a high level, even if the development of new development in unbuilt areas was strongly reduced.



Geodesy, Fig. 9 Numbers of transactions – city of Duisburg



Geodesy, Fig. 10 Cash flows of transactions – city of Duisburg

From both pictures (Figs. 9 and 10), you can deduce the economic importance of the real estate economy. In a city of 500.000 inhabitants happen to be concluded not more than about 4000 contracts per year, which means not more than about 8.000 contract parties. It is a small part of the town's people to exchange a fortune of 500 Mio € up to 1 billion € per year. A number of 250–500

unbuilt plots with a transaction value of 50 to 80 Mio € changed the ownership. These plots are predominantly the basis of a development of the city further on.

The number of contracts for developed, built plots is about 1.500; their value balances around 500 Mio €.

1200 to 1500 contracts are concluded concerning residence and part ownership with a value of about 60 Mio €.

Behind the curtain of this development, over the years there are a lot of different influences changing from year to year. The last ascent of values is reckoned to be based on the fear of people for their financial assets, not coming from a bigger need on the renter's side.

From "concrete gold" people expect to gain the best long-term maintenance of value.

Market economy is a contract economy. The exchange of property only means the top of the economic contracts in the real estate economy. The broad base of contracts is the renting contracts.

An exchange by contract of about one billion € releases work for the service institutions register and cadastral office, notary, broker, and public appointed surveyor of about 50 Mio €. The

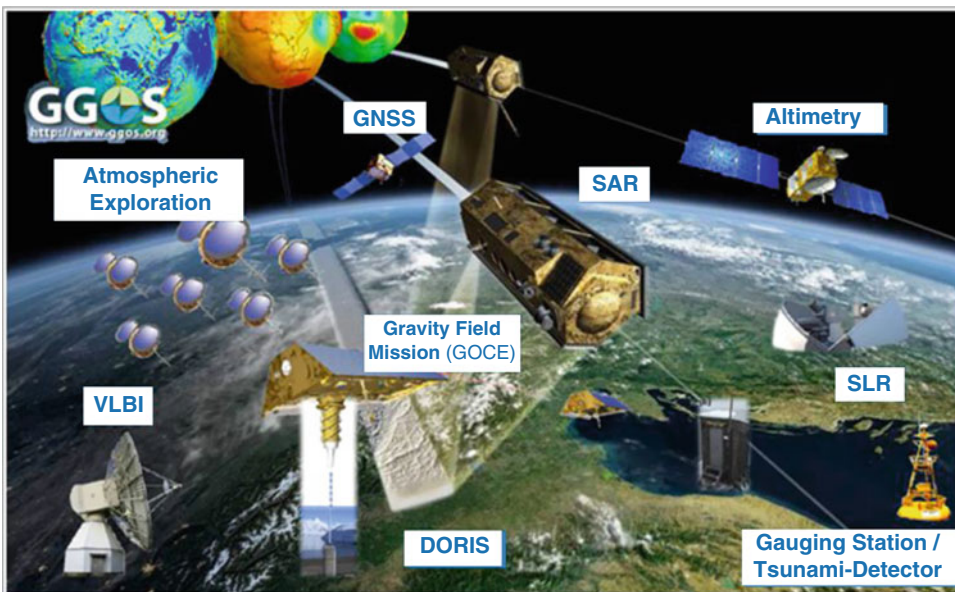
transaction tax for the community as well costs the parties 5% which means about 50 Mio €.

Geodetic Technology

Geodetic technology is more productive than ever in history (Schuster 1985). The number of sensors and items grew enormously: satellites for GNSS, altimetry, gravity field missions, side-looking airborne radar (SLAR) or synthetic aperture radar (SAR), and Doppler Orbitography and Radio-positioning Integrated by Satellite (DORIS) (Fig. 11). Not to forget the equipment for atmospheric exploration, the Very Long Baseline Interferometry (VLBI) and the tide gauges or even complex systems like tsunami detectors are all being used in the low and medium earth orbits.

With remote sensing a new geodetic technique came up for evaluation of the satellite photography, laser and radar surveys.

New sensor technology presents a variety of technical devices with regional impact like surveying planes or helicopters, steered by IMUs (inertial measurement units) with a variety of aircraft cameras and lasers. Most of the sensors built have a local impact: GNSS Rovers for dm



Geodesy, Fig. 11 Geodetic technology in the earth near space (Courtesy to Prof. Harald Schuh, GFZ Potsdam)



Geodesy, Fig. 12 Geodetic sensors with local perimeter

or cm – accuracy, combined positioning and sensor systems for vehicle fleets or personal tracking as well as tachymeters as total stations and laser scanners of different type (Fig. 12).

In the metrological scale, we find many robotics and high-precision measuring instruments (laser tracker *ao*), necessary for 3D measurements, and in-house laser systems in a local reference system.

The geodetic methods are directed upon maps and scientific results, but also upon data pools consisting of maps and alphanumeric information like property cadastre and surveying and engineering geodesy. They are used for the control of the building process. There is a big variety of procedures, products, and results. As far as they are based on measurements, they are to be distinguished by an implicit or explicit stochastic. For long the maps developed to graphic interactive systems (GIS). The orthophoto as an automatically generated aerial scalable photo is a worldwide proven and successful product.

Unexpectedly the unmanned aerial vehicle (UAV) or shortly drone started its triumphal march into the geodetic technology: equipped with cameras of most different quality or even small lasers, they catch up the objects in the space near to the surface (Haala and Schwieger 2017). The point clouds produced lead to new economic useful results, i.e., for monitoring, quantity or mass calculation, urban models of

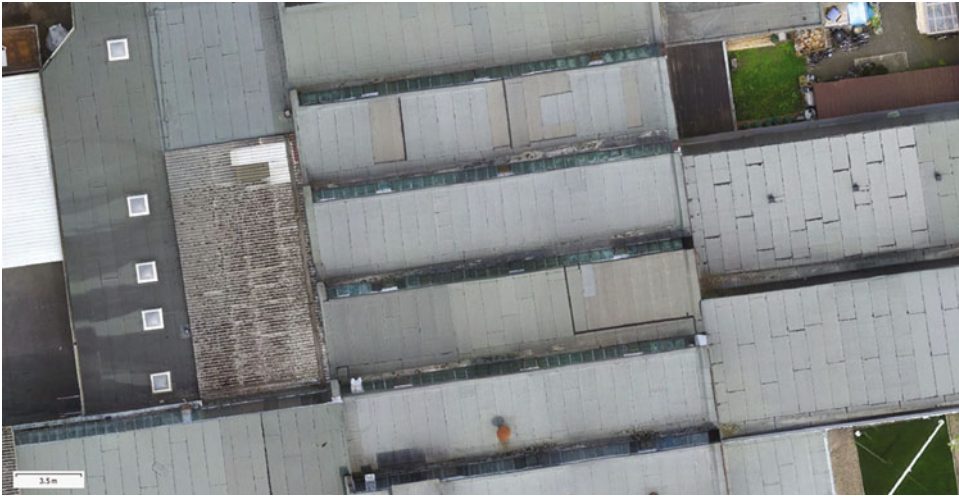
higher level of detail (LOD), roof landscapes as orthophotos (Fig. 13), a geodetic base for the building information modeling (BIM), or disaster control. Powerful geodetic software modulates the transformation of the point clouds.

Geodetic laser scan or even laser slam methods (without any fixed standpoint) capture new fields of application spirited up by the revolutionary building information models, which need and are able to process Big Data results during the construction process (Fig. 14).

Parallel to mechanical engineering, the quick comparison between the actual status of a construction in progress can be realized with an authentic point cloud by laser slam, which is called “trusted living point cloud” (Fig. 15).

The actual expansion of the sensor technology and the velocity of the data capturing enable the geodesy to do its bit for the “smart construction site.” The cross-linking of sensors, devices, machines, and finally the workforces is beginning (Fig. 16). The result is a seamless communication between sensors and server over the Internet. The sensors deliver the current necessary information for workforce and fleet management and support the accounting of the site’s activities. Complex algorithms and modern mathematical filter methods have to be composed.

Inertial measurement units (IMUs) in connection with great effort in terms of data filtering will play a great role in the future geodetic technology,



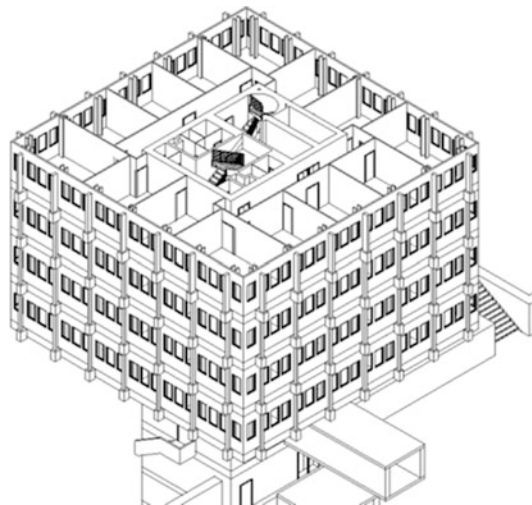
Geodesy, Fig. 13 Roof landscape as an orthophoto (acc. $<\pm 1$ mm) captured by UAV

because they are able to fix the coordinates of a moving object down to [mm] accuracy together with GNSS as a frame. So the coordinates of any moving object can be created coincidentally. One scientific challenge of the hour is the high-precision time adjustment of the different moving sensors.

The New Age of Geo-Information

Geodesy means basically the technique, which connects freely selected points with an imaginary line. This technique got social and economic importance with the property cadastre. The border points were surveyed, and the imaginary line between them was the property boundary. It is the borderline between the spheres of activity between neighbors. These borderlines are the most important legal and economic reference frame (Schuster 1997). Most of the authoritative actions and private measures relate to that system. Coordinate systems – as described above – are from that view auxiliary systems, which ease to comprehend the economic life.

Two hundred years ago, the installation of such a reference system “property borders” needed a huge economic effort of the authority. Today, the sum of new technologies enables the geodetic community to give a coordinate to all points, on



Geodesy, Fig. 14 Building captured by laser slam – base for BIM

which the society wants to imprint its will. Since that procedure happens quickly, the movement of objects or even that of men can be documented in such coordinate systems.

Vice versa every coordinate created bears a certain information, the geo-information (3. Geofortschrittsbericht der Bundesregierung 2012).

Seen from an economic viewpoint, the elements of the unnumbered nature and that of human privacy become finite-measurable. This means that water, air, wood, etc. a hundred years

meta-levels are needed, which means the graphic design in gliding scales and gliding adapted degree of abstraction.

Meta-data as well mean numerical description, the naming of maps and the description of the accuracy of the coordinates or related structures.

The public authorities gain a lot of advantages for their work on future topics like climate, energy, mobility, sustainability, and demographics. The European authorities started early with INSPIRE and Copernicus (former GMES) (Alessandro and Claude 1999). On national and regional level, the European structure got sub-structures (i.e., GDI-DE) partitioned to responsibilities.

The INSPIRE directive gives the legal frame for that Europe-wide geo-data infrastructure.

Worldwide suppliers of geo-information like Google or Microsoft today seem to be unavoidable, and they disperse in the economy with new business models (Barwinski and Schuster 1988; Schuster 1997).

Discussing the naming of this new field of work, international groups decided to change the

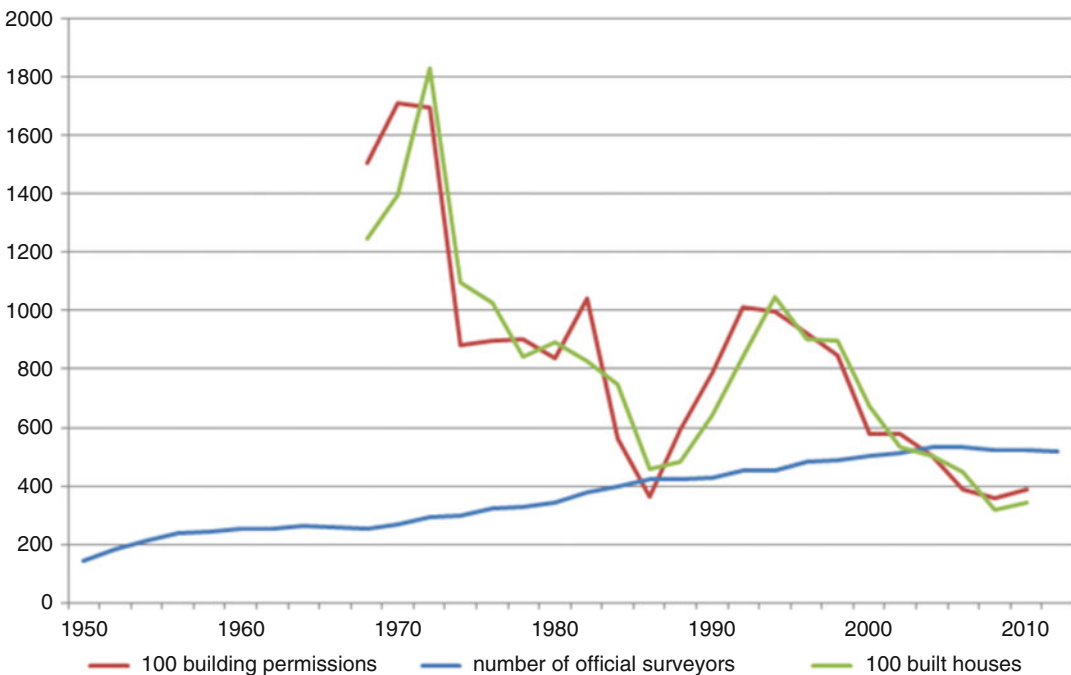
term to “geomatic” to show that the geodesy has expanded its field.

Geodesy as an Economic Sector

It is comparatively simple to value the costs of an economic sector on the European continent, because the professionals remain for a very high percentage of their professional life in their profession once chosen.

The “Market Report” presented by the Comité de Liaison des Géomètres (CLGE) and Geometer Europas (GE) (Schuster et al. 2003) comes to the conclusion that the volume of costs, deduced from the salary payments to geodetic professionals in the year 2000, is about 24 billion €. This sum comprises those who are busy with the maintenance of the public property system and public maps as public servants, as well as those who are active in the private business life in form of surveying, evaluation, and software (Schuster 2003).

Given that the predominant part of demand in geodesy is due to the building applications, Fig. 17 shows the dramatic changes in demand



Geodesy, Fig. 17 Dramatic changes of supply and demand

since the 1950s and the steadily growing number of supply (number of private offices) in a state like North Rhine-Westphalia with about 18 Mio inhabitants.

The private service structure of the supply answered with the reduction of personnel and an furthermore improved its offering with better quality of comprehensive services on the basis of the extending public regulations; the public structure passed several waves of reorganization but benefitted from the extending relevant legislation.

So, the demand for geodetic supply is relevant to the legal situation of land as a whole. The differences between countries can be huge.

In a developed country like the Netherlands, the geodetic community speaks about “crowdsourcing,” which means that the owners themselves are able to identify the property boundaries between their parcels on the basis of a well-explained cadastral information of the public authority.

The Internet now has brought up a new situation, where some hundred experts, i.e., from Google and Bing, can theoretically reach the whole mankind with their products. The funds for this investment and its maintenance come mostly from sources outside of the professional sphere (i.e., marketing). The underlying business model couples different sectors, which is not unusual in the economy (Schuster 2004).

Whereas the old business model of the geodesy was relatively closed and strongly regulated and could be broken down into

- Public mapping
- Maintenance of cadastre
- Cadastral surveys
- Sequent control of building geometry,

the business field expanded its supply strongly by the vast new field of sensors, the practically unlimited possibility for the geo-referencing of any object, and the use of the Internet. The enlargement of the supply is financed in the middle-class economy by the old business, big public investments, and the hope, which was induced by the technological progress. The

demand is lagging behind, which means generally low prices for the services.

The situation in various European countries is different, and it is scarcely straightforward because of the degree of regulation, not to mention the clearness in detail. Beside such difficulties one can say that the old geodetic business model was pillared by:

- Ground tax
- Transaction traffic
- Settlements
- Military needs
- Sequent control of building geometry

Nowadays is upcoming a growing income for the use of most different geo-information. One quickly spreading example is the area cadastre. With this GIS the communities are able to split the fees for percolation water and sewage. New services are created. With the cadastre of trees, of greens, and of public easements, they help to manage the real estates and have an organizational added value.

Many people look forward to a growing activity in geodesy. But the answer to the question is not yet clear, until new “biting” business models and broader economic trends will show up. Undoubtedly, they will deform the old economic basis.

References

- Allessandro A, Claude L (1999) (Europäische Kommission) (eds) Spatial reference systems for Europe. A joint initiative of MGRIN and the Space Application Institute
- Barwinski K, Schuster O (1988) Landinformationssystem und Freie Berufe. BDVI Schriftenreihe des Bundes der Öffentlich bestellten Vermessungsingenieure, vol 1
- Bundesministerium des Innern: 3. Geo-Fortschrittsbericht der Bundesregierung Stand Okt (2012)
- Haala N, Schwieger V (2017) UAV – Anforderungen und Möglichkeiten. DVW Schriftenreihe des Deutschen Vereins für Vermessungswesens, vol 86
- Kahmen H (2005) Angewandte Geodäsie: Vermessungskunde, 20th edn. Gruyter, Walter de GmbH, Berlin. ISBN 3-11-018464-8
- Kerl V, Schuster O (2003) Necessary outline conditions for property financing through covered bonds and

- mortgage banks. GEOhaus GbR, Berlin/Mülheim an der Ruhr
- Ledersteger K (1969) *Astronomische und physikalische Geodäsie, Handbuch der Vermessungskunde*, Bd 5, 10. Aufl. Metzler, Stuttgart
- Schuster O (1981) *Der Berufstand des Vermessungsingenieurs und seine Vorläufer*, Schriftenreihe des Vermessungstechnischen Museums. Konrad Wittwer Verlag, Dortmund, S 83–101
- Schuster O (1985) Satellitennutzende Vermessung – Beginn einer neuen Ära. BDVI-FORUM Zeitschrift des Bundes der Öffentlich bestellten Vermessungsingenieure. p 185
- Schuster O (1997) *Freie Berufe und Kultur – Schlaglichter. BDVI – Schriftenreihe*, vol 6
- Schuster O (2003) *Geodäsie in der Wirtschaft „FuB“-Flächenmanagement und Bodenordnung* Heft 2
- Schuster O (2004) *Geodesy in the German economy*. Eur J Law Econ 17(2):191–208
- Schuster O (2005) *Spatial data and globalisation*, Atlantic Think Tank VI. Massachusetts Institute of Technology, Cambridge, MA
- Schuster O, Ouranos E, Busch M, Höflinger E(†) (2003) Market report – report about the Market of Surveying in Europe, Editors: Comité de Liaison des Géomètres (CLGE) und GEOMETER EUROPAS (GE)
- Seitz M, Bloßfeld M, Angermann D, Schmid R, Gerstl M, Seitz F (2016) The new DGF-TUM realization of the ITRS: DTRF2014 (data). Deutsches Geodätisches Forschungsinstitut, Munich. <https://doi.org/10.1594/PANGAEA.864046>
- Sella GF, Dixon TH, Mao A (2011) REVEL: a model for recent plate velocities from space geodesy. J Phys Res 107(B4):ETG 11-1–ETG11-30
- Torge W (2009) *Geschichte der Geodäsie in Deutschland*, 2. Aufl. De Gruyter, Berlin. ISBN 978-3-11-020719-4

Geographical Indications

Marco Ricolfi

Department of Legal Studies, Department of Law,
Turin University, Turin, Italy

Definition

The issues at hand. Geographical indications (GIs) are signs (or symbols) used in connection with the sale of goods which convey an association, direct or indirect, with a place or location, which may be relevant for certain features of the good in question.

Rules concerning GIs may be found both at the municipal (or domestic) level and in international agreements. The international framework for dealing with GIs is to be found in the 1883 Paris Convention, in the 1891 Madrid Arrangement and in the 1958 Lisbon agreement. After the adoption by the European Union of the *sui generis* regime for GIs and designations of origin in 1992, a multilateral equilibrium was found in the 1994 TRIPs agreement.

When the same geographical symbol is used simultaneously by many businesses, a number of issues arise which may be dealt with in several ways. The different alternatives are discussed hereafter along with the reasons which may dictate the choice among them. They all rely, however, on a common assumption: that GIs should, at least in principle, not be appropriated by one single business at the expense of all its rivals. Therefore, we first deal with this assumption, which is shared by most legal systems.

Geographical Symbols and Individual Trademarks

Most legal systems prevent a single business to register as a trademark a geographical symbol (or to otherwise obtain exclusive legal rights over it). The reason for this prohibition is straightforward and has two prongs. Let us assume that the business trying to appropriate the geographical symbol is in fact established in the geographical area to which the symbol alludes to. If this is the case, and the geographical origin is responsible for features which are relevant for consumers' choices, then the grant of a trademark registration would generate an unwarranted competitive advantage for the benefit of one business to the exclusion and detriment of all its competitors. The matter gets even worse if we consider the second prong: if the business holding the trademark is not even located in the geographical area to which the symbol alludes and, therefore, the goods originating from it do not possess the corresponding features, then the public is deceived.

Usually, these difficulties are tackled by four related rules:

- There is a bar on the registration of signs which convey an association between features of the good and a given place (in US law: primarily geographically descriptive symbols).
- Should in fact the goods bearing the trademark in violation of the previous rule also come from an area different from the one designated by the symbol, this would also amount to prohibited deceptive behavior.
- However, the two rules above do not prevent the registration of a geographical symbol which, on its face, has no association whatsoever with features of the goods on which it is affixed; Montblanc fountain pens is a good example of this situation.
- Also, in some – but not in all – legal systems registrability may be obtained through the acquisition of a “secondary meaning,” which, in this case, indicates that the relevant public no longer associates the indication with a geographical place but with a business origin which has obtained recognition in the marketplace.

One Geographical Symbol, Many Users: The Legal Options

All the difficulties associated with the use of a geographical symbol, discussed in the previous paragraph, fade away when the geographical symbol is used simultaneously by several businesses, all located in the area which is relevant for the characteristics of the goods. This is an occurrence which is more frequent in connection with agricultural goods, such as wines, cheese, fruit, and the like, but may occasionally be relevant also for manufactured goods which embody handicraft traditions, such as pottery and fabrics (Cuccia and Santagata 2004). This does not mean that the simultaneous use of the same geographical indication by different businesses is unproblematic. To the contrary, this situation may generate a number of difficulties, which, while different from the ones arising from the use by a single business, still require the adoption of appropriate legal devices to avoid the emergence of recurring shortcomings, which include consumer deception

and unwarranted competitive advantage, again, but are not limited to them.

Three are the main options available to deal with the issues arising by the use of geographical symbols by multiple businesses, which may consist in the setting up of property rights, in the resort to liability rules, and in the adoption of a *sui generis* regime. While the choice between these different options ultimately rests on international considerations, rather than on the comparative assessment of the economic costs and benefits of the different approaches, for analytical reasons we will first describe the different options available at the domestic level in an economic analysis of law (EAL) perspective, reserving for a later stage consideration of the international factors which determine the choice.

The Domestic Level: Property Rights

It is possible to set up property rights enabling multiple businesses to use the same geographical symbol in a coordinated and nondeceptive manner in two different ways: collective trademarks and certification marks. Both are known in common law systems as well as in civil law countries. However, the latter tend to favor collective trademarks, the former certification marks (Belson 2002; on the status of collective and certification trademarks in the TRIPs Agreement see Pires de Carvalho 2006; on the corresponding notions Spada 1996). In either case, the geographical sign is registered as a trademark; as a result, any use by an unauthorized third party of a sign which is identical or similar for the goods indicated in the certificate is automatically considered as an infringement.

What distinguishes collective and certification trademarks is that in the former, the holder of the trademark is an entity (a corporation, a consortium, an association) and the users are members of the same; in the latter, the holder of the trademark is an independent third party, with no links with the users, which is legally prevented from engaging in the business for which the trademark is registered. This third party is contractually

bound to certify that users comply with the applicable standards, as agreed in an appropriate document (the “specification”).

In short, collective trademarks are membership-based; certification trademarks are contract-based. All the differences in the rules concerning collective and certification trademarks flow directly from this distinguishing feature.

Collective Trademarks

Each user obtains authority from the entity which holds the trademark to use it in conformity with the “specification,” which is adopted and submitted with the application for registration. In principle, only businesses which are members of the entity holding the registration are allowed to use the collective trademark. On the one hand, this rule is consistent with the more communitarian underpinnings of collective trademarks: only businesses which operate in the geographical area which is relevant for the features of the goods or otherwise share a characteristic which is signaled by the collective trademark are allowed to resort to it. On the other hand, making membership of the trademark-holding entity a precondition for use entails the possibility of abuse. More specifically, the risk is recurring that businesses complying with all the requirements are denied the opportunity to become members of the collective trademark holding entity and therefore to obtain a license to use it. Legal systems tackle this risk in several ways (Sarti 2011; Spada 1996). In some systems, an “open door” policy is mandated: applications for registration of collective trademarks are granted only on condition that the by-laws of the applicant entity provide that each business meeting the requirements is allowed to become a member and to obtain a license to use. In other systems, even non-members are entitled by law to use the geographical indication corresponding to the collective trademark, provided the sign is used as a descriptor of geographical origin rather than as trademark. Legal systems which do not adopt these approaches may resort to antitrust laws to make sure that the denial of admission to membership of a business meeting all requirements is prevented.

Certification Trademarks

The entity holding the certification trademark may be public or private and may come in any legal form. What is of essence is that its relationship with users is contract-based: compliance with rules set in the specification entitles to use of the trademark; no discrimination is allowed in the certification. This is the rational basis of the rule whereby the certifying entity may not be in the business sector for which it gives the certification.

The Domestic Level: Tort

Businesses located in a certain area which is relevant for some of the features of the goods they supply have not only the option to obtain a property right in the geographical indication, but, as an alternative, they may also resort to tort law. This is particularly so in legal systems which give protection against unfair competition. However, the scope of protection afforded under liability rules is more limited than the one based on property rules. A predicate required before a business operating in the relevant area may claim protection against use by a third party to associate with the location is that the geographical sign has already reached a certain degree of recognition or reputation. It should be noted in this connection that, when protection is sought at the international level, recognition in the country of destination (import), not of origin (export), is controlling. The businesses operating in the relevant area must also prove that the third party resorting to the corresponding geographical symbol is taking advantage of a reputation it has not earned (“free ride”), to the detriment of lawful users (“injury”) and, additionally, that the public is likely to be deceived. This requirement entails another limitation in protection; disclaimers, such as corrective additions on the label (e.g., Roquefort, produced in Ontario, Canada) or qualifying language (e.g., Roquefort-type), may avert liability.

A Comparative Assessment

There is not much literature comparing the costs and benefits of protection of GIs under property

rules as opposed to liability rules. This is rather unsurprising, as the protection of GIs is not prevailing in the Anglo-Saxon world (Faulhaber 2005), and European scholars are wary of treading in waters not tested by American literature (Ricolfi 2009a). However, in a welfare maximization perspective, we may quite safely assume that the property and liability rules are effective in a roughly equivalent way in pursuing the goal of lowering consumer search costs. As to the other function of trademarks, to provide an incentive to invest in quality (or, rather, a combination of quality and price which meets consumers' desires), it is submitted that, on the one hand, property rules tend to be superior. Resort to a standard set by the relevant community in a formal and transparent way, the specification, is likely to avert resort to free-riding better than a reference to more nebulous "best practices" to be established during litigation. However, the costs of setting up a collective trademark registration system are likely to be substantial (Van Caenegem 2004). On the other hand, assuming that adjustment over time is a desirable feature for GIs, where standards may change over time, it is arguable that tort-based systems are superior to property-based systems, as they tend to be more flexible. This is so because they rely on the understanding associated with a specific geographical symbol at the time a controversy arises rather than on the notion codified once and for all in a specification.

The Sui Generis Regime of Protection of Geographical Indications

In the quest for protection of geographical symbols, there is a third alternative which tops the property and liability rules-based systems just sketched out: the *sui generis* regime adopted back in 1992 by the EU, which is remarkable in itself and as an (alleged) blueprint for the multi-lateral provisions adopted by Artt. 22 ff. TRIPs. The current version of the EU regime is to be found in reg. no. 1151/2012. All in all, this regime is more similar to property than to liability rules-based systems, even though it deviates from them in that there is no person or entity which may be

described as the owner or holder of a sign. The collective monopoly, bestowed on all businesses which can show their location in the relevant area and compliance with the requirements set by the specification, is the result of a public law proceeding initiated at the local level and completed by the grant of title by the EU Commission. What is provided by this set of rules is a *sui generis* regime, characterized by low access requirements and strong levels of protection. This feature is not very usual in intellectual property law, where, by and large, the level of protection is commensurate to the access requirement. The deviation shown by the EU *sui generis* regime is accounted for by a clear political choice intended to give a high level of protection to agricultural communities and to their local traditions in growing and producing foodstuffs.

Three components of the regime stand out.

1. As far as *access requirements* are concerned, the regime, Art. 5 of reg. no. 1151/12, applies to so-called protected designation of origin (PDO) which, as in the prior international framework, refers to signs identifying products whose quality or characteristics are essentially or exclusively due to a particular geographical environment with its inherent natural and human factors (the so-called *milieu*). The protection extends, however, also to "protected geographical indications" (PGI), the requirement for which are remarkably lower as they depend not on objective criteria (the *milieu*) but on the subjective factor of the reputation and even admit to some relaxation as to the localization of the production steps;
2. While in theory "generic terms shall not be registered" as PDOs or PGIs, Art. 6 of reg. no. 1151/12, in practice even descriptive terms, like "Feta," for the characteristic cheese of Greek origin, have been found to be registrable even though they had been in common usage outside Greece for long time (the ruling by the European Court of Justice was adopted in 2005). Under Art. 13 PDOs and PGIs, once registered, are no longer subject to genericization.
3. Registration as PDO or PGI confers an exclusivity over the sign entailing very strong

protection. While the protection for ordinary trademarks is granted only against the risk of confusion, here the grant extends to usage which falls short of such a risk (e.g., mere “evocation”) and does not take into account the presence of disclaimers which may prevent consumers’ confusion (Art. 13 of reg. no. 1151/12).

Originally the *sui generis* regime was thought of as a tool to facilitate the replacement of member States rules by a single EU protection title. However, Art. 11 of the current regulation opens up access to this EU title also to non-EU entities, in furtherance of the TRIPs-mandated prohibition of discrimination against non-EU nationals.

The comparative assessment carried out above may be extended to the *sui generis* regime. Overall, the regime is closer to property than to tort: access requires registration; once registration is obtained, no proof of deception of the public or of injury to lawful users is required. What is remarkable in the regime is the political nature of the process which leads to the registration. For example: one single Greek island may register a dozen GIs for olives and oil, not so much because there is a business rationale for this proliferation but because local politicians and communities choose to engage in an exercise where most of the costs are borne by the general taxpayer. This is hardly surprising: a preferential treatment for agricultural communities is at the basis of the French-German alliance which has been underpinning the design of the EU from its beginnings.

The International Level

One might assume that lawmakers are in a position to choose among the options described in above in order to maximize the welfare of their constituency and that they accordingly proceed to select the regime which minimizes costs and maximizes the benefits. In real life, lawmakers have another, overriding variable to consider: the position in international trade of the political entity they are legislating for. In this connection, two

considerations should be factored in. *First*, the choice has a very strong mercantilist component: usually countries tend to support exports and discourage imports; and this applies also to goods bearing geographical indications. In this regard, producers’ interests may prevail over consumers’, as predicted by collective action theory. *Second*, goods originating from a given country (country A) may derive competitive advantage (or disadvantage) not so much from rules adopted in country A as from rules applicable in the export markets countries B, C, ... N. As a result, it is submitted that, while EAL is a powerful analytical tool to examine the competitive consequences of the different options, it is public choice theory which ultimately accounts for the choice among the various options.

This is the backdrop against which the international framework of norms concerning GIs has to be assessed. In doing so, one should never forget that here we are dealing with power structures rather than with economic optimality.

The 1891 Madrid Agreement: The Tort-Based Approach The 1883 Paris Convention, which is considered – along with the 1886 Berne Convention for copyright – as one of the two pillars of international intellectual property protection, only mentioned geographical indications without adopting specific rules concerning them. As a follow-up to it, the 1891 Madrid Agreement was adopted, which visualized protection of GIs as an issue of unfair competition. Access requirements for protection are low. Also simple (and indirect) indications of geographical origin are protected; therefore no specific link between the features and the place of origin is required. Correspondingly, protection is weak: it is confined to cases of actual fraud; even more to the point, the relevant recognition and reputation of the goods designated by a geographical indication has to be established on the sales market, not in the country of origin. In this perspective, the Madrid rules are acceptable to countries which tend to import, rather than to export, goods bearing GIs.

The 1958 Lisbon Agreement for the Protection of Appellations of Origin The countries which

tend to export, rather than to import, goods bearing a geographical indication pushed in the opposite direction half a century later, finally adopting the Lisbon Agreement in 1958. This convention presents features which are the obverse of the ones observed in the Madrid Agreement. The access requirements for protection are high: only signs identifying products, the quality or characteristics of which are exclusively due to the *milieu*, are eligible and their protection is contingent on registration in appropriate national registers. The scope of protection is correspondingly high: it extends to usurpation and imitation; infringement is not ruled out by disclaimers or rectifying language (e.g., “like” or “type”); and, most importantly, it has an extraterritorial dimension, in that, once the appellation of origin is registered in one State party to the Agreement, protection extends to all the other States automatically, without the need to establish either recognition or reputation on the sales market.

The problem with the Lisbon Agreement is that it only includes like-minded countries. The 27 States which are parties to it are notable for the importance and quality of their agricultural production, which makes them export-, rather than import-oriented. For their own part, import-oriented jurisdictions are loath to join an agreement which would increase the level of protection of foreign productions to the detriment of local businesses. In 2015, the Geneva Act has tried to take care of these difficulties, lowering the access requirements to the system. Whether this move will in fact enable the participant countries to reach the required critical mass remains to be seen.

A Half-Way House: The GI Provisions in TRIPS

EU officials like the idea that the TRIPs provisions concerning GIs (Artt. 22–24) follow the blueprint of the European *sui generis regime* (Gervais 1998). This is to a very large extent a delusion. While a regime resembling the EU approach is envisaged for wines and spirits (Art. 23), all other GIs are dealt with as a tort or unfair competition issue: no protection is available without deception of the public (Art. 22). As a result, the EU has placed itself into a situation where it is under the obligation to extend the benefits of its

generous *sui generis* regime to all businesses which are resident in a country which is a TRIPs member (Artt. 3–4) while its own businesses cannot claim a protection going beyond tort law for European GIs abroad, which was already available to them under the 1891 Madrid Agreement all the time. The EU appears all the more misguided as it tries to lure developing countries to follow it into a bargain in which it is a clear loser (Ricolfi 2009b).

Future Directions: Conflicting Agendas in GI Protection

There is no doubt that the EU and the Anglo-Saxon world are pushing in opposite directions as far as GIs are concerned (O’Connor 2004). The latter favors GI rules which avoid competitive disadvantage deriving from location for their powerful agro-businesses. The former extends into the twenty-first century the traditional pro-agricultural communities attitude inherited from the French-German arrangements at the basis of the European architecture. What we do not know as yet is how this divergence will fare when brought into the much wider compass of global competition. Initially, the rival views have played out in the bilateral and regional free-trade agreements (Rangel-Ortiz 2014). Until the US elections in 2016, it would have seemed that the American approach was taking the upper hand also in international trade negotiations, such as the Trans Pacific Partnership (TPP) and the Trans-Atlantic Trade and Investment Partnership (TTIP). However, the situation is very much in a state of flux now, as a result of the repudiation of these treaties by the Trump presidency.

We may wonder what role was – and is being – played by Economic Analysis of Law in the process. Perhaps unsurprisingly, EAL-based tools may contribute to understanding what are the costs and benefits of the different alternatives at hand, but as the choices are being taken at the super-national, rather than domestic level, they seem to be suggested much more by power dynamics rather than by the pursuit of welfare optimization.

References

- Belson J (2002) Certification marks, special report. Sweet & Maxwell, London
- Cuccia T, Santagata W (2004) Collective property rights for economic development: the case of the ceramic cultural district in Caltagirone, Sicily. In: Colombatto E (ed) *The Elgar companion to the economics of property rights*. Edward Elgar, Cheltenham, 473 ff
- Faulhaber LV (2005) Cured meat and Idaho potatoes: a comparative analysis of European and American protection and enforcement of geographic indications of foodstuffs. *Columbia J Eur Law* 11:623 ff
- Gervais D (1998) *The TRIPs Agreement. Drafting history and analysis*. Sweet & Maxwell, London, p 119
- O'Connor B (2004) The legal protection of geographical indications. *Intellect Prop Q* 1:35
- Pires de Carvalho N (2006) *The TRIPs regime of trademarks and designs*. Wolters-Kluwer, The Hague, 214 ff. and 369 ff
- Rangel-Ortiz H (2014) Patent and trademark rights in commercial agreements entered by the United States with Latin American nations in the first decade of the twenty-first century: Divide et vinces. In: Ghidini G, Peritz JR, Ricolfi M (eds) *TRIPs and developing countries. Towards a new IP world order*. Edward Elgar, Cheltenham, 72 ff
- Ricolfi M (2009a) Geographical symbols in intellectual property law: the policy options. In: *Festschrift für Ulrich Loewenheim zum 75. Geburtstag, Schutz von Kreativität und Wettbewerb*. Verlag Beck, Munich, 231 ff
- Ricolfi M (2009b) Is the European GI policy in need of rethinking? *IIC* 40:123 f
- Sarti D (2011) Il marchio collettivo. In: Ubertazzi LC (ed) *La proprietà intellettuale*. In: *Trattato di diritto privato dell'Unione europea* directed by G. Ajani and G.A. Benacchio, vol XI. Giappichelli, Torino, 131 ff
- Spada P (1996) Il marchio collettivo "privato" tra distinzione e certificazione. In: *Scritti in onore di G. Minervini, vol II, Impresa, concorrenza, procedure concorsuali*. Morano, Napoli, 475 ff
- Van Caenegem W (2004) Registered GIs: intellectual property, agricultural policy and international trade. *Eur Intellect Prop Rev* 26:170 ff
- Goebel B, Groeschl M (2014) The long road to resolving conflicts between trademarks and geographical indications. *TMR* 104:829 ff
- Knaak R (2015) Geographical indications and their relations with trade marks in EU law. *ICC* 46:843 ff
- Libertini M (2010) L'informazione sull'origine dei prodotti nella disciplina comunitaria. *Riv Dirit Ind* 1:289 ff
- Mantrov V (2014) *EU law on indications of geographical origin: theory and practice*. Springer, Cham
- Ricolfi M (2015) *Trattato dei marchi. Diritto europeo e nazionale, vol 2*. Giappichelli, Torino, 1754 ff
- Sarti D (2009) Segni e garanzie di qualità. In: Ubertazzi B, Muñoz Espada E (eds) *Le indicazioni di qualità degli alimenti*. Giuffrè, Milano, 113 ss
- Sironi GE (2014) Marchi collettivi. In: Scuffi M, Franzosi M (eds) *Diritto industriale italiano, vol 1, Diritto sostanziale*. CEDAM, Padova, 312 ff
- Tosato A (2013) The protection of traditional foods in the EU: traditional specialties guaranteed. *Eur Law J* 19:545 ff

German Law System

Sonja Elisabeth Haberl
University of Ferrara, Ferrara, Italy

Abstract

The following entry provides an overview of some elected aspects of the German law system. From different points of view, the German system has been deeply influenced by the ordoliberal ideas developed within the Freiburg School in the early 1930s of the twentieth century. One of the core ordoliberal concepts that have to be discussed within the constitutional framework is that of a social market economy. Social market economy became the interpretive framework for the economic and social order of Western Germany in the aftermath of World War II, and today, it represents not only a key concept at national level but also within the European Union. No less important is the role of the so-called Private Law Society, another key concept of ordoliberal thinking. Its main elements are clearly reflected in the *Bürgerliches Gesetzbuch* (BGB) of 1900 which is based on the idea of the citizen as a *homo oeconomicus*. Notwithstanding its traditional approach – libertarian, unsocial, and

individualistic – the BGB, a child of the abstract conceptualism of the Pandectist school, has been able to survive till today. This is the merit of judge-made law and, in particular, of the theory of the indirect horizontal effect of fundamental rights in relations governed by private law. In recent times, the BGB has even assumed a highly visible role as a possible model within the harmonization of European contract law. The entry finishes with a description of the German system of legal education, a state-oriented and judge-centered bureaucratic model which is still embedded in the model of a “uniform jurist,” the so-called *Einheitsjurist*.

The Constitutional Framework

The memory of the collapse of the ill-functioning, weak, and helpless Weimar democracy which facilitated the slide into a totalitarian dictatorship has profoundly influenced the framers elaborating the post-1945 constitutional order of a West German state (Kielmansegg 1990). The document which should offer appropriate guarantees for the development of a solid democracy, able to defend itself against its enemies, was elaborated by a Parliamentary Council, composed of 65 members elected by the state parliaments, and approved by the *Länder* rather than by popular referendum. It entered into force on 23 May 1949 and was named *Grundgesetz* (Basic Law) and not *Verfassung* (Constitution) in order to underline its provisional character while waiting for unification. Despite this, the Basic Law outgrew its temporary character by means of the accession of the five East German *Länder* to the FRG in October 1990, and with only a few provisions revised and others inserted, it became an all-German constitution within the new Berlin Republic.

The Discussions About the Basic Law’s Economic Order: Social State, Social Market Economy, and Ordoliberalism

Unlike the Weimar Constitution, the Basic Law does not contain any explicit reference to social

rights. The only indirect references are included in Art. 20, which defines Germany a “democratic and *social* federal state,” and in Art. 28 which requires the constitutional order in the *Länder* to conform “to the principles of a democratic and *social* state governed by the rule of law” (the so-called principle of homogeneity). The reasons for the mere affirmation of the social state principle are closely connected to the highly conflicting positions between the constitution-shaping political forces with regard to the economic order which the Grundgesetz should refer to. While the Social Democratic Party tried to promote the concept of economic democracy, based on a form of planned economy, workers’ participation in management, and the nationalization of important economic interests (Bommarius 2011), the Christian Democratic Party argued in favor of a system of neoliberal democracy, in which state intervention should be aimed at creating the necessary conditions for the functionality of the market mechanism. Since the various political forces involved in the constitution-making process were not able to reach a consensus on this issue, they set aside the prescription of a specific economic model in the Grundgesetz. As the Federal Constitutional Court later stated, the Basic Law is “neutral” as regards the economic order (4 BVerfGE 7 (1954)). Instead, they contented themselves with reference to the principle of the social state, thus giving rise to the paradoxical fact that the Basic Law ascribes a fundamental rank to “the social,” yet without defining it more precisely by means of specific social rights.

After a short time, a new key word was coined through which the task of the German State to perform as a “social state” should be fulfilled (Zacher 1987): the ideas underlying the concept of a social market economy are theoretically and ideologically deeply rooted in the ordoliberal ideas developed within the Freiburg School during the Nazi dictatorship. In harsh opposition to the Weimar party and intervention state (Stolleis 2002), its proponents – among others the economists Eucken, Rüstow, Röpke, and the lawyer Böhm – were united in the idea of militating against social and political pluralism and arguing in favor of a “strong state,” called to efficiently

manage the economy by building and enforcing a legal regime representing an *ordo* intrinsic to economic life (Joerges and Roedl 2004). The ordoliberals thus propagated the so-called third way as the proper alternative between laissez-faire liberalism and collectivist forms of political economy, calling for the establishment of a system of undistorted competition in order to enhance the functioning of the economic order.

It was Alfred Müller-Armack who substantiated the ordoliberal ideas in the enticing slogan social market economy and transformed it into policy together with Ludwig Erhard under the chancellorship of Konrad Adenauer. Social market economy refers to a strategy based on voluntary market transactions with undistorted competition and private law mechanisms as its fundamental elements, to be complemented only by a certain kind of state intervention: contrary to what one could think, in fact, the meaning of the attribute “social” does not refer to a policy of social justice associated with a welfare state. Müller-Armack, by arguing for the “total mobilisation of all forces” (Müller Armack 1933), to be achieved by enabling individuals as self-responsible and self-determined entrepreneurs (Bonefeld 2012; Somma 2013), primarily referred to the efficiency of a market economy, i.e., to the fact that the latter generates economic growth and wealth, thus directly and automatically bringing about social achievements (Joerges and Roedl 2004). In addition, he required “a system of social and societal measures” as social peacekeeping tools which however had to be *marktkonform*, i.e., consistent with the competitive order (Müller-Armack 1966).

Social market economy has become a core concept of the foundational period of the FRG, whose success story in the postwar period is frequently associated with it. With its insertion, in 1990, into the treaty establishing a Monetary, Economic, and Social Union between the FRG and the GDR, the term achieved the rank of a legal norm for the first time. It is also included in the Treaty of the European Union (Art. 3) which sets among its objectives that of a “highly competitive social market economy.”

The System of Fundamental Rights and the Bundesverfassungsgericht as the Guardian of the Grundgesetz

The outstanding importance of the fundamental rights included in the *Grundgesetz* is emphasized by their location – they are all placed at the head of the document (Arts. 1–19), with the guarantee of human dignity at its core. In the words of the Federal Constitutional Court (Bundesverfassungsgericht), the paramount importance of Article 1 par. 1 – which declares human dignity as inviolable and provides for its obligatory respect and protection by all state authorities – can be explained only by the historical experience and spiritual-moral confrontation with the previous system of National Socialism. In order to avoid another abyss, the *Grundgesetz* has constructed a “value-bound order in which the individual person and his/her dignity is placed at the center of all its rules and regulations” (39 BVerfGE 1 (1975)). Art. 1 par. 1 thus expresses the highest value of the *Grundgesetz*, which not only has an effect on the following fundamental rights but also informs the substance and spirit of the entire document (Häberle 1987).

The whole system and structure for the protection of fundamental rights aim to correct the mistakes of the past. While the fundamental rights included in the Weimar Constitution were only declaratory and thus not judicially enforceable, the Basic Law’s fundamental rights are transformed in subjective rights by Art. 1 par. 3 which states that they “shall bind the legislature, the executive and the judiciary as directly applicable law.” In addition and also directed at remedying the deficiencies of the Weimar Constitution, the *Grundgesetz* contains precise rules and limits which have to be respected in the context of the restriction of a fundamental right, whose “essential core” (*Wesensgehalt*) may in no case be infringed (Art. 19).

On the other hand, the *Grundgesetz* provides for the forfeiture of certain fundamental rights, should they be abused “in order to combat the free democratic basic order” (Art. 18). Besides the possibility of a party ban (Art. 21) and the

prohibition of associations (Art. 9), the forfeiture of basic rights is considered to be one of the core elements regarding the conception of democracy in the Basic Law, qualified as a “militant democracy” (*streitbare oder wehrhafte Demokratie*), which “expects its citizens to defend the free democratic basic order and does not accept the misuse of fundamental rights aiming to undermine it” (28 BVerfGE 36 (1970)). In order to further strengthen this concept, the Parliamentary Council opted for the textual anchoring of a so-called eternity clause (*Ewigkeitsklausel*) according to which certain core elements of the Constitution are unamendable (Art. 79). The federal state principle, human dignity, and the basic principles of state order mentioned in Art. 20 belong to the elements which represent the identity of the Basic Law and are thus immune to any constitutional revision.

The supreme guardian of the Constitution is the Bundesverfassungsgericht (BVerfG). It was established in 1951 and is vested with extraordinary powers. It has not only the competence for constitutional disputes between federal organs and between the federation and the *Länder* but also the right to control the constitutionality of laws and to deal with individual constitutional complaints which can be initiated by a person who alleges to be negatively affected in his/her fundamental rights by an action of the public authority. The proceedings for controlling the constitutionality of laws can either be initiated by an ordinary court (concrete judicial review) or upon application of the Federal Government, a *Land* government, or one fourth of the members of the *Bundestag* (abstract judicial review). In addition to these competences, the BVerfG also rules “in the other instances provided for in this Basic Law” (Art. 93). It is the competent organ for declaring the forfeiture of basic rights (Art. 18), for cases of impeachment of federal judges (Art. 98) and of the federal president (Art. 61). Its function as a guardian of the Constitution perhaps finds its most evident expression in Art. 21 according to which the BVerfG shall rule on the question of the unconstitutionality of political parties.

The Federal System and the Division of Powers

The federation established after reunification in 1990 consists of 16 *Länder*, and each of them has its own constitution upon which the state institutions are based. Although both the *Bund* and the *Länder* are vested with the three branches of public power, these are not separate and distinct from each other. On the contrary, Germany’s system of cooperative federalism gives a significant example of what Fritz Scharpf termed *Politikverflechtung* (Scharpf 2009), referring to the complex interrelationship, interdependencies, and overlappings between the competences of the federation and the states. The Federalism Reform I of 2006 aims to disentangle these interdependencies, on the one hand by more precisely delimiting the allocation of legislative competences between the federal and the state level and on the other by reducing the risk of blocking situations which can potentially arise during the legislative process because of a divergence of majorities between *Bundestag* (Federal Diet) and *Bundesrat* (Federal Council). The latter consists of at least three and at the most six members of every *Land* government and has the ability to block federal legislation whenever it affects the interests of the *Länder*. In that case, the Grundgesetz requires the *Bundesrat*’s consent, and the latter has an absolute veto which cannot be overridden by an equivalent vote of the *Bundestag* (Art. 77). In order to decrease the *Bundesrat*’s possibility to block federal lawmaking, the federalism reform has significantly reduced the percentage of laws requiring its consent. Besides that, it profoundly altered the rules concerning the division of legislative competences. While the previous so-called “framework” legislation was abolished, both the matters of the federation’s exclusive (Arts. 71, 73) and concurrent (Art. 72) legislation were expanded.

The strong interconnection between federal and state level is at its greatest within the administration of justice. The courts of first and second instance (*Amtsgerichte*, *Landgerichte*, and *Oberlandesgerichte*) are state courts, whereas only the highest courts are established at the

federal level. Hence, there is a hierarchical division between the federal level and the *Länder*, and the federal courts do only control questions of legality. In addition to decentralization, there is also a high degree of specialization within the German court system. There are five different judicial branches – the ordinary, administrative, financial, labor, and social jurisdiction – with respective federal courts (*Bundesgerichte*) at their head which are scattered in different cities throughout Germany.

The German Civil Code and Private Law Society

The significance and importance of private law is emphasized by the ordoliberal concept of a Private Law Society (Böhm 1966), according to which the French Revolution marks the moment in which public order ceases to be the core governance mechanism in a hierarchically organized society, and consensus and private transactions become the main instruments for governing most parts of social life. However, since economic freedom, according to ordoliberal thinking, does only exist through order – it is an “ordered freedom” (Bonefeld 2012) – the new society, characterized by the equality of its members, party autonomy, freedom of contract, and freedom of competition, needs a strong state, called to assure the orderly conduct of self-interested entrepreneurs.

The German Civil Code of 1896 (*Bürgerliches Gesetzbuch* – BGB), which came into force on 1 January 1900 after 20 years of work, clearly reflects the core elements of a Private Law Society. The principle of freedom of contract, though it is nowhere expressly proclaimed, dominates the law of obligations and the idea that contracting parties are formally free and equal implicates that contracts thus formed must be adhered to in all cases. The typical citizen for the BGB is the *homo oeconomicus*, a person “who can be expected to have business experience and sound judgement, capable of succeeding in a bourgeois society with freedom of contract, freedom of establishment and freedom of competition” (Zweigert and Kötz 1994). Whenever the content of a contract was the

result of a bargaining process, the former was considered to be fair – it was the contractual mechanism itself that was said to guarantee the correctness of its outcome (Schmidt-Rimpler 1941). Put succinctly, the bargaining process between the parties was seen as “the epitome of fairness” (Markesinis et al. 2006), and the parties were considered to be the best guarantors of their respective rights. Thus, the idea that individuals are free to engage in private transactions as they see fit only finds limits in the case of the violation of statutory prohibitions (§ 134), *bonos mores*, or in those cases in which one party has exploited the plight, inexperience, or lack of judgment of the other (§ 138), whereas the draftsmen expressly rejected the doctrine of *laesio enormis*, holding that the idea of a *iustum pretium* did not conform to the idea of freedom of contract.

Severe criticism was expressed by those who argued against the dominance of laissez-faire concepts in favor of a more solidaristic approach, calling for the limitation of freedom of contract, which was accused of favoring only the propertied classes while suppressing the socially weaker ones (Anton Menger). Nonetheless, at the end, only a few “drops of socialist oil” (Otto von Gierke) were added to the soulless individualism of the Code. Party autonomy should not be considerably restricted, neither by means of statutory nor judge-made law. In fact, the role that the few open-textured general clauses inserted into the BGB should once have played was a role the drafting fathers had not counted on.

The BGB is divided into five books, each of which is devoted to a different subject, with a complicated system of cross-referencing between them. This applies especially to the first book, the so-called General Part (*Allgemeiner Teil*) which contains certain basic institutions – such as the provisions concerning natural and legal persons (including the concepts of consumer (§ 13) and business (§ 14) which were introduced in 2000 by a statute implementing various EEC consumer directives), juristic entities and foundations, and general rules about legal acts (*Rechtsgeschäfte*) – that have an effect on all the other four books (*Law of Obligations*, *Law of Property*, *Family Law*, and *Law of Succession*).

In particular, the provisions regarding legal acts – an artificial concept which represents the leading topic of the *Allgemeiner Teil* – provide evidence of the BGB's abstract conceptual calculus. The fundamental component of a *Rechtsgeschäft* is the declaration of intent, and this is why it does not only include “normal” types of contract, such as sale or lease and the so-called real contracts necessary to create or transfer real rights over another's property, but also the contract of family law such as marriage, as well as unilateral acts such as that of making a will (Zweigert and Kötz 1994).

In its concepts and its language, the BGB is the child of the abstract conceptualism of the Pandectist school. It does not deal with particular cases in a concrete manner, but adopts an extremely abstract, logical, and accurate legal jargon which is admired for its rigor of thought and precision, but criticized because of its lack of comprehensibility, especially for the ordinary citizen.

The BGB During the Twentieth Century

The question immediately arising in this context is how a Code which was said to be more “the cadence of the Nineteenth than the upbeat to the Twentieth Century” (Radbruch), thus doing no more than prudently summing up the past rather than boldly anticipating the future (Zitelmann 1900), could have survived till this day, transiting the German Empire, the Weimar Republic, Nazi Germany, two world wars, and German reunification without being subjected to fundamental revision; a Code which was described and criticized as conceptualist, libertarian, unsocial and individualistic, and unaware of the changing economic and social climate in the late nineteenth century.

The few drops of socialist oil inserted into the BGB indeed soon proved to be insufficient, in particular in the field of labor law. Already during the Weimar Republic, it became clear that the existing regulation was inadequate and incomplete and that legislative reforms were necessary in order to better protect dependent workers by means of

provisions concerning security of employment, worker participation, and minimum rates of pay (Zweigert and Kötz 1994). Thus, labor law for the most part developed outside the Code. But also in other areas of law, such as the law of housing, the need to take more account of the social needs of the time and to better protect the economically weaker parties in society led to a considerable increase of legislation outside the BGB; only since the 1960s parts of it have been gradually incorporated into the Code.

In certain cases, even the Code itself proved to be a useful instrument for coping with the changing social circumstances. In the aftermath of World War I, in a period of great instability, a sheet anchor to approach the arising great social and economic problems was represented by the general clauses inserted in the BGB. In particular § 242, which states that the performance of a contract has to be according to the requirements of good faith (*Treu und Glauben*), was envisaged by the judiciary as a possible loophole in order to afford the revaluation of debts, which proved necessary in consequence of the staggering hyperinflation of the early 1920s. Since currency laws explicitly established the nominalistic principle *Mark ist gleich Mark* (a paper Mark repaid during the inflation is equal to a gold Mark invested before inflation) and government rejected revaluation, it was left to the judiciary to find a solution for the constantly growing problem of the inflation effects on monetary relations. Although the legislator had never intended the principle of good faith to be used in this manner, in 1923, the *Reichsgericht* used it to request a revision of monetary obligations owed under contracts (RGZ 107, 78).

The significance and importance of the general clauses as a means which allows the courts to keep the Code up to date have also emerged in other areas. To cite just one example, in the context of standard terms of contract whose proliferation was the aftereffect of the Industrial Revolution at the end of the nineteenth century, the general clauses (especially § 242) turned out to be the decisive instrument in order to efficiently control the increasing number of cases dealing with standard terms. Many of the rules developed by the judiciary

in the course of time eventually turned into statutory form within the Act on General Conditions of Business of 1977 (*Gesetz zur Regelung der Allgemeinen Geschäftsbedingungen – AGBG*) whose provisions were partly inserted into the second book of the BGB in the course of the 2002 Act Modernizing the Law of Obligations (for details, see below).

However, the downside of open-textured principles with no clearly defined content became evident during the Nazi regime when the arbitrary misuse in particular of the principles of good faith and good morals, reinterpreted by the judges in order to direct the law in a way which served their nationalist ideology, led to the reassessment of a whole legal order by means of interpretation, resulting in the phenomenon of an *entartetes Recht* (Rüthers 1989), a “degenerated law,” entirely based on the National Socialist *Weltanschauung*. Though the Nazis had initially warned against the general clauses, there was now a real proliferation of undetermined terms in a number of new measures and special laws in order to serve as “entry points” for their ideology. Even more, the *Führer* principle and the *völkische Rechtsidee* (the popular notion of the law) according to which law was no longer perceived in terms of the rights of the individual but as the rights of the people as determined by the state became supra-positive and preexisting principles able to limit the existing written law. Doctrinal rationality and legal certainty got completely lost (Stolleis 1998). The plans of the Academy of German Law to create a *Volksgesetzbuch*, a “people’s Code,” in order to replace the BGB, however, did not progress beyond the draft.

The BGB initially continued to be in force in both parts of Germany also after World War II. In the GDR, it was gradually replaced only some years after the construction of the Wall. On 1 April 1966, the Family Code led to the abrogation of the fourth book of the BGB; 10 years later, the BGB was entirely replaced by the *Zivilgesetzbuch* (ZGB) which entered into force on 1 January 1976 (and was replaced by the BGB in 1990). In the FRG, the entering into force of the *Grundgesetz* in 1949 has fundamentally inspired the evolution of German Private Law, through to its constitutionalization. In this

context, again, the role of the general clauses was – and continues to be – a fundamental one. Today, the general clauses are commonly described as “gateways” (*Einfallstore*) for the Basic Law’s value judgments (*Wertesystem*) in the realm of private law. In other words, the fundamental rights enshrined in the *Grundgesetz* do not have any direct effect on private law relationships, but it is the courts’ duty to take them into account while interpreting the general clauses of the BGB in order to provide them with more concrete content. The idea of the fundamental rights’ *mittelbare Drittwirkung* (indirect horizontal effect) was laid down for the first time in the famous Lütth Case of 1958 (7 BVerfGE 198), in which the BVerfG stated that the Basic Law contains an “objective order of values” (*objektive Werteordnung*) which must be looked upon as a fundamental constitutional decision affecting all areas of law, including private law.

The theory of the “radiating effect” (*Ausstrahlungswirkung*) of the basic rights in relations governed by private law has played a fundamental role in all subsequent court decisions on the subject and has given rise to a substantive body of case law developed by the courts. That it has also led to far-reaching consequences within the law of contract becomes particularly clear in the context of the famous *Bürgschaft* Case of 1993 in which the Federal Constitutional Court obliged civil courts to intervene on the basis of the general clauses of good morals and good faith – using the principles of party autonomy (Art. 2) and the social state (Art. 20) enshrined in the Constitution as interpretative guidelines – in a case in which the “structural imbalance of bargaining power” had led to an “exceptionally burdensome contract” for the weaker party (89 BVerfGE 214 (1993)).

The Europeanization of the BGB: Legal Developments in Private Law in the Twenty-First Century

The Act Modernizing the Law of Obligations (*Schuldrechtsmodernisierungsgesetz*) of 1 January 2002 is considered to be one of the most sweeping reforms of the BGB since it was enacted. The necessity to transpose several EC Directives,

among which the Consumer Sales Directive of 1999, was used as an occasion to realize the often called for fundamental reform of the law of obligations (the so-called *grobe Lösung*). This explains why most parts of the new provisions go back to and are largely based on reform proposals which had never progressed beyond the draft stage, prepared by a commission of leading academics and practicing lawyers in the early 1990s already. Because of its extremely short legislative process, the modernizing reform was heavily criticized by those who had argued in favor of a *kleine Lösung*, i.e., a legislation transforming only the Consumer Sales Directive without totally reforming the whole law of obligations. The Modernizing Act not only affected key elements of the general law of obligations and the contracts of the sale of goods, but also led to the integration into the BGB of many of the special statutes concerning consumer rights which over the years had developed outside the BGB and are the result of a transposition of EC Directives in this field. Since central parts of the new law of obligations have transposed European Directives, the reform is largely understood as a Europeanization of the BGB (Schulze and Schulte-Nölke 2001); the reform has thus also aimed to turn the BGB into a model in the context of the harmonization of European contract law. Having said that, the highly visible influence of the BGB within the framework of the Draft Common Frame of Reference (DCFR) is certainly also due to the leading role of German academics within the Joint Network on European Private Law entrusted with its elaboration.

The influence of EC/EU law has constantly grown over the last years, in particular in the field of consumer protection law which is increasingly inspired by the so-called information model. In German literature, the instrumentalization of the consumer as a means to promote the internal market on the basis of a model of procedural justice – more precisely of the idea that a sufficiently informed consumer is able to reach a completely rational decision – is said to have brought about a “dematerialization” and “re-formalization” of (national) consumer protection law (Micklitz 2012). The most recent example of this tendency is represented by the Consumer Rights Directive 2011/83/EU which was

implemented in June 2013 by means of an amendment of the law of obligations.

Legal Education and Legal Professions: The German Construct of *Einheitsjurist*

The particularity and main characteristic of the German system of legal education is the fact that it is embedded in the model of a “uniform jurist,” the so-called *Einheitsjurist*. The admission to all legal professions requires exactly the same formal qualification, so education is identical for everyone, regardless of the legal profession to be adopted in the future. Since federal legislation, i.e., the German statute of judges (*Deutsches Richtergesetz – DRiG*), *only* lays down general provisions and standards in the field of university education and practical training, each *Land* disposes of more detailed regulations. This has given rise to quite different rules – mainly related to procedural aspects of examination – in the 16 *Länder* (Keilmann 2006).

The legal education system is a two-phase one. It is a state-oriented and judge-centered bureaucratic model with traditionally great emphasis on the technical skills needed in the judiciary which has only been slightly reduced in recent times. It has its roots in the Prussian system of legal education and justice which already at the beginning of the eighteenth century was characterized by a state-controlled meritocratic selection for entry into the public administration on the basis of state examinations.

University studies in law end with an examination which today is called “first examination” (*erste Prüfung*) and no longer “first state exam” (*Erstes Staatsexamen*) because it is no longer entirely organized by the *Länder’s* Court of Appeals and the state offices for the Law Examinations (*Landesjustizprüfungsämter*). The modification brought about by the Act on the Reform of Legal Education of 2002 is an aggregate final exam mark, composed of an examination conducted by the law faculties in certain areas of specialization which students can choose (*universitäre Schwerpunktbereichsprüfung*) and a state exam which covers compulsory subjects (*staatliche Pflichtfachprüfung*). The former

represents 30, the latter 70 % of the final mark. Low average marks, high failure rates, and the assumption among students that university courses do not suitably prepare them for the exam have caused a run on the expensive service of private law teachers, the so-called *Repetitoren*. This dualism of abstract metatheory and practical legal training, which goes back to ideas diffused by Wilhelm von Humboldt at the beginning of the nineteenth century (Keilmann 2006), is heavily criticized but does not seem to draw to a close. Today, even the law faculties themselves offer special preparation classes in order to keep up with the private *Repetitoren* (Korioth 2006). The fact that privatization has affected the legal education system is also proved by the presence of the Bucerius Law School, the first and up to now the only private law school in Germany which was founded in 2000 and is located in Hamburg.

Legal studies are followed by a biennial practical training (*Referendariat* or *Vorbereitungsdienst*), organized and paid for by the *Land* in which it is undertaken. This period, during which the *Referendare* are instructed in various areas (civil courts, criminal courts or public prosecutor's office, public administration, and law firms) in order to obtain preparation for the main legal professions, ends with the "second state examination" (*zweite Staatsprüfung*). After successfully passing it – candidates only have two attempts, but failure rates are less than 15 % – the young lawyer is called *Assessor* or *Volljurist* (fully qualified jurist) and, having obtained "the qualification for the office of a judge" (§ 5 DRiG), he/she is theoretically qualified for any legal profession. That means, for example, that there is no additional bar exam, and to practice as a *Rechtsanwalt* (attorney), only a formal admission is required. However, access to a legal career is strongly determined by examination grades. Since for a career in the judiciary or civil service it is a precondition to achieve a top mark in the second state examination (the so-called *Prädikatsnote*) – a criteria which is fulfilled by fewer than 10 % of the aspirants – the overwhelming majority work as attorneys; only a small percentage has the option of becoming a judge (*Richter*), a public prosecutor (*Staatsanwalt*), a notary (*Notar*), or undertaking an

academic career. As a response to this reality, the Reform of 2002 has extended the practical training during the *Referendariat* at a law firm from 3 to a minimum of 9 and a maximum of 13 months.

Attorneys who have practiced as lawyers for at least 3 years can specialize in a particular field of expertise in order to attain the title of *Fachanwalt*. In a few *Länder*, attorneys can also become the so-called *Anwaltsnotare* (lawyer notaries) who combine the profession of an attorney with that of a notary on the condition that they have practiced the forensic profession for 5 years. They are the second major type of notaries beside the so-called *Nurnotare* (fulltime notaries). Both groups have to pass a third exam, the so-called *notarielle Fachprüfung*. Due to the limited number of posts available, competition is hard, and the door to this profession is only open to the top law graduates.

As seen above, the judiciary follows a separate career path: it is immediately entered after the second state examination by those who have achieved excellent final marks, and only rarely does a practicing lawyer transfer to become a judge. The initial appointment to a local or a regional court will only be for a certain period, usually for 3 years, during which the candidate is a *Richter auf Probe*, i.e., his employment status is "on probation." A *Richter auf Lebenszeit*, a lifetime judge, can be promoted to higher regional courts on the condition that he has worked for at least 10 years in the courts of first instance. Whereas the initial appointments and this kind of promotion are made by the state's ministers of justice, appointments to the five federal courts involve the competent Federal Minister and a so-called *Richterwahlausschuss*, a committee for the selection of judges which consists of the competent *Land* ministers and an equal number of members elected by the *Bundestag*.

References

- Böhm F (1966) Privatrechtsgesellschaft und Marktwirtschaft. *ORDO* 17:75–151
- Bommarius C (2011) Lecture – Germany's Sozialstaat principle and the founding period. *German Law J* 12: 1879–1886. Available at <http://www.germanlawjournal>.

- com/index.php?pageID=11&artID=1389. Bonefeld 2012
- Bonefeld W (2012) Freedom and the strong state: on German ordoliberalism. *New Polit Econ* 17:633–656
- Häberle P (1987) Die Menschenwürde als Grundlage der staatlichen Gemeinschaft. In: Isensee J, Kirchhof P (eds) *Handbuch des Staatsrechts*, vol 1. C.F. Müller, Heidelberg, pp 815–861
- Joerges C, Roedl F (2004) “Social market economy” as Europe’s social model? European University Institute. Available at <http://ssrn.com/abstract=635362>
- Keilmann A (2006) The Einheitsjurist – a German phenomenon. *German Law J* 7:293–312. Available at <http://www.germanlawjournal.com/index.php?pageID=11&artID=712>
- Kielmansegg PG (1990) The basic law – response to the past or design for the future? In: Lehmann H, Ledford K (eds) *Forty years of the Grundgesetz*. German Historical Institute, Washington, DC, pp 5–18. Available at <http://www.ghi-dc.org/publications/ghipubs/op/op01.pdf>
- Korioth S (2006) Legal education in Germany today. *Wis Int Law J* 24:85–107
- Markesinis BS, Unberath H, Johnston A (2006) *The German law of contract: a comparative treatise*. Hart Publishing, Oxford/Portland
- Micklitz HW (2012) Do consumers and businesses need a new architecture of consumer law? A thought-provoking impulse. European University Institute, Florence. Available at http://cadmus.eui.eu/bitstream/handle/1814/23275/LAW_2012_23_Rev.pdf?sequence=3
- Müller-Armack A (1933) *Staatsidee und Wirtschaftsordnung im neuen Reich*. Juncker und Duennhaupt, Berlin
- Müller-Armack A (1966) *Wirtschaftsordnung und Wirtschaftspolitik*. Rombach, Freiburg
- Rüthers B (1989) *Entartetes Recht. Rechtslehren und Kronjuristen im Dritten Reich*. C.H. Beck, München
- Scharpf F (2009) *Föderalismusreform: kein Ausweg aus der Politikverflechtungsfalle?* Campus-Verlag, Frankfurt/New York
- Schmidt-Rimpler W (1941) Grundfragen einer Erneuerung des Vertragsrechts. *Archiv für die civilistische Praxis* 147:130–197
- Schulze R, Schulte-Nölke H (eds) (2001) *Die Schuldrechtsreform vor dem Hintergrund des Gemeinschaftsrechts*. Mohr Siebeck, Tübingen
- Somma A (2013) Private law as biopolitics: ordoliberalism, social market economy, and the public dimension of contract. *Law Contemp Probl* 76:105–116. Available at <http://scholarship.law.duke.edu/lcp/vol76/iss2>
- Stolleis M (1998) *The law under the swastika*. The University of Chicago Press, Chicago
- Stolleis M (2002) *Geschichte des öffentlichen Rechts in Deutschland: Weimarer Republik und Nationalsozialismus*. C.H. Beck, München
- Zacher HF (1987) Das soziale Staatsziel. In: Isensee J, Kirchhof P (eds) *Handbuch des Staatsrechts der Bundesrepublik Deutschland*, vol 1. C.F. Müller, Heidelberg, pp 1045–1112

- Zitelmann E (1900) Zur Begrüßung des neuen Gesetzbuches. *Deutsche Juristen-Zeitung* 5:2–6
- Zweigert K, Kötz H (1994) *An introduction to comparative law*. Clarendon, Oxford

Gibbard-Satterthwaite Theorem

Pierre Bernhard¹ and Marc Deschamps²

¹Biocore team, Université Côte d’Azur-INRIA, Sophia Antipolis Cedex, France

²CRESE EA3190, Université Bourgogne Franche-Comté, Besançon, France

Definition

One seminal question in social choice theory was: is it possible to find a social choice function such that each agent is always better off when telling the truth concerning his preferences no matter what the others report? In other words, can we find a strategy-proof voting rule? With at least three alternatives and two voters, the answer is clearly no under a very general framework, as was proved independently by Allan Gibbard and Mark Satterthwaite. Since then, the Gibbard-Satterthwaite theorem is at the core of social choice theory, game theory, and mechanism design.

Introduction

Since K. Arrow’s 1951 analysis, which marks the revival of the theory of social choice, economists investigate from an axiomatic point of view the aggregation of individual preferences in order to obtain a social welfare function (i.e., a complete and transitive ranking based on the individual preferences) or a social choice function (i.e., one alternative from the individual preferences). Such questions concern huge domains of human beings, for example, the family’s choice of the walls color in the living room, the choice of the Palme d’Or winner by the jury of the Cannes International Film Festival, or the vote for the President of the European Commission. Thus, in

addition to his (im)possibility theorem, K. Arrow has initiated a new domain in economic analysis.

Since then, the question of the strategic behavior of individuals during an election was raised again on several levels. Indeed for a very long time, since we find, for example, a letter from Pline the Younger in Roman Antiquity, the question of manipulation had attracted attention. Whether playing on who will be a voter, who can be a candidate, the choice of the voting rule, abstention, beliefs about preferences or preferences expressed during the vote, the idea that at least one individual can do a manipulation to his advantage has interested and concerned many thinkers.

Concerning this last form of manipulation, it was historically mentioned at least seven times before the Gibbard-Satterthwaite theorem (an expression first used by (Schmeidler and Sonnenschein 1978)). First, a story probably apocryphal tells that Borda responded to one of his critics who pointed out that his method was manipulable that it was intended for use by honest people. There is also a reference to this question in Ch. Dodgson (alias Lewis Carroll), who said in a specific voting system, “This voting principle makes an election more of a skill game than real test of voters’ wishes” (see Black (1958)). In the modern period, Black (1948) discusses the link between unimodal preferences and strategic votes, while (Arrow 1951, p. 7) explicitly states that he will not deal with this issue even though he later returns to it in a footnote [footnote 8, pp. 80–81]. Arrow’s general analysis, however, will lead (Vickrey 1960, pp. 517–519) to conjecture that immunity to strategic manipulation is logically equivalent to the association of the axiom of independence of irrelevant alternatives and that of positive association. Finally, it will be R. Farquharson in his 1958 doctoral dissertation, published in 1969, (Farquharson 1969), who will introduce the distinction between “sophisticated strategy” and “sincere strategy,” and then in an article with M. Dummett in 1961 when they make the conjecture: “It seems unlikely that there is any voting procedure in which it can never be advantageous for any voter to vote ‘strategically,’ i.e., non sincerely” (Dummett and Farquharson 1961,

p. 34). In an interview given in 2006 to R. Fara and M. Salles, M. Dummett confesses that he felt at this time that proving this conjecture would be extremely difficult and that is why they did not try the demonstration. We refer the reader to (Barberà 2010) for more details on this historical part.

Thus, for more than 20 years after (Arrow 1951), all scholars of social choice theory seem to have been convinced that the question of the manipulation of preferences by an individual (i.e., the fact that he does not express his true preferences in order to lead to a social choice that satisfies him better than would have been obtained if he had been honest) was an important question, easy to explain, but very difficult to prove.

As an illustration of this question of the manipulation of preferences by an individual, we can give the following example (from (Feldman 1979, p. 459)):

Imagine a committee made up of 21 members each having one vote and whose actual preferences presented in descending order can be broken down into three groups.

Type 1	Type 2	Type 3
A	B	C
B	C	B
C	A	A
10 voters	9 voters	2 voters

In a majority election where voters indicate their real preferences, A gets 10 votes, B gets 9 votes, and C gets 2 votes; this leads to the election of A. However, if anticipating this result, the voters in group 3 manipulate their preferences and vote for B, this will lead to 10 voters for A and 11 voters for B and so B will be elected. This strategic choice of the voters of the group 3 allows them to obtain a more favorable result than they would have obtained if they had voted sincerely. **I**

Therefore, a question immediately comes to mind: is it possible to imagine a social choice function such that each individual, regardless of the others’ choices, is always better off by expressing his true preferences? In other words, is there a social choice function such that always telling the truth is a strictly dominant strategy for

each individual? The answer to this question was independently given by the philosopher Allan Gibbard (1973) and the economist Mark Satterthwaite (1975) (work summarizing part of his doctoral dissertation defended in 1973) and it is unfortunately negative. What we call since the Gibbard-Satterthwaite theorem can be broadly stated as follows: *it is not possible to find a social choice function that is both nonmanipulable and nontrivial.*

This last term deserves an explanation. By a trivial function we mean one of the following solutions (or a solution equivalent to it): 1/ a dictatorship (i.e., all the power of decision resides with a single individual and he has no indifferent choice), 2/ the permanent choice of an alternative whatever the preferences expressed by the individuals are (which could, for example, correspond to a tradition which would be imposed in all circumstances), 3/ the perfect unanimity of all the voters (which in fact means having perfect clones and therefore no diversity in the preferences), or 4/ the majority between two alternatives only, whatever the number of other alternatives existing and the votes that are expressed for them. Thus, if all individuals know their preferences and those of others and that there are at least three possible alternatives and at least two voters, there is no social choice function that implies that each individual always has an interest in expressing his or her real preferences. We thus find here the conditions of Arrow's theorem and the same conclusion since when there are only two alternatives, majority voting is at the same time a nondictatorial and nonmanipulable social choice function.

Since then, the Gibbard-Satterthwaite theorem has been considered, along with Arrow's theorem, as one of the two most famous results of social choice theory and has led to a very large literature in this field. It also plays a crucial role in public economics and in the theory of incentive mechanisms, which can be broadly seen as a social engineering approach of finding the rules to achieve a specific outcome from agents interacting strategically and having private information (see Börger (2015)). Indeed, because of

this theorem, the incentives of individuals must be considered as relevant constraints in the design of any mechanism.

Gibbard-Satterthwaite Theorem

Many proofs of this theorem have been proposed and it is possible to consider that they take one of the following four paths: 1/ that used by A. Gibbard and which uses Arrow's theorem, 2/ that used by M. Satterthwaite thanks to a combinatorial argument and recurrences on the number of individuals and alternatives, 3/ that considering this theorem as the consequence of the fact that nonmanipulability requires strong monotony (see Moulin (1988)), and 4/ that developed by S. Barberá and his coauthors using the concept of pivotal agents. For a first presentation of the question of manipulation, we refer to Feldman (1979) and for a review of this literature we refer to Sprumont (1995) and Barberá (2010).

Following Gibbard's approach, we will prove the Gibbard-Satterthwaite theorem as a corollary of Arrow's (im)possibility theorem. The presentation we retain will distinguish the formal setup, the links between the two theorems, the proof of the Gibbard-Satterthwaite theorem, and that Arrow's theorem in turn can be seen as a corollary if we directly prove the Gibbard-Satterthwaite theorem.

Formal Set-up

Let $\mathcal{A} = \{a, b, \dots\}$ be a set of three or more alternatives. Let P be the set of linear orders over $\mathcal{A}$, and $\mathcal{P} = P^n$. An element $\Pi = (P_1, P_2, \dots, P_n) \in \mathcal{P}$ is called a *profile*, and the P_i the *individual preferences*. The order in P_i is denoted $a \succ_i b$ for "player i prefers a to b ." We further define

Definition 1

The domination set of $a \in \mathcal{A}$ in P_i

$$\mathcal{D}(a, P_i) = \{x \in \mathcal{A} \mid a \succ_i x\}$$

Moreover, for $\Pi = (P_1, \dots, P_n)$, $\mathcal{D}(a, \Pi)$ stands for the product set of the $\mathcal{D}(a, P_i)$. Thus,

let $\Pi' = (P'_1, \dots, P'_n)$ be another profile, the notation $\mathcal{D}(a, \Pi') \supseteq \mathcal{D}(a, \Pi)$ means: $\forall i \in \{1, \dots, n\}$, $\mathcal{D}(a, P'_i) \supseteq \mathcal{D}(a, P_i)$.

Definition 2

- A social welfare function (or SWF) is an application $f: \mathcal{P} \rightarrow P$.
- A social choice function (or SCF) is an application $F: \mathcal{P} \rightarrow A$.

The order relation in $f(\Pi)$ will be denoted $a > b$ or, if needed $a >_{f(\Pi)} b$. We need to define the following properties of SWF or SCF:

Definition 3

- The SWF f is Pareto efficient (or satisfies the unanimity rule) if

$$[\forall i \leq n, a \succ_i b] \Rightarrow a > b.$$

- The SCF F is Pareto efficient, (or satisfies the unanimity rule) if whenever an alternative $a \in A$ is the most preferred alternative in all individual preferences, it results that $F(\Pi) = a$.
- The SWF f is independent of irrelevant alternatives (IIA) if, for any a and b in A , the relative ranking of a and b in $f(\Pi)$ only depends on their relative rankings in the P_i , irrespective of the rankings of other alternatives.
- The SCF F is monotonic if, given two profiles Π and Π' ,

$$[F(\Pi) = a \text{ and } \mathcal{D}(a, \Pi') \supseteq \mathcal{D}(a, \Pi)] \Rightarrow F(\Pi') = a.$$

- The SCF F is strategy-proof if, when Π and Π' differ only in changing P_i into P'_i , it results that either $F(\Pi) = F(\Pi')$ or $F(\Pi) \succ_i F(\Pi')$.
- The SWF f is called dictatorial if there exists a player k such that, $\forall \Pi \in \mathcal{P}$, $f(\Pi) = P_k$. (The social preferences are always player k ' preferences.)

- The SCF F is called dictatorial if there exists a player k such that, $\forall \Pi \in \mathcal{P}$, $\forall x \neq F(\Pi)$, $F(\Pi) \succ_k x$. ($F(\Pi)$ is player k 's most preferred alternative.)

Arrow and Gibbard-Satterthwaite Theorems

Theorem 1 (Arrow) If a SWF is Pareto efficient and IIA, then it is dictatorial.

Theorem 2 (Gibbard-Satterthwaite) If a SCF is strategy-proof and onto (i.e., its range is all of A : $\forall a \in A, \exists \Pi \in \mathcal{P}$ such that $F(\Pi) = a$), it is dictatorial.

Lemma 1 (Muller and Satterthwaite) If a SCF is strategy-proof and onto, it is monotonic and Pareto efficient.

Proof Let Π and Π' be two profiles, $F(\Pi) = a$, $\mathcal{D}(a, \Pi') \supseteq \mathcal{D}(a, \Pi)$, but assume that $F(\Pi') \neq a$. Make the change from Π to Π' one player at a time in numeric order. Denote by Π_k the profile obtained after changing P_k to P'_k . (And $\Pi_0 = \Pi$). At some point, we have that $F(\Pi_{i-1}) = a \neq b = F(\Pi_i)$. By strategy-proofness, it follows that $a \succ_i b$ in P_i while $b \succ_i a$ in P' . But by hypothesis, if $a \succ_i b$ in P_i it is a fortiori true in Π' , a contradiction. Therefore, the SCF is monotonic.

Because F is assumed to be an onto function, for any given $a \in A$, there exists a profile Π such that $F(\Pi) = a$. Build Π' by moving a at the top of the preferences of all players. By monotonicity, it still holds that $F(\Pi') = a$. Now, get Π'' by shuffling at will the preferences of all players below a , leaving a at their top position. Because of the definition of monotonicity, and specifically its IIA-like character, it still holds that $F(\Pi'') = a$. Therefore, the SCF is Pareto efficient. **I**

Proof of the Gibbard-Satterthwaite Theorem

This note is devoted to the proof of the Gibbard-Satterthwaite theorem viewed as a corollary of Arrow's theorem. We assume therefore that the latter is known. Given the above Lemma 1, we need to prove

Lemma 2 *If a SCF is Pareto efficient and monotonic, it is dictatorial.*

Our method of proof will be as follows: we assume that a SCF F is known which is Pareto efficient and monotonic. From it we construct a SWF f which we show to be Pareto efficient and IIA, therefore dictatorial, which will imply that F is dictatorial.

Let F be a SCF. Given a profile Π , our social ordering $f(\Pi)$ is built via the following algorithm:

Algorithm

- Let $\Pi_1 = \Pi$.
- For i ranging from 1 to n , do:
 - define $a_i = F(\Pi_i)$,
 - define Π_{i+1} by moving a_i at the bottom of all individual preferences in Π_i .
- Define $f(\Pi)$ as: for all $i \in \{1, \dots, n-1\}$, $a_i >_{f(\Pi)} a_{i+1}$.

Proposition 1 *If the SWF F is Pareto efficient and monotonic, the above algorithm produces a linear order on $\mathcal{A}$.*

Proof What we need to prove is that all elements of $\mathcal{A}$ will be ranked, i.e., that for all $i \in \{1, \dots, n\}$, and for all $j < i$, $a_i \neq a_j$.

Let first $i < n$. Therefore, not all alternatives have been numbered as one of the a_k . Define Π'_i as follows: Take an alternative b which has not yet been ranked. Raise it at the top of all individual preferences. This does not change the domination sets: $\mathcal{D}(a_j, \Pi'_i) = \mathcal{D}(a_j, \Pi_i)$ since in Π_i , all the alternatives that have already been selected in the algorithm, including a_j (remember that $j < i$), are stacked at the bottom of all individual preferences. Therefore, by monotonicity, if $F(\Pi_i) = a_j$, it also holds that $F(\Pi'_i) = a_j$. But by Pareto efficiency, $F(\Pi'_i) = b$, a contradiction.

Finally, in Π_n , $n-1$ different alternatives have been placed at the bottom of all individual preferences; thus, the same and last alternative is alone at the top of all and is therefore selected by F by Pareto efficiency.

The following propositions end our proof:

Proposition 2 *If the SCF F is Pareto efficient and monotonic, the SWF defined by the algorithm is*

1. *Pareto efficient*
2. *Independent of irrelevant alternatives (IIA), and therefore dictatorial by Arrow's theorem*

Proof

1. Let $a, b \in \mathcal{A}$, and assume that in Π , $\forall i \in \{1, \dots, n\}$, $a \succ_i b$. Assume that at some step of our algorithm, $F(\Pi_i) = b$, but a has not yet been chosen. In Π_i , raise alternative a at the top of all individual preferences. This does not change the domination sets of b in any individual preferences. Therefore, F should still select b . But by Pareto optimality, it should select a . A contradiction.
2. Let $i < j$ and therefore by our algorithm, $a_i >_{f(\Pi)} a_j$. In Π , move another alternative b in some individual preferences, call the new profile Π' . We claim that in the order created by our algorithm applied to Π' , it still holds that $a_i > a_j$.

Assume that $F(\Pi') = b$. Then at step 2, b is brought at the bottom of all individual preferences, and, as compared to profile Π , the domination sets of a_1 have been enlarged or kept unchanged for all individual preferences. Therefore, $F(\Pi'_2) = a_1$. If $b \neq a_2$, at step 2, for the same reason a_2 will be selected, and so forth until the end of the algorithm. Therefore, we will still select a_i before a_j .

Assume $F(\Pi') = c \neq b$, and assume $c \neq a_1$. This is possible only if in one individual preferences at least, b has been moved from below a_1 to above it. In these individual preferences only, bring b back below a_1 . Call this profile Π'' . $\mathcal{D}(a_1, \Pi'') \supseteq \mathcal{D}(a_1, \Pi)$. Therefore $F(\Pi'') = a_1$. But in going from Π' to Π'' , b has only been moved down in some individual preferences. Therefore $\mathcal{D}(c, \Pi'') \supseteq \mathcal{D}(c, \Pi')$. Therefore, $F(\Pi'') = c$. Hence, $a_1 = c$ contrary to the hypothesis. Hence, $F(\Pi') = a_1$.

Repeat this argument at each step before b is selected, and once this happens, repeat the

previous argument. It follows that, except for b , all other alternatives are chosen in the same order as for Π . **I**

It follows that, by Arrow's theorem, f is dictatorial. There is a player k such that for all Π , $f(\Pi) = P_k$, and in particular $F(\Pi)$ is player k 's preferred alternative. And this proves Lemma 2, and consequently the Gibbard-Satterthwaite theorem. **I**

Complement: Arrow's Theorem as a Corollary of Gibbard Satterthwaite

It can also be shown that if the Gibbard-Satterthwaite theorem has been proved directly, Arrow's theorem is an easy corollary, thus establishing the equivalence between the two results. The process is symmetrical from the above one, deriving a SCF from a Pareto efficient and IIA SWF by simply picking the top alternative in the social preferences and showing that it is onto and strategy-proof.

We might also mention that Lemma 1 has an easy reciprocal.

Conclusion

Since its demonstration, the Gibbard-Satterthwaite theorem has generated a huge literature on the question of the manipulation of preferences, distinguishing, in particular, the cases where the social choice function is manipulable by an individual from the case where it is manipulable by a coalition of individuals. It has also been extended to take into account set-valued (Duggan and Schwartz 2000) and nondeterministic choice functions (Gibbard 1977), that is, those where the result depends on the votes of individuals but also by chance. (Often both set-valued and non-deterministic.) Our framework requires that social choice function is onto. It has been proven since that the results hold as soon as the social choice function has at least three alternatives in its range.

In our opinion, three main ways of circumventing this theorem have been followed. The first is to restrict the preference domain (with functions defined on restricted sets of preference profiles, it is

possible to find social choice functions that are both nondictatorial and nonmanipulable, e.g., unimodal preferences *à la* Black). The second is to change the goal. Indeed, the framework in which the Gibbard-Satterthwaite theorem is situated is very strong since it seeks a social choice function for which telling the truth is a dominant strategy for each individual. The path followed by implementation theory is to simply ask that it be a Nash strategy, or a perfect subgame strategy, or a Bayesian Nash strategy, depending on the informational context of the individuals. Finally, more recently, a third way explains that this problem is real but may not be very important empirically. Indeed, besides the integrity, ignorance or stupidity of individuals that can prevent them from performing manipulations, the fact that a social choice function is manipulable does not imply that it will be manipulated. And since Bartholdi et al. (1989), economists consider that it may be empirically impossible for individuals to decide how to manipulate even when they have all the information to do so, as the problem may be NP-hard.

Cross-References

- [Electoral Systems](#)
- [Heterarchy](#)
- [Liberty](#)
- [Political Competition](#)
- [Political Economy](#)
- [Politicians](#)
- [Public Goods](#)
- [Public Interest](#)
- [Rationality](#)
- [Simple Majority](#)

References

- Arrow KJ (1951) Social choice and individual values. Wiley, New York
- Barberà S (2010) Strategy-proof social choice. Barcelona economic working paper 420. University of Barcelona, Spain
- Bartholdi J, Tovey C, Trick M (1989) The computational difficulty of manipulating an election. Soc Choice Welf 6:227–241

- Black D (1948) On the rationale of group decision making. *J Polit Econ* 56:23–34
- Black D (1958) The theory of committees and elections. Cambridge University Press, Cambridge, UK
- Börger T (2015) An introduction to the theory of mechanism design. Oxford University Press, New York
- Duggan J, Schwartz T (2000) Strategic manipulability without resoluteness or shared beliefs: Gibbard-Satterthwaite generalized. *Soc Choice Welfare* 17:86–93
- Dummett M, Farquharson R (1961) Stability in voting. *Econometrica* 29:33–44
- Farquharson R (1969) Theory of voting. Yale university Press, New Haven
- Feldman AM (1979) Manipulating voting procedures. *Econ Inq* 17:452–474
- Gibbard A (1973) Manipulation of voting schemes: a general result. *Econometrica* 41:587–601
- Gibbard A (1977) Manipulation of schemes that mix voting with chance. *Econometrica* 45:665–682
- Moulin H (1988) Axioms of cooperative decision making. Economic Society monographs. Cambridge University Press, Cambridge, MA
- Satterthwaite M (1975) Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *J Econ Theory* 10:187–217
- Schmeidler D, Sonnenschein H (1978) Two proofs of the gibbard-satterthwaite theorem on the possibility of a strategy-proof social choice function. In: Gottinger HW, Leinfellner W (eds) Decision theory and social ethics, Issues in Social Choice. Reidel Publishing Company, pp 227–234. Published version of a 1974 working paper. Dordrecht/Holland/Boston: USA London: England
- Sprumont Y (1995) Strategy-proof collective choice in economic and political environments. *Can J Econ* 28:68–107
- Vickrey W (1960) Utility, strategy and social decision rules. *Q J Econ* 74:507–535

Global Warming

Alfred Endres

FernUniversität Hagen, Hagen, Germany

Abstract

This essay analyses global warming from the perspective of environmental economic theory: National activities to curb greenhouse gas emissions are public goods. It is well known that the expectations to arrive at a satisfactory provision of these kinds of goods by

uncoordinated decision-making are low. Nations try to overcome this dilemma by international treaty-making. However, economic analysis suggests that the genesis and persistence of international climate agreements are challenged by problems of national rationality and stability as well as by the leakage effect. On the other hand, economic analysis reveals some reasons for a more optimistic view. Among those are incentives to develop and introduce green technologies and incentives to adapt to changes in the world climate.

Synonyms

[Greenhouse effect](#)

Introduction

Human economic activity is directed to produce commodities and services which are valued by people. Many of these goods are sold and bought in markets where consumers pay a market price to reimburse the production costs to the supplying firms and contribute to the profits of these firms. Jointly with these intended products of the economic activity, unwanted side products are generated, like waste and emissions into environmental media (e.g., air and water). The generation of these side products can be steered in terms of quantity and quality, but it is unavoidable in principle.

There are millions of different types of substances which are emitted into the environment as by-products of the productions process. A certain subset of these emissions constitutes the group of *greenhouse gases (GHG)*. Among this type of gases are carbon dioxide (CO₂), nitrous oxide (N₂O), chlorofluorocarbons (CFCs), methane (CH₄), and ozone (O₃). There is a broad consensus in the natural sciences that the emission of these greenhouse gases changes the atmospheric conditions around the globe to such an extent that the world climate is adversely affected. Particularly, increases in the stock of greenhouse gases in the atmosphere contribute to an increase of the average temperature on earth (*global warming*). This

is a phenomenon of fundamental concern in the society and for policy makers. It is also an important issue of research undertaken and policy advice given by economists. Climate change affects the living conditions of the people directly, and it also changes the conditions under which many commodities and services are produced (e.g., in agriculture and tourism).

From the perspective of sciences, climate change is an issue which has so many dimensions intertwined in a very complex way that it can only be dealt with using an interdisciplinary approach. Of course, in this interdisciplinary process of communication, collaborators from the social sciences (like economics) cannot really assess the quality of the research in the natural sciences. So economists usually take for granted what the mainstream of natural science research takes to be true. This is particularly so with regard to the quantitative effect that certain greenhouse gases have on the world climate and in terms of the effect that certain changes in the climate have on the living conditions on earth (e.g., sea level rise and increases in extremely adverse weather conditions).

There is a broad consensus in the world that global warming has anthropogenic origins and is extremely dangerous to human kind. Even though not each consequence of higher average temperature in the world is detrimental, the aggregate effect appears to be a severe threat to the living conditions on earth. (A standard reference in this context is the *Stern Review*, Stern (2007), see also Stern (2015))

The problem is very well known among the political leaders of the world and to the people, at least in the economically developed countries. (The beginning of the international efforts to protect the climate was marked by the *Earth Summit* in Rio de Janeiro in 1992.) On the other hand, there is no progress in terms of a reduction in the globally aggregated quantity of GHG emissions. In the public media, “uninterested and incompetent politicians” are often held responsible for this lack of progress. However, economic analysis shows that the climate problem is very hard to solve even if we assume that the involved governmental decision-makers are competent and seek to

serve the interest of the people they represent. This is due to an adverse incentive structure that impedes an international consensus on an effective climate change policy.

An Economic Perspective of Global Warming

Greenhouse gases are to be distinguished from most other pollutants in that they do not have a regionally specific dose–response profile. Instead, they are global in the sense that after their emission they disperse homogeneously around the atmosphere of the earth. Particularly, the regional structure of these pollutants’ generation is irrelevant for their effect on the global climate. Changes in the world climate are induced by the aggregate effect of GHG emissions and, particularly, by the stock of greenhouse gases in the atmosphere that is accumulated by the emission flow generated in each consecutive period of time. Thereby, global pollutants have properties that come very close to the economics textbook stylization of a “pure public good (bad),” which is an important issue in *microeconomic theory*, *public finance*, and *environmental economics*. The problems that arise when governments strive to agree upon curbing GHG emissions to a tolerable level can thereby be explained using insights from the economic theory of public goods.

Concerning greenhouse gases, each individual country is in a double role: On the one hand, the country is the generator of the problem emitting certain quantities of these gases. On the other hand, the country is the victim of its own activity suffering from the detrimental effect this activity has on the climate. However, since it is not a country’s specific climate which is under consideration here but the world climate, the country is not the only victim of its own activity. All other countries are also victimized by the pollution of the individual country under consideration. In terms of economic theory, each country generates an *internal effect* as well as an *external effect* by its GHG emissions. The problem arising here is that a government striving to maximize the welfare of the citizens that it represents designs a GHG

policy that takes the internal effect into account and ignores the external one.

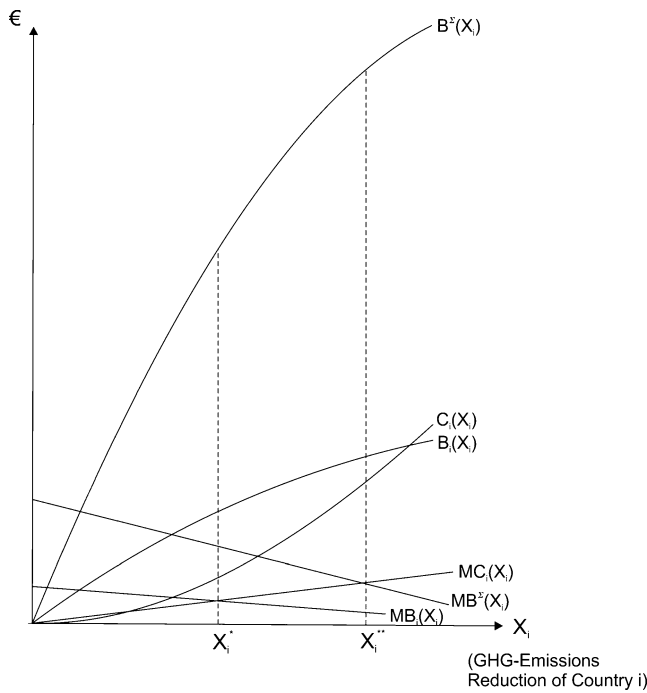
The welfare effects of a country's GHG policy are two-fold. On the one hand, the national economy has (the citizens have) to bear the costs of the policy. This reduces national welfare. On the other hand, the country benefits from its own policy in that the GHG-reducing measures are advantageous in terms of the world climate. Therefore, the climate change damages the country under consideration suffers from are reduced. The national welfare effect of the national climate policy is maximized when this policy is designed to maximize the difference between the national benefits (reduction in national climate change damages) and national climate change costs. The quantity of GHG reduction for which the aforementioned difference is maximized is called the *equilibrium quantity* of emission reduction for the country under consideration. It is fundamental to note that this nationally optimally designed climate policy is drastically different from the design of a national climate policy which is optimal from the point of view of the world as a whole. From this global point of view, in order to characterize optimal national climate policy, the cost of this

policy has to be compared with the benefits of this policy aggregated across all the countries of the world (instead of the country under consideration only). The quantity of country i 's emission reduction which maximizes the difference between aggregate benefits and costs to country i is called the *globally optimal quantity* of i 's emission reduction.

In economics, it is a standard exercise to clarify and specify these thoughts using formal mathematical analysis or graphical visualization. In what follows, we turn to the latter Fig. 1.

In Fig. 1, the term B_i denotes the benefit, B , of GHG emission reduction in some country, i . X_i denotes the quantity of this reduction in the aforementioned country. The function $B_i(X_i)$ denotes these benefits as they depend upon the quantity of the pollutants reduced. It is assumed in this figure that this function is known to the decision-makers. In reality (and in more elaborate theoretical expositions), it is to be acknowledged that the extent of damages and the manner in which they systematically depend on the extent of emissions reduction is highly uncertain. There is an extensive literature dealing with this issue, which is ignored here. In the present context, the figure is

Global Warming,
Fig. 1 The Benefits and
Costs of GHG abatment:
The National and the Global



used to illustrate the divergency between globally optimal emission reductions and the emission reductions chosen by individual countries striving to maximize the welfare of their respective citizens. In this context, it is sufficient to interpret the curves in Fig. 1 to be the best estimate of the relationships dealt with here that is available to the political decision-makers. Of course, a benefit curve like the one that has been drawn for country i can be drawn for any other country. This is not done in the figure. However, a curve $B^{\Sigma}(X_i)$ is drawn representing the vertical aggregation of the individual benefit curves of all countries. This aggregate curve shows how the sum of worldwide benefits from pollution reduction depends upon emissions reduction of country i . Additionally, Fig. 1 presents the curve $C_i(X_i)$ representing how the costs of emission reduction in country i depend upon the quantity of emissions that this country reduces. Comparing these curves, it is obvious that the globally optimal level of GHG reductions in country i is at X_i^{**} , the quantity where the difference between the aggregate benefit curve and the curve representing the abatement cost of country i is maximized. The quantity of emission reductions maximizing the welfare in country i (the “equilibrium quantity” for country i), however, is at X_i^* . In addition to these curves representing total benefits and costs, the figure shows the curves $MB_i(X_i)$, $MB^{\Sigma}(X_i)$, $MC_i(X_i)$. These curves are “partner curves” to the aforementioned curves in that they represent the marginal country-specific benefits, marginal aggregate benefits, and marginal country-specific costs instead of the total benefits and costs. These curves represent the country-specific benefits, aggregate benefits, and country-specific costs, respectively, for small increases of the quantity of emissions reduced. Technically speaking, these marginal curves are the first-order derivatives of the total curves. The equilibrium emission reduction quantity, X_i^* , for country i is defined by the intersection of the country’s specific marginal benefit curve with the country’s specific marginal cost curve. For any quantity lower than X_i^* , the marginal benefit of the last unit’s reduction is higher than the cost of this last unit. Therefore, national welfare could be increased by reducing

an additional unit. It follows that the starting point emission reduction level in this little thought experiment cannot be the equilibrium one. Analogously, for any level of pollution abatement beyond X_i^* , it can be argued that national welfare increases when pollution abatement is attenuated. By analogous reasoning, it can be shown that the globally optimal quantity of emission reduction, X_i^{**} , is defined by the intersection of the marginal aggregate benefit curve, $MB^{\Sigma}(X_i)$, and the marginal country-specific cost curve, $MC_i(X_i)$. In economic analysis (as applied to environmental problems and very many other issues), marginal analysis is a very common and powerful tool. It will be repeatedly used in a subsequent exposition.

International Environmental Agreements – Painful Genesis and Fragile Architecture

In the preceding section, the decisions of national governments have been stylized regarding the quantity of GHG reductions. It has been shown that a national GHG policy striving to maximize national welfare is unable to meet the global requirements of effective GHG reduction.

Even though we obviously do not have a “world government,” it is instructive to imagine what a government like that might do. The result of this reasoning may be used as a benchmark which serves for the results of real world decision processes. To build up this benchmark, it is usually assumed in economics that this world government is well informed and benevolent. Applying the established economic welfare criteria, the world government would choose a situation with two essential properties. The first property is that GHG emissions would be at their globally optimal level as characterized in the preceding section. The second property relates to the question, how much each individual country would have to contribute to the globally optimal aggregate emission reduction. Thriving to keep the worldwide total costs of GHG abatement at their minimum, the world government would put a burden to an individual country which would be

the higher, the lower the abatement cost of this country is. More precisely, the cost-effective distribution of country-specific emission reduction loads is characterized by the marginal abatement costs to be equal across all countries. This “equi-marginal condition” is very often used in economics (within the environmental context and beyond) and is explained more elaborately, e.g., in Endres (2011), pp. 216–220. If the countries of the world are aware of what has just been said and note that this solution cannot be achieved by country-specific “autistic” welfare maximization, they might try to get there by *voluntary agreement*.

The fundamental challenge to the design of international agreements to fight global warming is that they must be *self-enforcing* since international law is too weak to implement enforcement. This means that the international contract must be designed in a manner that it is attractive for each country to join in, and having joined, to keep the commitments. These two properties are called the *individual rationality* and the *stability* of an international treaty.

Very unfortunately, it does not follow from the global optimality property of an international GHG treaty that it is also individually rational and stable. To the contrary, for many countries there is an incentive to “cheat” upon the commitments they have signed in an international agreement by taking a *free ride* on the GHG abatement efforts of the other partaking countries (lacking stability). Moreover, some countries might have an incentive not to sign the treaty at all, even though this treaty is optimal from the global point of view (lacking individual rationality). This is particularly so for countries with low marginal abatement costs which do not suffer very much from global warming (or may even benefit from it).

There are many attempts in the literature to fill these rather theoretical considerations with empirical content. A prominent example is a study by Finus/van Ierland/Dellink (2006). (There is a textbook-style interpretation of the aforementioned original research in Endres (2011).) To investigate the incentive compatibility of a globally optimal GHG agreement, the authors of this study proceed as follows. They divide the world

into 12 “regions” and use data from the empirical literature to estimate the GHG abatement benefit and cost functions for each of those regions. These 12 regions (countries or groups of countries, respectively) are USA, Japan, European Union, other OECD countries, Central and Eastern European countries, former Soviet Union, energy-exporting countries, China, India, dynamic Asian economies, Brazil, and “rest of the world”. The result in terms of individual rationality is that three regions would deteriorate their welfare position when the situation would change from the pre-treaty situation (with each country maximizing its own welfare “autistically”) to the globally optimal contract. These regions are the “other OECD countries,” the “Central and Eastern European countries,” and the countries of the former Soviet Union. For these, it is not individually rational to join a coalition that agrees to realize the globally optimal global warming policy. The results of the stability test are even less encouraging: With the exception of the European Union and Japan, all regions are subject to an incentive to leave a coalition that has agreed upon the globally optimal GHG abatement policy. Of course, there are very many uncertainties regarding the physical and economic aspects of GHG emission and global warming. Therefore, the specific numbers of this study quantifying the costs and benefits of GHG abatement in the 12 regions cannot be really taken for granted (as the authors of this study are well aware of and concede). Still, the study illustrates the fundamental difficulties that impede the international coordination of an effective policy to fight global warming. This study (and other empirical treatments) is able to explain why the *Kyoto Protocol* suffers from so many weaknesses and why it is so difficult to agree upon an effective succeeding treaty.

In the preceding analysis, it has been shown that it is extremely difficult for sovereign countries to agree upon a globally optimally designed global warming treaty and to enforce it. Obviously, the globally optimal contract is a very ambitious goal. Thus, the question arises whether the international legislation process has better expectations if it strives for a somewhat more pragmatic goal. A pragmatic program

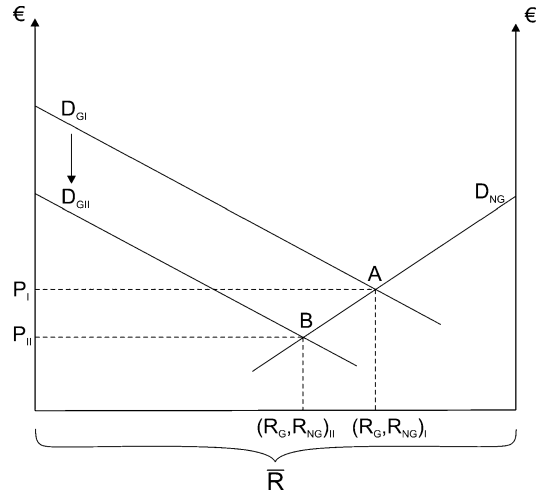
might seek to curb climate change such that the adverse consequences are “tolerable” and allocate the abatement costs among the countries in an “acceptable” way. It has been shown in an extensive literature that this change in the scenario does not lighten up the situation dramatically (See Finus/Caparrós (2015) and Finus/Rundshagen (2015) for fundamental issues and recent developments.). Incentive compatible international global warming agreements are likely to be either deep (ambitious reduction target) or wide (high participation rate of countries) and very unlikely to be both.

Leakage

Let us assume that a country or a group of countries, against all odds, would decide to take serious national measures to reduce GHG emission and then fully enforce these measures.

The country or the group of country could take a number of steps that would aim at reducing energy consumption and especially at replacing fossil energy sources with regenerative ones. However, these activities would result in a lower demand of fossil energy sources and thus, *ceteris paribus*, in a downward pressure on the prices. The decreasing world market prices would stimulate the demand of fossil energy sources in other countries. There, the increasing usage of these energy sources would, *ceteris paribus*, lead to a rise of the relevant emission though. This interdependence is called “leakage effect” in economics. It results in the dissipation of the self-limiting activities of pioneer countries. If this effect is anticipated, even the countries that would agree to limit themselves would be less willing to do so than without the leakage effect.

This reasoning can be illustrated by Fig. 2. The abscissa shows the stock of a fossil energy resource R fixed at a level of $\bar{R}$. There are two groups of countries in the stylized cosmos of this illustration, green countries (G) and nongreen countries (NG). The green countries are the ones mentioned in the introductory paragraph of this section. They take effective GHG curbing measures. In the pre-policy situation (the time before



Global Warming, Fig. 2 Leakage in the Resource Market

the green countries decide and act upon their measures), the demand curves for the fossil energy resource are D_{GI} for the green countries and D_{NG} for the nongreen ones. The market equilibrium price (P_I) and the market equilibrium distribution of the resource stock between the two countries $(R_G, R_{NG})_I$ are defined by the intersection of these two demand curves. The effect of the GHG-fighting policy in the green countries is that their demand for the fossil resource is reduced. In the figure, this is illustrated by the demand curve of the green countries shifting down from D_{GI} to D_{GII} . Accordingly, the market equilibrium shifts from A to B in the figure. In detail, that means that the equilibrium price goes down from P_I to P_{II} and the equilibrium distribution of the resource between the two groups of countries changes from $(R_G, R_{NG})_I$ to $(R_G, R_{NG})_{II}$. In terms of global warming, the result of this analysis is that total consumption of the fossil resource is unaffected by the efforts of the green countries. Thereby, GHG emissions jointly produced with the consumption of this resource are unchanged. The GHG emission “savings” generated by the policy of the green countries are completely used up by the additional consumption of the nongreen countries.

Of course, this is a very stylized reasoning which might not completely cover what is happening in the real world. But, it still illustrates the

“curse of partial solutions” to the global warming problem.

Moreover, the case in which the leakage effect is “transported” via the market for the resource is only one of several forms this detrimental effect might take on. Other forms are discussed in the literature, e.g., in Rauscher/Lünenbürger (2004).

The Dilemma of Intergenerational Allocation

In the preceding analysis, incentive problems of effective global warming policy have been discussed in a framework focusing on the present generation. However, there is also an intergenerational dilemma of global warming policy. GHG reductions conducted in the present are beneficial not only presently but also for future generations. Thereby, an intertemporal externality is generated. The allocative consequences of this externality are analogous to what has been said in the intragenerational context above: Equilibrium GHG reductions of the present generation are too low compared to globally optimal GHG reductions. In the present context, the “global optimality” must be interpreted to comprise the costs and benefits of GHG reduction in all generations. It is plausible without further elaboration that this intertemporal misallocation occurs if the present generation is oblivious with respect to the welfare of future generations. It may come as a surprise, however, that the intergenerational allocation problem does not vanish even if we assume that the preferences of the present generation are not so self-centered. Let us assume the opposite and take the present generation to be *future-altruistic* in the sense that it takes the welfare of future generations into account during its decision-making processes just as it takes its own welfare into account.

To see the point, let us look again at a single country. If a single country trying to optimize the situation for its own citizens and their descendants decides to use less fossil energy, it preserves the energy resources of mankind and causes less CO₂ emissions. However, it cannot be sure that the aforementioned “target group” will benefit from

this saving. It must rather anticipate that the saved resources will be consumed by its contemporaries in the other countries. As far as this does not happen, it is by no means granted that the descendants of the country under consideration will benefit from the savings. Instead, the future generations of another country might consume the resources saved by the country under consideration. By this, the aforementioned leakage effect gets a time dimension. Thus, the incentive to save resources is limited for a country that acts “selectively altruistic” insofar that it only cares for its own children and children’s children. This incentive distortion can be compared to the one, where individual self-restraint collapses due to the exploitation of *open access resources*. So the individual fisherman does not heed the advice he should bear in mind that he wants to go fishing again tomorrow: The decision-maker knows that present consumers (and, if anything is left, future consumers) will profit from his self-restraint and his own provision will suffer. This results in a phenomenon well known from the economics of natural resources, namely, the individually rational, but collectively unreasonable depletion of freely accessible fishing resources.

Now let us assume (neglecting any concerns based on economic theory) that not a single country would act future-altruistically but the worldwide population as a whole would seriously think about it. Unfortunately, even such an assumption cannot solve the intertemporal incentive problem sufficiently. The trouble is that a present generation, even if it comprises the whole earth, can never be sure whether the succession of coming generations might benefit from a better world climate because of their self-restraint regarding the consumption of natural resources. It is already the very next generation that is tempted to exploit the resources the present generation has left them and to waste them regardless of the CO₂ emissions. Since the generations are naturally limited to a certain period of time, there is nothing generation 1 can do to protect their future-altruistic sacrifices from the exploitation of generation 2. This *constitutional nonprotection* of investments intended to support coming generations is a substantial obstacle for making the investments

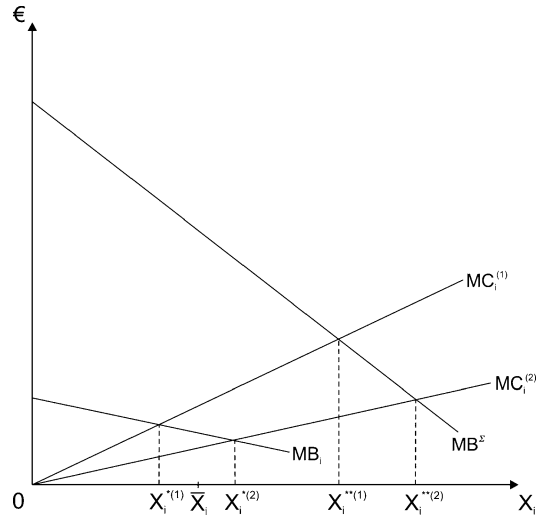
at all. The incentive structure for the present generation is discouraging and can be compared to an investor who thinks about getting economically engaged in a politically unstable country. Since he cannot be sure of how long his investments will be profitable, he is subject to an incentive not to invest at all or only in projects that might be profitable in the short run. Climate protection is a *very* long-term project, though.

The Dismal Science and Its Strive for Climate Change Optimism

Economics is often called the *dismal science*. The pessimistic assessment of the expectations to curb global warming by international GHG policy, as given above, is an obvious tribute to this trademark. However, there are also some strands of economic thinking which might be able to lighten up the picture. Two important examples are *technical change* and *adaptation*.

In the preceding analysis, it has been tacitly assumed that the costs to reduce GHG emissions are given in the sense that they do not change over time. More precisely (and technically speaking), the cost functions mapping alternative quantities of GHG abatement into alternative amounts of cost are invariant. Therefore, in Fig. 1, above, there is only one marginal cost function. This representation is no longer appropriate if we allow for technical change. An important form of this change may be stylized in that with a superior technology more greenhouse gases can be reduced at a given cost than using old-fashioned technology. An equivalent understanding is that a given quantity of greenhouse gases can be reduced at lower cost using the new technology compared to using the old one. In the language of Fig. 1, this kind of technical progress makes the marginal cost curve go down. We illustrate this change in Fig. 3 denoting the marginal cost function, which is valid if the old technology is used, $MC_i^{(1)}$. Accordingly, the cost function illustrating economic circumstances if the new technology is used is called $MC_i^{(2)}$.

It is plausible that technological change making GHG reduction cheaper tilts the scale of



Global Warming, Fig. 3 GHG Abatement and Technical Change

national decision-making into the direction of reducing more greenhouse gases with the new technology than with the old one. In Fig. 3, this is illustrated in that the equilibrium reduction quantity for country i increases from $X_i^{*(1)}$ to $X_i^{*(2)}$. From the perspective of economic theory, this does not reduce the size of the misallocation, since not only the equilibrium quantity increases due to technical change but also the globally optimal quantity. The latter quantity increases from $X_i^{**(1)}$ to $X_i^{**(2)}$. From an ecological perspective, however, technical change is still improving the situation because the higher GHG reduction quantity contributes to mitigate the global warming problem. To illustrate, imagine a GHG abatement target $\bar{X}_i$ which is extra-economically motivated. Then it is possible (and implied in how the level of $\bar{X}_i$ is entered in Fig. 3) that the equilibrium abatement quantity of country i which is valid for the old technology falls short of this extra-economically determined goal, but the equilibrium quantity which holds for the new technology exceeds this goal. Technically speaking, $X_i^{*(1)} < \bar{X}_i < X_i^{*(2)}$ applies.

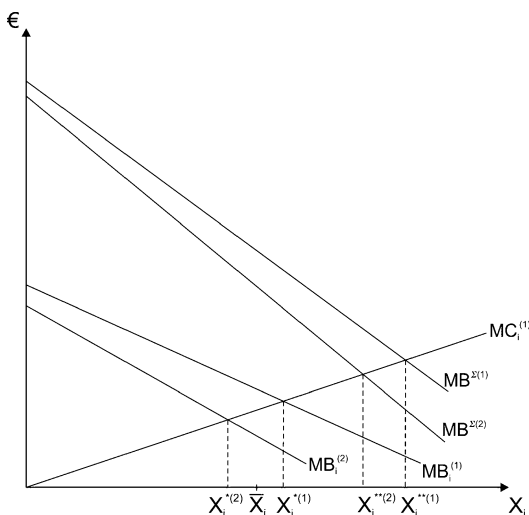
Another important issue modifying the fundamental analysis presented in sections I–V above is greenhouse gas adaptation. The idea is that the damage associated with a given level of GHG

emissions (a given level of average world temperature increase) is not given, as has been tacitly implied in the preceding analysis. To the contrary, human decision-makers might attenuate the detrimental effects of climate change by prudently adapting to this change. One of many examples is agriculture. The detrimental effect of an increase in the average temperature is reduced (and might even turn to be negative) if the portfolio of crops grown in a certain region is “reoptimized” in the sense that it is adjusted to the changed climate circumstances. Obviously, this is a completely different approach to global warming than that which has been followed in the earlier sections of this essay. Instead of trying to put a break onto the extent (and speed) of climate change, the issue of adaptation is to try to cope with given climate change as well as possible. Of course, the two approaches may be combined.

If, by successful adaptation, the damage of increasing global temperature is reduced so is the benefit of GHG abatement. In terms of the graphical illustrations that have been used above, the benefit curves shift downwards. In Fig. 4, the marginal benefit curves indexed “(1)” denote the situation without adaptation, and the curves indexed “(2)” show the benefits which occur if adaptation is applied. As the curves are drawn in Fig. 4, equilibrium GHG abatement for country i

goes down from $X_i^{*(1)}$ to $X_i^{*(2)}$ in the process of adaptation. Globally optimal GHG reduction goes down from $X_i^{**(1)}$ to $X_i^{**(2)}$.

So it is plausible (but very well worth noting because often overlooked in the public discussion) that successful adaptation to climate change attenuates the incentive for GHG mitigation. This is not to be regretted if global welfare maximization is the relevant policy goal. Since adaptation makes global warming less harmful, it is to be welcomed that adaptation tilts the scale of GHG policy decision-making into the direction of less GHG abatement. However, there is a dangerous element in that. This is revealed if we turn from global welfare maximization as a policy goal to the aforementioned goal of achieving an extra-economically motivated standard of GHG abatement. To illustrate this, let $\bar{X}_i$ represent this kind of a standard and assume that $\bar{X}_i$ is “located” between the equilibrium reduction quantity of country i without adaptation, $X_i^{*(1)}$, and the equilibrium quantity with adaptation, $X_i^{*(2)}$. Given these relationships, $X_i^{*(1)} > \bar{X}_i > X_i^{*(2)}$, then adaptation induces policy changes which lead to the result that the extra-economically determined standard is missed. Ironically speaking, one might add: it turns out to be very difficult to take the “dismal” out of the dismal science.



Global Warming, Fig. 4 GHG Mitigation and Adaptation

References

- Endres A (2011) Environmental economics – theory and policy. Cambridge University Press, New York
- Finus M, Caparrós A (eds) (2015) Game theory and international environmental cooperation. Edward Elgar, Cheltenham
- Finus M, Rundhagen B (eds) (2015) Game theory and environmental and resource economics (Part I). Special Issue in Honour of Alfred Endres: Environmental and Resource Economics 62(4), 657–978
- Finus M, van Ierland E, Dellink R (2006) Stability of climate coalitions in a cartel formation game. Econ Gov 7:271–291
- Rauscher M, Lünenbürger B (2004) Leakage. In: Böhringer C, Löschel A (eds) Climate change policy and global trade. Physica-Verlag, Heidelberg/New York, pp 205–230
- Stern N (2007) The economics of climate change: the Stern review. Cambridge University Press, Cambridge
- Stern N (2015) Why are we waiting? The MIT Press, Cambridge, Mass./London

Further Reading

- Barrett S (2003) *Environment and statecraft: the strategy of environmental treaty-making*. Oxford University Press, Oxford and New York
- Cherry TL, Hovi J, McEvoy DM (eds) (2014) *Toward a new climate agreement. Conflict, resolution and governance*. Routledge, Abingdon/New York
- Finus M, Kotsogiannis C, McCorriston S (eds) (2013a) The International dimension of climate change policy. *Special Issue: Environ Resour Econ* 56(2), 151–305
- Finus M, Kotsogiannis C, McCorriston S (eds) (2013b) International coordination on climate policies. *Special Issue: J Environ Econ Manag* 66(2), 159–382
- Nyborg K (2012) *The ethics and politics of environmental cost benefit analysis*. Routledge, Abingdon/New York

Globalization

Danny Cassimon¹, Peter-Jan Engelen^{2,3} and Hanne Van Cappellen⁴

¹Institute of Development Policy and Management (IOB), University of Antwerp, Antwerp, Belgium

²Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

³Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

⁴Institute of Development Policy (IOB), University of Antwerp, Antwerp, Belgium

Definition

Globalization is the ongoing process of increasing interconnectedness between cross-boundary actors, driven by flows of people, ideas, goods, and capital. Globalization reduces the relevance of these national boundaries and stimulates the emergence of complex networks that foster the exchange and integration of technologies, economies, governance, communities, and culture.

Introduction

In this contribution, we first define and conceptualize globalization by distinguishing between de

jure and de facto forms of globalization and by highlighting its main dimensions. From this conceptualization, we then derive and discuss how we can and are measuring globalization, again in its crucial dimensions; using these measures, we analyze and discuss how globalization has evolved over time and across countries or country groupings. We then look at the consequences of globalization and ways to govern it, from the perspective of three complementary perspectives, the market-centered, state-centered, and people-centered perspectives.

Conceptualizing Globalization

De Jure and De Facto Globalization

An important additional distinction to make both conceptually and empirically is the distinction between de jure and de facto globalization (see e.g., Kose et al. 2009). De jure globalization is expressed in the legal sphere and can be measured through the intensity of regulations, restrictions, and controls on, for example, international trade, capital flows, communications, or the ratification of international treaties. De facto globalization on the other hand measures the actual flows of knowledge, ideas, goods, and people. Both concepts are not necessarily high correlates of each other as high de jure globalization does not automatically lead to high de facto globalization. In certain instances, deliberate repression of de jure globalization may still lead to observed de facto globalization, as ways to circumvent legal restrictions can be found. In the same way does an environment with relatively low de jure restrictions not necessarily stimulate high actual flows. It is thus important when researching or analyzing globalization to be explicit about the kind of globalization one is referring to, while acknowledging that combining both views offer a more complete perspective.

Dimensions of Globalization

Next to distinguishing between “written” de jure globalization and “actual” de facto globalization, it is insightful to disaggregate the different dimensions of globalization. One of the most obvious

transboundary connections is through international trade. Together with financial connections such as FDI, these constitute the two main categories of economic globalization.

The second dimension of globalization can be found in the political sphere. Political globalization describes the process of an emerging transnational political system that can be observed by the involvement of nations in international organizations, the adherence to international policy-making or political presence beyond the national borders, such as embassies or international NGOs.

The third dimension of globalization is social globalization. Social globalization describes the interconnectedness of people across national or regional boundaries. This interconnectedness can express itself through the ease of physical movement, both temporary and permanent, but also through the interconnectedness via social and communication media.

The fourth dimension, often incorporated in social globalization, is cultural globalization, which describes the process of increasing transmission and assimilation of values, lifestyle, culture, and consumption across boundaries.

Historically, globalization and the interconnectedness between nations was often characterized by asymmetrical North-South relations, especially in terms of trade and labour. However, South-South globalization is increasingly gaining importance, mainly due to the fast growth of China and India and an increase in regional trade agreements in the Global South (Mansoor Murshed et al. 2011).

Measuring Globalization

The KOF Index of Globalization

The KOF Index is the most often used index of globalization. A detailed review of over 100 empirical studies using the KOF index is given by Potrafke (2015), and a list of studies using the KOF index can be found at <http://globalisation.kof.ethz.ch/papers>. Starting from the premise that globalization is a broad and multifaceted concept, the KOF index incorporates a wide range of variables in its index, divided into

an economic, social, and political dimension (see Table 1). The KOF index was released in 2006 and consequently updated from 2007 on. The latest update, released in January 2018, introduced a considerable amount of innovation. The most important change is the introduction of the distinction between *de jure* and *de facto* globalization, an important distinction that is increasingly studied in the literature (see for example Quinn et al. 2011). The previous versions already used variables that represented a mix of both *de facto* and *de jure* measures, but no explicit distinction was made.

The KOF index is available yearly for the period 1970–2015. The index as well as each subindex is a score ranging from 0 to 100 with 100 indicating the maximum level of globalization, following the procedure of panel normalization. The weights for each variable are calculated using principal component analysis over a period of 10 years, while the weights on the subindices remain the same for the entire panel. Missing values are calculated using linear interpolation or substitution with the closest value. For more details and information on the definition, the dimensions and the calculation of the KOF-index, see Gygli et al. (2018).

De Facto Versus De Jure Globalization

An analysis of the *de jure* side of the index versus its *de facto* counterpart shows that it is indeed an insightful distinction to make. When looking at overall globalization for the entire sample, we see that after moving closely together for the first 20 years of the time series, from the nineties on *de jure* globalization started to increase at a higher rate than *de facto* globalization (Fig. 1, panel a). When looking at the three dimensions separately, we see diverging trends in the *de facto* and *de jure* counterparts (Fig. 1, panel b–d). A common trend in all panels of Fig. 1 is the considerable flattening of globalization in the last decade, possibly due to the lingering effect of the financial crisis. For some subindices, there is even a slight downward trend to be identified, which is typically coined as “*de-globalization*.”

To see that *de facto* and *de jure* concepts are not necessarily high correlates, one can look at the trend of financial globalization (a sub-index of economic globalization, see Table 1) over the

Globalization, Table 1 2018 KOF Globalization Index: Structure, variables and weights. (From The KOF Globalization Index – Revisited, KOF Working Paper, No. 439 by Gygli et al. 2018)

Globalization Index, <i>de facto</i>	Weights	Globalization Index, <i>de jure</i>	Weights
<i>Economic globalization, de facto</i>	33.3	<i>Economic globalization, de jure</i>	33.3
<i>Trade globalization, de facto</i>	50	<i>Trade globalization, de jure</i>	50
Trade in goods	40.9	Trade regulations	32.5
Trade in services	45.0	Trade taxes	34.5
Trade partner diversification	14.1	Tariffs	33.0
<i>Financial globalization, de facto</i>	50	<i>Financial globalization, de jure</i>	50
Foreign direct investment	27.5	Investment restrictions	21.7
Portfolio investment	13.3	Capital account openness 1	39.1
International debt	27.2	Capital account openness 2	39.2
International reserves	2.4		
International income payments	29.6		
<i>Social globalization, de facto</i>	33.3	<i>Social globalization, de jure</i>	33.3
<i>Interpersonal globalization, de facto</i>	33.3	<i>Interpersonal globalization, de jure</i>	33.3
International voice traffic	22.9	Telephone subscriptions	38.2
Transfers	27.6	Freedom to visit	31.2
International tourism	28.1	International airports	30.6
Migration	21.4		
<i>Informational globalization, de facto</i>	33.3	<i>Informational globalization, de jure</i>	33.3
Patent applications	35.1	Television	25.2
International students	31.2	Internet user	31.9
High technology exports	33.7	Press freedom	13.2
Internet bandwidth	29.7		
<i>Cultural globalization, de facto</i>	33.3	<i>Cultural globalization, de jure</i>	33.3
Trade in cultural goods	22.6	Gender parity	31.1
Trademark applications	13.3	Expenditure on education	30.9
Trade in personal services	25.6	Civil freedom	38.0
McDonald's restaurant	23.2		
IKEA stores	15.3		
<i>Political globalization, de facto</i>	33.3	<i>Political globalization, de jure</i>	33.3
Embassies	35.7	International organisations	37.0
UN peace keeping missions	27.3	International treaties	33.0
International NGOs	37.0	Number of partners in investment treaties	30.0

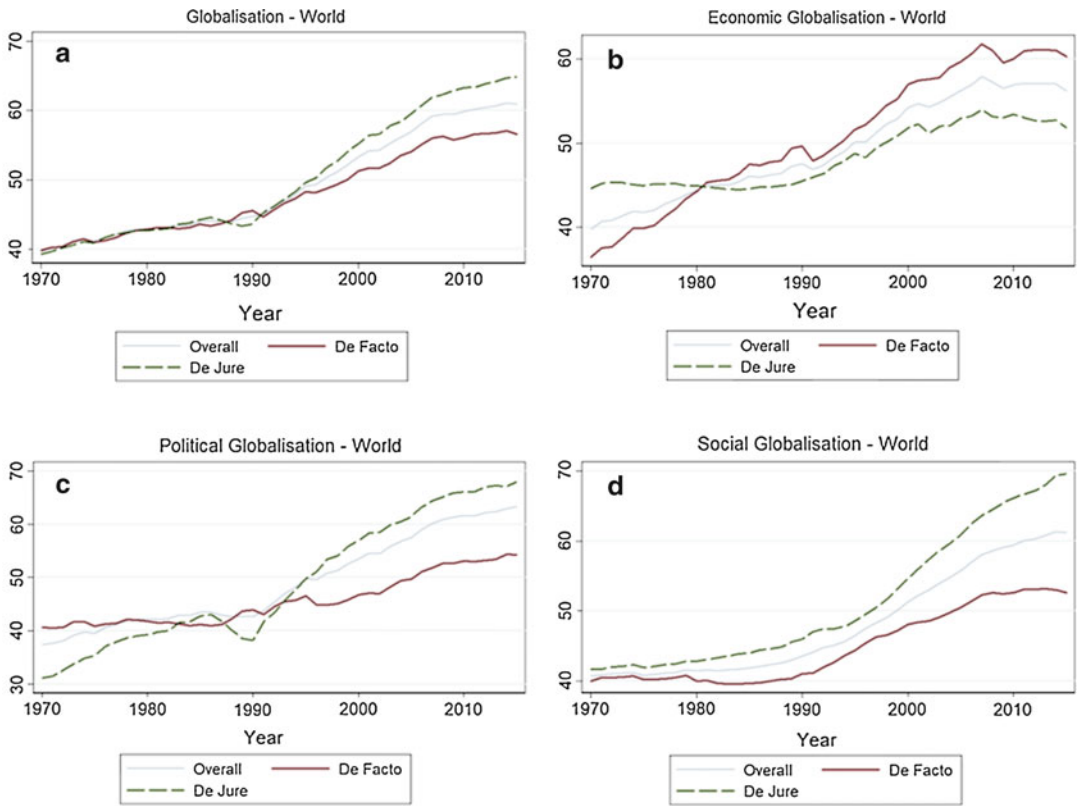
Notes: Weights in percent. Weights for the individual variables are time variant. Depicted are the weights for the year 2015. Overall indices for each aggregation level are calculated by the average of the respective *de facto* and *de jure* indices.

entire period for the entire sample. Figure 2 shows that the *de facto* financial globalization (measured as FDI, portfolio investment and international debt, reserves and income payments) rose at a high rate over the time period, almost doubling from a score of 34 in 1970 to 63 in 2015. The *de jure* counterpart (measuring investment restrictions and capital account openness), however, did not exhibit a clear trend, and even decreased slightly over the period from 48 to 46. This example

illustrates the importance of the choice of certain variables for measuring and analyzing globalization.

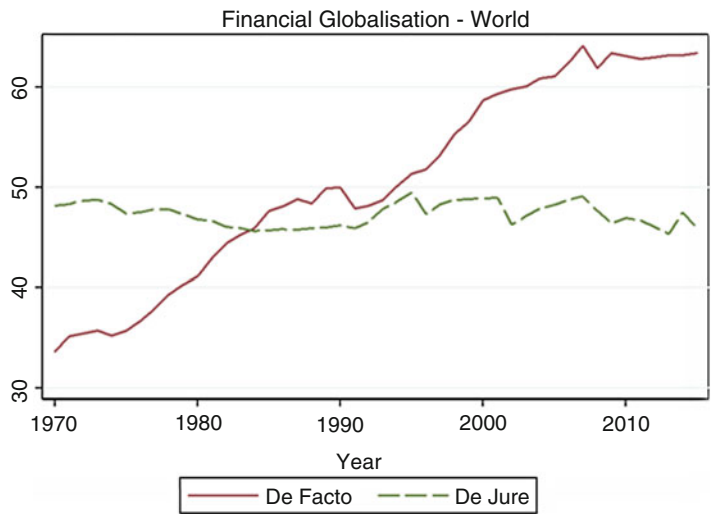
The KOF Index for Different Income Groups

Looking at globalization per income group shows that there is a clear correlation between income level and globalization: the curves retain the same relative position, with the higher income group having a higher globalization score (Fig. 3). This is in contrast with a comparison based on regional



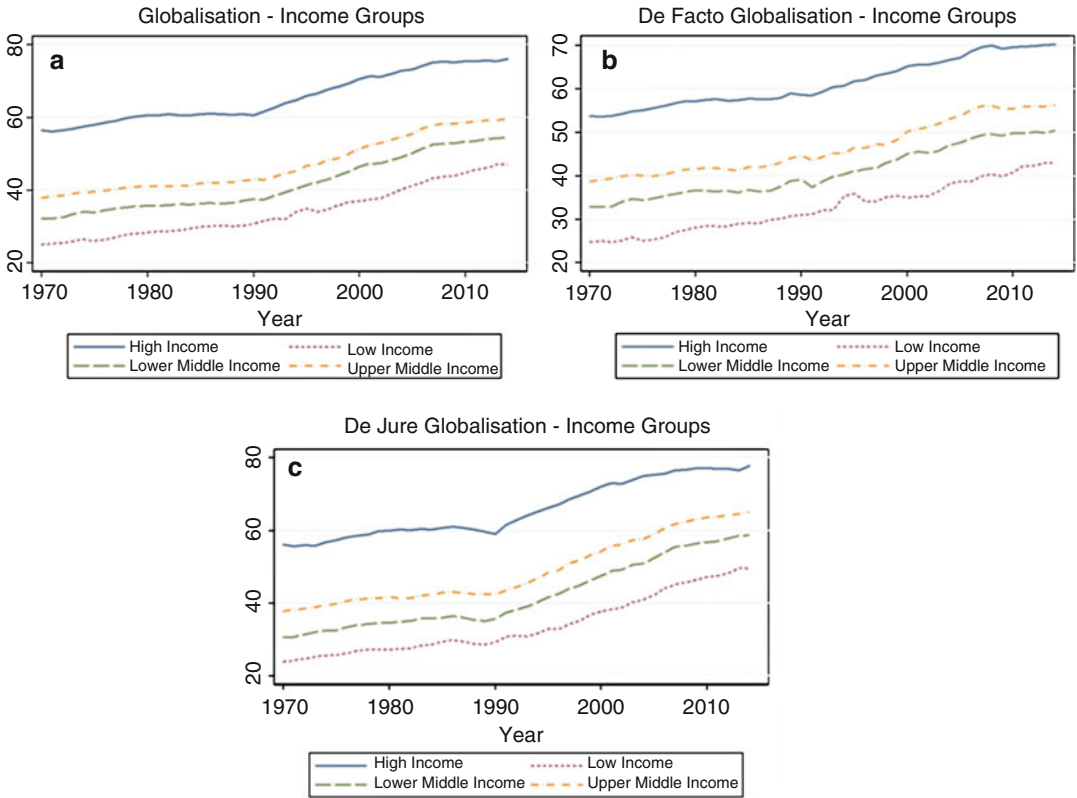
Globalization, Fig. 1 KOF de facto, de jure and overall globalization indices, 1970–2015 – World average

Globalization, Fig. 2 KOF financial de facto and de jure globalization indices, 1970–2015 – World average

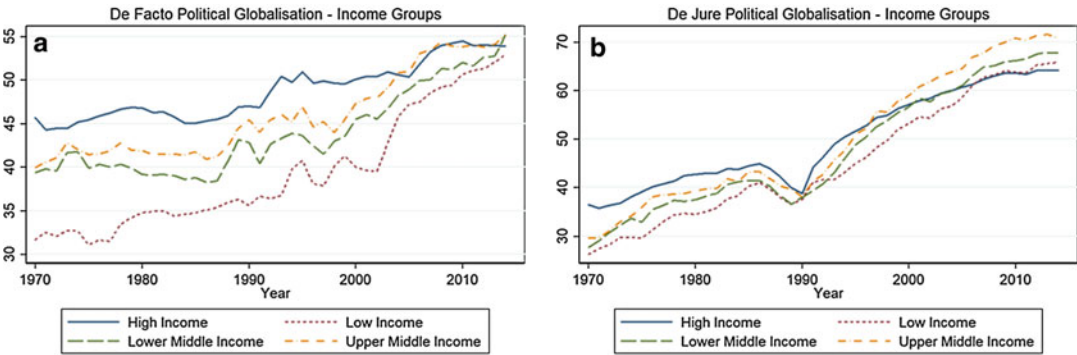


groups, for which the curves often look somewhat messier, and exhibit a less consistent ranking. This is possibly due to the influence of particular

regional events. However, richer regions are often on the higher end of the globalization spectrum.



Globalization, Fig. 3 KOF overall, de facto and de jure globalization indices, 1970–2015 – Income group averages



Globalization, Fig. 4 KOF de facto and de jure globalization indices, 1970–2015 – Income group averages

In Fig. 3, panel a, we see that overall globalisation per income group rose steadily over the observed timespan, with a slight increase in the growth rate starting from the nineties. When looking at the different dimensions of globaliza- tion per income group, we see again that the

subdivision between de jure and de facto is use- ful, as they often paint a somewhat different picture. An interesting example of this is political de facto and de jure globalization (Fig. 4). When looking at de jure political globalization, we see that the different income groups behave

fairly similar, exhibiting a strong growth over the given time period, with the exception of a drop occurring at the end of the eighties but recovering only a few years later. When looking at the *de facto* political globalization however, we see that although convergence is present, the different income groups start from a far more diverging pattern in the beginning of the time period.

Other Measures

A few other measures have been developed to measure globalization. As such there is the CSGR Globalization Index (Lockwood and Redoano 2005), the GlobalIndex (Raab et al. 2008), the Kearney/Foreign Policy index, and the Maastricht Globalization Index (Figge and Martens 2014). These indices as well as the KOF indices, take the national boundaries as its unit of measurement. This is in contrast with the Person-Based Globalization Index (PBGI) proposed by Caselli (2012). A person-based index counters overemphasis of small countries as well as underemphasis of bigger countries. However, it should be noted that weighting by population is more insightful for some dimensions than for others. Social globalization for example might be interesting to measure on a headcount basis, while for political globalization, which depends largely on national policies, this would make less sense.

Perspectives on Globalization and Its Consequences for Governance

In order to describe the consequences of globalization and ways to try to “govern” or manage it, it is useful to distinguish between at least three different perspectives on globalization, or processes that are triggered by it, described by, e.g., Woods (1998, 2000) as market-centered, state-centered, and people-centered.

The “Market-Centered” Perspective of Globalization

The “market-centered” perspective describes a process of intensification of cross-border

exchanges of technology, goods, services, finance, ideas, and ownership, led by multinational enterprises organized in global production and distribution chains, which erodes national markets, making virtually all goods and services becoming essentially “tradables” (as opposite to non-tradables, i.e., goods or services that are only produced and consumed in the local market) and leads to (the “triumph of”) global markets. Essentially, it focuses on the perspective of globalization as leading to the global expansion of capitalism. From this perspective, a quite positive prognosis on the results of globalization is derived: the demolition of *de jure* barriers, the spread of technological advances, combined with efficiency-maximizing behavior of global firms leads not only to long-term production and consumption benefits, and thus higher welfare, but also points at the new opportunities that open up for firms, workers, and consumers all over the world, and particularly also in less developed countries and regions, to gain from being inserted and/or upgraded in these global production chains (Bhagwati 2005; Wolf 2004).

At the same time, while increasing overall welfare, it is clear that this process will produce not only winners but also losers, and that benefits and the costs of globalization are often indeed incurred asymmetrically, necessitating the necessity and ability for complementary measures to be added in an effective way in order to compensate these marginalized and losing ones and try to reintegrate them in a valid and sustainable way in the global capitalist system. In essence, critics in general, and the anti-globalist movement in particular, heavily question the ability of these mitigating effects to be included and, when indeed included, to be effective, and/or, more generally, heavily criticize the ability of global capitalism to provide long-term benefits in a sustainable way (see, e.g., Della Giusta et al. 2006 for a collection of critical writings). A less extreme criticism would suggest positive effects for welfare at modest increases of globalization, that would slowly bottom out at higher levels, and may even become negative at “hyper-globalization” level, hinting that there may something as “too much globalization,” as too much of a good thing (for recent

evidence see, e.g., Lang and Mendes Tavares 2018).

A detailed overview of studies analyzing the effect of globalization focusing mainly on this more economic perspective can be found in Dreher et al. (2008) and, more recently, Potrafke (2015). The reviewed studies measure the influence of globalization on macroeconomic performance such as the size of the government, economic performance, labor markets, and capital markets, to name a few. These analyses are mostly performed on a specific geographical unit and as a consequence findings often differ accordingly. When looking at the world as a whole, the general accepted view is that globalization has a net positive effect (Dreher et al. 2008; Wolf 2004). However, the assessment of the effect of globalization on the world as a whole is ultimately a normative question, as the benefits and the costs of globalization are indeed often incurred asymmetrically. As such, studies like Wade (2004) and Dreher and Gaston (2008) conclude that globalization increases income inequality. Goldin and Reinert (2012) research in depth the complex relationship between poverty and globalization and condition the positive effect of globalization on the presence of compensatory policies or mechanisms that mitigate the effects on the actors on the losing end.

Clearly, this market-centered, somewhat “economistic” perspective is too narrow, as if once market forces are unleashed, institutions and policies will follow (Woods 1998), nation-states will see their role diminish and even evaporate and people will become global consumers, being inserted in global labor chains and sharing global values. In other words, state-centered and people-centered perspectives provide not only reactive forces but also complementary insights.

Adding State-Centered and People-Centered Perspectives of Globalization

A state-centered perspective may acknowledge that indeed, the power of nation-states may reduce in some ways as some policies may become less effective in global markets, or some state capacities, such as the ability to tax, may erode, but several counterbalancing forces occur. First of

all, as highlighted in the *de jure* globalization concept, the global integration process has been driven by explicit state policy choices, especially in Western countries. Secondly, states have been using these policies also to protect their citizens, and particularly also the losers of globalization among them, against the negative effects of globalization. Thirdly, in the absence of global markets’ ability to make good rules and regulations themselves, it requires states to come together to make and enforce these global rules and regulations. As such, rather than eroding state sovereignty, it mainly transforms it, by shifting some power to a supranational level, be it global or regional.

The need for this cooperative behavior at global level has been enhanced by globalization itself, as more of the pressing challenges of the world, such as environmental degradation, crime, and terrorism or “illegal” migration, have become globalized. As such, actions to solve them have become international public goods, necessitating interventions, such as treaties, new agencies, or regulations at regional or global level (e.g., Kaul et al. 2003). In this new globalized world order, it is important to see who exactly determines the rules in these new supranational forms of cooperation, avoiding a race to the bottom where the weakest link dominates the quality of rules. Moreover, it is also important then to monitor the extent to which less-powerful countries are amply represented or able to voice their viewpoints and concerns at the global table; in other words, what is the quality of global governance (see, e.g., Claes and Knutsen 2011).

The focus on these more nation-state centered perspectives has also shown that the evolution towards global markets did not necessarily lead to a globalized consensus on what constitutes good policies. To give one example, the “Washington consensus” has been rather successfully, leading to the gradual acceptance of more heterodox policies, also in emerging and developing countries, and increasingly being rivalled by the Peking one. And more prominently, globalization has not gone hand-in-hand with the worldwide introduction, at nation-state level, of Western-style forms of democracy.

A “people-centered” perspective adds additional insights, complexities, and counteractions, as it does not only impact people’s standards of living but also their values and the opportunities to voice them (see, e.g., Whalley 2008). Again, the proposition that (market-centered) globalization would also lead to the globalization of Western values, and a global culture, has been counteracted by strong upsurges of national and/or religious identity reassertion. It did not lead to a smoother introduction of worldwide human rights values, nor did it realize successful bottom-up promotion of democracy, from the grassroots to nation-state levels. More importantly, a people-centered perspective depicts intensified attention to the asymmetric gains of globalization and the voicing of those left behind in the process. As already stated by Singer and Wildavsky (1993), it is producing a polarized world with “zones of peace” and “zones of turmoil”; these zones of turmoil not only reflect places with a high concentration of losers of globalization in a more narrow sense but also more broadly speaking, zones than may encompass whole regions or countries that are in conflict, or post-conflict, where the state has imploded, and non-state actors have taken over basic service delivery, through all kinds of forms of hybrid governance. Clearly, this will produce increased global inequality. At the same time, increased globalization has also facilitated voicing of concerns and protest against this situation, and also facilitated attempts to escape from these zones of turmoil to the peaceful ones, leading to increased migration moves across the globe, which at the same time start to challenge the very fundamentals of the “market-centered” perspective of globalization. In its most recent form, it is also leading to an increasing wave of forms of populism also in core Western countries such as in Europe and the USA, albeit of different faces, that again threaten to challenge the core fundamentals of market-centered globalization (Rodrik 2017).

This more complex interlinkage between globalization, nation-state power, and democracy is conceptually shown by Rodrik (2011) in his so-called globalization paradox or trilemma, hinting at the fact that we cannot simultaneously

pursue democracy, nation-state self-determination, and hyper-market-centered globalization, advocating for a system of “smart globalization” in which we deliberately limit this market-centered globalization, in order to be able to have more political (nation-state) and social (democratic) support from and for those it is supposed to help.

Cross-References

- [De Jure/De Facto Institutions](#)
- [Economic Development](#)
- [Economic Integration](#)
- [Public Goods](#)

References

- Bhagwati J (2005) In defense of globalisation. Oxford University Press, Oxford
- Caselli M (2012) Trying to measure globalisation, experiences, critical issues and perspectives. Springer Briefs in Political Science. Springer Netherlands, Dordrecht
- Claes DH, Knutsen CH (2011) Governing the global economy. Politics, Institutions and Economic Development. Routledge, London and New York
- Della Giusta M, Kambhampati U, Wade RH (eds) (2006) Critical perspectives on globalization. The globalization of the world economy, vol 17. Edward Elgar, Cheltenham, 656p
- Dreher A, Gaston N (2008) Has globalization increased inequality? *Rev Int Econ* 16(3):516–536
- Dreher A, Gaston N, Martens P (2008) Measuring globalisation: gauging its consequences. Springer Science & Business Media, New York
- Figge L, Martens P (2014) Globalisation continues: the Maastricht globalisation index revisited and updated. *Globalizations* 11(6):875–893
- Goldin I, Reinert K (2012) Globalization for development. Meeting new challenges. Oxford University Press, Oxford, 337p
- Gygli S, Haelg F, Sturm J-E (2018) The KOF globalisation index – revisited, KOF working paper no. 439
- Kaul I, Conceição P, Le Goulven K, Mendoza R (eds) (2003) Providing global public goods. Managing globalization. UNDP, New York, 646p
- Kose A, Prasad E, Rogoff K, Wei SJ (2009) Financial globalization: a reappraisal. *IMF Staff Papers* 56(1), pp 8–62
- Lang V, Mendes Tavares M (2018) The Distribution of gains from globalization, IMF working paper WP/18/54
- Lockwood B, Redoano M (2005) The CSGR globalisation index: an introductory guide. Centre for the Study of Globalisation and Regionalisation working paper 155(04), pp 185–205

- Mansoor Murshed S, Goulart P, Serino L (eds) (2011) South-south globalization. Challenges and opportunities for development. Routledge studies in development economics, vol 90. Routledge, London/New York, 348p
- Potrafke N (2015) The evidence on globalisation. *World Econ* 38:509–552
- Quinn DP, Schindler M, Toyoda MA (2011) Assessing measures of financial openness and integration. *IMF Econ Rev* 59(3):488–522
- Raab M, Ruland M, Schönberger B, Blossfeld HP, Hofäcker D, Buchholz S, Schmelzer P (2008) GlobalIndex: A sociological approach to globalization measurement. *Int Sociol* 23(4):596–631
- Rodrik D (2011) The globalization paradox: democracy and the future of the world economy. W. W. Norton, New York, 346p
- Rodrik D (2017) Populism and the economics of globalization, NBER working paper no. 23559
- Singer M, Wildavsky A (1993) The real world order: zones of peace, zones of turmoil. Chatham House, Chatham
- Wade RH (2004) Is globalization reducing poverty and inequality? *World Dev* 32(4):567–589
- Whalley J (2008) Globalisation and values. *World Econ* 31(11):1503–1524
- Wolf M (2004) Why globalization works. Yale University Press, New Haven, 398p
- Woods N (1998) Editorial introduction. *Globalization: Definitions, debates and implications*. Oxford Development Studies, 26(1):5–13
- Woods N (2000) The political economy of globalization. In: Woods N (ed) *The political economy of globalization*. MacMillan Press, Houndmills, pp 1–19

Golden Age Piracy

► Piracy, Old Maritime

Good Faith

Hans-Bernd Schäfer^{1,2} and Hüseyin Can Aksoy³

¹Bucerius Law School, Hamburg, Germany

²Faculty of Law, St. Gallen University, St. Gallen, Switzerland

³Faculty of Law, Bilkent University, Bilkent, Ankara, Turkey

We thank Anne van Aaken, Hein Kötz, Alan Miller, Luciano Benedetti Timm, and Reinhard Zimmermann for valuable comments on an earlier draft. All errors are ours.

Abstract

Good faith is a blanket clause under which courts develop standards of fair and honest behavior. It gives ample discretion to the judiciary and entitles a court to narrow down the interpretation of statutes or contracts and even to deviate from codified rules, from the wording in the law or the contract or to fill gaps. The law and economics literature relates bad faith to opportunistic behavior and it is accepted that in specific cases where the application of default or mandatory rules leads to opportunism or where both the law and the contract are silent on risk, the judge can resort to the good faith principle. As a result the parties may delegate part of the contractual drafting to the legal system in addition to having reduced apprehension regarding the possibility of opportunistic behavior from the other side. This allows them to keep contracts relatively short and reduces the costs of defensive strategies.

Definition

Good faith is a blanket clause in civil law under which courts develop standards of fair and honest behavior in the law of obligations, especially in contract law and – to some extent – in the law of property. Courts also define legal consequences resulting from the violations of such standards, such as reliance damages, expectation damages, injunction, imposing a duty of conduct, and invalidation or validation of a contract. The good faith principle entitles a court to narrow down the interpretation of statutes or contracts and even to deviate from codified rules, from the wording in the law or the contract or to fill gaps. Courts use it as a last resort, if and only if the contract itself or the rules of contract law lead to grossly unfair results. “Good faith” also has found its way into jurisdictions other than those of civil law, as well as into the public international law of contracts and international law in general.

Introduction

The roots of the good faith principle can be traced back to Roman times. Even today Cicero’s writing

on good faith (De Officiis 3.17 (44 BC)) from more than 2,000 years ago reads as fresh as if it had just been published. It comes close to the modern description of “opportunistic behavior” that economists have developed in recent decades: “These words, good faith, have a very broad meaning. They express all the honest sentiments of a good conscience, without requiring a scrupulousness which would turn selflessness into sacrifice; the law banishes from contracts ruses and clever maneuvers, dishonest dealings, fraudulent calculations, dissimulations and perfidious simulations, and malice, which under the guise of prudence and skill, takes advantage of credulity, simplicity and ignorance” (quoted from Association Henri Capitant and Société de législation comparée 2008). The contemporary content of the principle is not well defined. Its scope is also subject to debate, with some scholars regarding it as a rule of contract law or of the law of obligations and with others relating it to the whole body of civil law.

Objective good faith is an objective standard of behavior, especially for contracting parties. Contract law, with its mandatory and default rules, tries to allocate risk and imposes contractual, pre-contractual, and post-contractual duties and risks fairly and honestly in the way self-interested parties would have chosen had they considered the problem *ex ante* at the time of contract formation. However, due to the features of a specific case, the application of these legal rules can possess unintended and absurd consequences in individual cases. And even though the black letter rule mostly leads to the desired consequences, it might lead to the opposite in some. In such cases, the good faith principle allows courts to deviate from the black letter rules of contract law or from the contract itself. If the courts find the application and the result of a legal norm in a specific case to be absurd, because they still allow opportunistic exploitation or grossly inefficient risk allocation, they can deviate by using the good faith principle. And they can give the law flexibility if unforeseen contingencies transform fairness into dishonesty and efficient into grossly wasteful allocation of risk. This not only provides adaptability to new circumstances but also saves the parties’ transactions costs.

More specifically, by knowing that a court has the option to intervene, parties are less inclined to write every contingency into the contract, thus making contracting less costly and defensive measures of parties less necessary.

The principle is not without pitfalls and therefore subject to criticism. Firstly, it gives ample discretion to the judiciary, which can be and has been misused for importing ideology, including totalitarian ideology, into contract law or for promoting private opinions of judges about what they regard as just. Secondly, it might foster judicial activism if the judiciary develops the law proactively in such a way that it replaces parliament. Thirdly, the principle might be inappropriately used to redistribute wealth from the rich to the poor party by adopting a deep pocket approach. Finally, following Hayek, judges who intervene in a contract might not as outside observers have the information for reliably acting in the *ex ante* interest of honest parties, even if they have the best intention in doing so.

The Good Faith Principle in the Legal Doctrine

The good faith principle is widely used in all civil law countries, for instance, in Germany, Switzerland, Turkey, France, Italy, and the Netherlands. It is also recognized in the US Uniform Commercial Code as well as in unification instruments of private law, including the UN Sales Law, the UNIDROIT Principles for Commercial Contracts, and the Principles of European Contract Law. Good faith is relatively new in American law, but seemingly it also caught on quickly. On the other hand, English law does not recognize a general duty of good faith. In fact, English lawyers think that the courts should only interpret contracts and refrain from allocating risks by way of the good faith principle (Musy 2000; Goode 1992). This does not imply that English contract law cannot correct for absurd consequences of routine decisions, yet the English principles, such as “fair dealing,” are narrower than the good faith principle and do not transfer the same amount of discretionary power to the judiciary. Scholars of comparative law have expressed the view that if one looks at how hard

cases of contract law are actually decided in common law and civil law jurisdictions, the differences might be less pronounced than it seems on the surface (Zimmermann and Whittaker 2000). Judge Bingham expressed a similar view with the following words in *Interfoto Picture Library Ltd. v. Stiletto Ltd.* (1988) 2 W.L.R. 615 (p. 621): (England has no) “overriding principle that in making and carrying out contracts parties should act in good faith ” but added that on the other hand “English law has, characteristically, committed itself to no such principle but has developed piecemeal solutions in response to demonstrated problems of unfairness.”

A major point of critique of the principle of good faith is its generality and broad scope. This is also in close relationship with the critique that the judiciary can arbitrarily interfere with the contract by using this principle. However, it could be argued that the risk of intolerable legal insecurity and unpredictability is eliminated. In those jurisdictions where the principle is accorded an important role, it is split up into categories and subcategories (Hausheer and Jaun 2003; Grüneberg 2010). Through this subcategorization, an out-of-hand use of the principle by the courts in these jurisdictions is seldom, with each subcategory applying the facts of the case to a series of judge-made tests that possess particular legal consequences on passing the test.

In jurisdictions that recognize a general principle of good faith, there exists a variety of established legal dogmatic forms that define terms and conditions under which the principle of good faith is to be used. Moreover, they also define particular legal consequences if a defendant has violated a standard of behavior. These legal dogmatic forms, all of which are derived from the general good faith principle, provide it with an internal structure, which checks it against willful use. These subcategories are mainly as follows: culpa in contrahendo; contract with protective effects for a third party; liability for breach of trust; adaptation of the contract to changed circumstances (*clausula rebus sic stantibus*); side obligations from a contract (breach of which constitutes a case of breach of contract); obligation to contract; principle of trust in formation,

interpretation, and gap filling of legal transactions; misuse (abuse) of rights; and forfeiture.

The well-established subcategories of good faith substantially eliminate the risk of a judge’s arbitrary interference. As a matter of fact, there is a general scholarly agreement on the conditions for resorting to any of the subcategories as well as on the legal results that follow. Still, despite such concrete dogmatic forms, in the jurisdictions where the good faith principle is recognized, it is accepted that such general principles should be used as a last resort, if and only if the formal rules of the contract law lead to absurd consequences. If the good faith principle is a monster, as scholars (Zimmermann and Whittaker 2000) once claimed, it has been domesticated as a farm animal. From an economic perspective, this means two things. First, the good faith principle is a delegation norm for the judiciary. It entitles judges to manually control the law if it does not fulfill its functions. It serves parties if the judges are knowledgeable, loyal, and not corrupt but does a disfavor to them otherwise. Secondly, the development of a tight internal structure is a commitment by the judiciary to credibly signal to the legal community that it applies the principle with self-restraint and self-discipline, ruling out willful, ideological, or biased decisions.

The Good Faith Principle as a Norm to Curb Opportunism

The law and economics literature mainly relates the good faith principle to the prevention of opportunism. According to Summers (1968), good faith is an excluder which “has no general meaning or meanings of its own, but which serves to exclude many heterogeneous forms of bad faith.” Burton (1980) relates bad faith to the exercise of discretion by one of the contractual parties concerning aspects of the contract, such as quantity, price, or time. Accordingly, “Bad faith performance occurs precisely when discretion is used to recapture opportunities forgone upon contracting – when the discretion-exercising party refuses to pay the expected cost of performance.” Similarly, while defining good faith,

Miller and Perry (2013) refer to the enforcement of the parties' actual agreement. More specifically, the authors argue that the good faith principle protects the reasonable expectations of the parties, which they had while contracting. They criticize however that in the USA in contrast to definitions, courts often resort to community standards, based on what people actually do, and that this can undermine the function of contract. They argue that the good faith principle in US contract law should be based on normative ideas like welfare maximization or the golden rule and not on purely empirical observations. Mackaay (2009) defines bad faith as the legal term for opportunism. Richard Posner describes the principle in a similar way: "The concept of the duty of good faith like the concept of fiduciary duty is a stab at approximating the terms the parties would have negotiated had they foreseen the circumstances that have given rise to their dispute. The parties want to minimize the costs of performance. To the extent that a doctrine of good faith designed to do this by reducing defensive expenditures is a reasonable measure to this end, interpolating it into the contract advances the parties' joint goal" (Justice Posner's opinion from the decision *Market Street Associates Limited Partnership v. Frey*, 941 F.2d 588, 595).

Elements of Opportunistic Behavior

Williamson (1975) famously defines "opportunism" as "self-interest seeking with guile." In fact, strategic manipulation of information and misrepresentations are to be deemed as opportunistic. The same applies to hidden action aimed at reducing the quality of performance.

According to Dixit, opportunistic behavior is the "whole class of actions that tempt individuals but hurt the group as a whole" (Dixit 2004). Opportunistic behavior is inefficient because it increases transaction costs and reduces the net gain from the contract (Sepe 2010; Mackaay 2011). The risk of opportunism encourages parties to take defensive precautions and write longer contracts to prevent opportunistic behavior as well as the harms that might arise from it. Opportunistic behavior could even incentivize parties to take strict precautions, such as foregoing

a contemplated contract altogether. If many contractors were to choose this option, this could destroy entire markets (Mackaay 2011).

Muris (1981) defines opportunism as behavior which is contrary to the other party's understanding of their contract – even if not necessarily against the implied terms of the contract – and which leads to a wealth transfer from the other party to the performer. Cohen (1992) defines it as "any contractual conduct by one party contrary to the other party's reasonable expectations based on the parties' agreement, contractual norms, or conventional morality." Similarly, Posner (2007) defines opportunism as taking advantage of the other party's vulnerability.

Mackaay and Leblanc (2003), who regard good faith as the opposite of opportunism, propose a test to operationalize opportunistic behavior: "an asymmetry between the parties; which one of them seeks to exploit to the detriment of the other in order to draw an undue advantage from it; the exploitation being sufficiently serious that, in the absence of a sanction, the victim and others like him or her are likely substantially to increase measures of self-protection before entering into a contract in the future, thereby reducing the overall level of contracting."

The Necessity to Curb Opportunistic Behavior with a Blanket Clause

The role of contract law can be explained as an endeavor to guarantee the fairness of the contract in the sense of avoiding opportunistic behavior (Posner 2007; Kaplow and Shavell 2002) and through the cost-efficient allocation of risk (Harrison 1995; Schwartz 2003; Schäfer and Ott 2004). In accordance with the purpose of contract law, if the good faith principle were used exclusively to curb opportunistic behavior and to allocate risks in a cost-efficient way in cases in which the established rules of contract law permitted opportunism, it would be regarded as a valuable corrective mechanism giving flexibility and innovativeness to contract law without questioning its rationale for facilitating win-win constellations under fair conditions.

A court can enhance efficiency through the good faith principle in three ways: by (i) not applying a mandatory rule (for instance, by

declaring a non-notarized real estate sales contract as valid), (ii) refraining from the application of a default rule (for instance, by declaring the rule under which partial delivery can be rejected as invalid), and (iii) allocating risks when the law or the contract is silent (for instance, by imposing a non-competition clause) (Schäfer and Aksoy 2015). In the first two cases, the law specifies the risk, but it is clear that the specification of the risk in such a particular case is questionable and gives rise to opportunism. In other words: in such a specific case, applying the norm delivers absurd results, with the blanket clause of the good faith principle on the other hand providing the judge with flexibility that he would otherwise not have had. Finally, both contract and default rules may be silent on a matter consequentially leading to a result that may be neither fair nor cost-saving. In such cases, the principle of good faith can provide efficient risk allocation.

Finally, in line with the legal reasoning, Mackaay (2011) argues that good faith is a last resort tool for preventing opportunism. The author states that there are various anti-opportunism concepts in the law, but when none of these concepts manage to prevent opportunism, the judge will use the good faith principle. Economically speaking, the blanket clause of good faith is a delegation of decision-making authority to the judiciary. If contract law fails to fulfill its aim of curbing opportunistic behavior and allocating risk in a cost-efficient way, thus to guaranteeing fairness and honesty, the judiciary is given the competency for guaranteeing the rationale of contract law, even if this implies deviation from legal rules or from the contract itself. It is obvious that such a delegation norm can work only in jurisdictions in which judges are loyal to the spirit of the law, as well as knowledgeable and non-corrupt.

Good Faith and the “Social Function of Contract”

The self-restriction employed by courts when using the good faith principle has contributed significantly to its international recognition, expansion, and widespread use in many countries and codes and has helped to silence criticism. It is used to maintain the fairness of the contract. It is

employed as a last resort; it has developed a rich internal structure, thus preventing willful use. Under these conditions, the good faith principle saves transactions costs, and the business community as well as the whole legal community can regard it as a valuable service.

This proposition comes, however, with a caveat. In some civil law countries, including those of continental Europe, it has been argued that courts have sometimes used the good faith principle to achieve a proper distribution of wealth by interpreting the law or filling gaps in contracts in favor of parties such as consumers, employees, and tenants. Sometimes, such decisions can be interpreted as a correction of market failures, but sometimes they are indicative of a deep pocket approach, meaning redistribution of wealth to the poorer party, even if this neither corrects unfairness nor market failure and is contrary to the terms of the contract itself and to black letter law. This is motivated by distributive justice but comes at the cost of protecting only the insiders not the outsiders and increasing rather than decreasing the costs of using the market making it difficult or even impossible for people without reputation to make promises.

In some countries, the principle is openly used to promote the redistributive goals of social justice, which are usually achieved through the welfare state and statutory laws. In Brazil a case law is emerging in which the good faith principle is not merely used to maintain fairness in the sense of honesty. Under the legal principle of the “social function of contract” the good faith principle sometimes helps to achieve social justice by redistributing income *ex post*. This can destroy the mutual benefit from a fair contract in which all partners abstain from opportunistic behavior. It seems that such features of “good faith” can be found more often in countries without an effective welfare system or public and private insurance systems. If contracting parties must, however, expect that under certain conditions, in spite of their fairness and honesty, they are stripped from the benefits of the contract *ex post* by the use of the good faith principle, they might abstain from such contracts altogether. This can lead to dysfunctional markets and considerable costs for

society. The unintended consequences of such interventions, for instance, price ceilings, have been widely discussed in the literature.

Timm (2008) has reviewed some of the Brazilian case law, which spans from not allowing interest on interest, imposing interest rates *ex post* which are lower than those for government bonds or an injunction requiring an insurance company to cover a surgery or treatment not provided according to the policy, or even preventing an unpaid utility from cutting the supply of electricity. In a lawsuit initiated in Rio de Janeiro, a tenant failed to pay rent for consecutive months. Hence, the landlord sued the tenant bringing a claim for eviction from the property, which the tenant ran as an asylum for the elderly. Despite the Brazilian Landlord-Tenant Law (Law No. 12112/2009) which set forth that a failure to pay rent for consecutive months is a valid reason for eviction, the Appeal Court of Rio de Janeiro ruled for the prolongation of the deadline for an indefinite period of time in order to protect the elderly residents, thus making the landlord a charitable donator. Such decisions help the individual claimant, but not the poor as a group as they deny them the possibility to make credible promises and commitments making market transactions more difficult for them.

According to Nóbrega, the new 1988 Constitution brought about judicial activism by way of establishing a principle- and standard-oriented “social welfare legal order in which the judge had the mission of maintaining social justice (Nóbrega 2012). Consequently, the Brazilian Civil Code regulated both the good faith principle and the social function of contract. Within this scope, the social function of contract, which is derived from the principle of good faith, is regulated by Article 421 of the Brazilian Civil Code as follows: “the freedom to contract shall be exercised by virtue, and within the limits, of the social function of contracts.” Although the law refrains from defining the “social function of contract,” a prevailing definition of this broad concept is made by Diniz (2007) as a kind of contractual “super-principle,” comprising precepts of public order, good customs, objective good faith, contractual equilibrium, solidarity, distributive

justice, etc. More specifically, the author argues that the social function of the contract should comprise every constitutionally and/or legally recognized value that might be said to have a “collective” or “nonindividualistic” character.

Conclusion

The law and economics literature relates bad faith to opportunistic behavior, which increases transaction costs, reduces the net gain from the contract, and allows one party to achieve unfair re-distributional gains. Contract law is mainly the endeavor to curb opportunistic behavior and to maintain the cost-efficient allocation of risk. In specific cases where the application of default or mandatory rules leads to opportunism or where both the law and the contract are silent on risk, the judge can resort to the good faith principle. Its use allows for the correction of opportunistic behavior for all those cases in which the black letter law fails to do so. It provides ample discretion to the judiciary and has – from a historical perspective – been misused and heavily criticized. This has not, however, prevented it from becoming an important legal norm in a rising number of jurisdictions and codes.

The good faith principle lost its fearsome features mainly for three reasons. First, its contemporary use is constrained to preserve fairness and honesty among parties, but does not burden honest parties with obligations from community values or social justice in the sense of redistributing income *ex post* for social purposes. Some exceptions exist however: sometimes camouflaged, sometimes – like in Brazil – easily visible, and more explicit. Secondly, courts use the good faith principle as a last resort when all other routines of judicial decision-making are exhausted yet still lead to absurd legal consequences that allow for opportunistic exploitation or grossly inefficient risk allocation. Finally, the good faith principle has a deep interior structure with subcategories providing routines, tests, and legal consequences resulting from these tests. This makes it difficult for judges to misuse the principle for promoting private or ideological

views about justice and becoming disloyal to the law. Decisions of an individual judge using the good faith principle without observing these self imposed dogmatic constraints are likely to be uplifted by a higher court. These features have led to the increasing confidence that the good faith principle provides a valuable service to parties and enables them to delegate part of the contractual drafting to the legal system. Moreover, it has reduced their apprehension regarding the possibility of opportunistic behavior from the other side. This allows them to keep contracts relatively short and reduces the costs of defensive strategies. In comparison, in jurisdictions which either reject the principle (England) or in which the principle has no long-lasting tradition (USA), there is less delegation and contracts are extensive, are more expensive to draft, and contain long laundry lists of required or impermissible behavior from parties. In the literature, other reasons have been discussed as to why English and US private contract law implies less delegation to the judiciary. This is only one of them.

References

- Association Henri Capitant and Société de législation comparée (2008) *European contract law – materials for a common frame of reference: terminology, guiding principles, model rules*. Sellier European Law Publishers, Munich
- Burton SJ (1980) Breach of contract and the common law duty to perform in good faith. *Harv Law Rev* 94(2):369–404
- Cohen GM (1992) The negligence-opportunism tradeoff in contract law. *Hofstra Law Rev* 20(4):941–1016
- Diniz MH (2007) *Curso de direito civil brasileiro*, vol 3. Teoria das obrigações contratuais e extra-contratuais, 23 edn. Revista e atualizada de acordo com a Reforma do CPC. Saraiva, São Paulo
- Dixit AK (2004) *Lawlessness and economics – alternative modes of governance*. Princeton University Press, Princeton
- Goode R (1992) The concept of “good faith” in English law. <http://www.cisg.law.pace.edu/cisg/biblio/goode1.html>. Accessed 18 Feb 2014
- Grüneberg C (2010) § 241. In: Palandt O (ed) *Bürgerliches Gesetzbuch*. Verlag C.H. Beck, München
- Harrison JL (1995) *Law and economics*. West Publishing, St. Paul
- Hausheer H, Jaun M (2003) Die Einleitungsartikel des ZGB – Art. 1-10 ZGB. *Stämfli*, Bern
- Kaplow L, Shavell S (2002) *Economic analysis of law*. In: Auerbach AJ, Feldstein M (eds) *Handbook of public economics*, 1661–1784. Elsevier, Amsterdam
- Mackaay E (2009) The economics of civil law contract and of good faith. https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/3016/Mackaay_Trebilcock-Symposium%20_3_.pdf?sequence=1. Accessed 18 Feb 2014
- Mackaay E (2011) Good faith in civil law systems – a legal-economic analysis. CIRANO – Scientific Publications 2011s-74. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1998924. Accessed 18 Feb 2014
- Mackaay E, Leblanc V (2003) The law and economics of good faith in the civil law of contract. <https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/125/Article%20papyrus.pdf?sequence=1>. Accessed 18 Feb 2014
- Miller AD, Perry R (2013) Good faith performance. *Iowa Law Rev* 98(2):689–745
- Muris TJ (1981) Opportunistic behavior and the law of contracts. *Minn Law Rev* 65(4):521–590
- Musy AM (2000) The good faith principle in contract law and the precontractual duty to disclose: comparative analysis of new differences in legal cultures. <http://www.icer.it/docs/wp2000/Musy192000.pdf>. Accessed 18 Feb 2014
- Nóbrega FFB (2012) Custos e Benefícios de um Sistema Jurídico baseado em Standards: uma análise econômica da boa-fé objetiva. *Econ Anal Law Rev* 3(2):170–188
- Posner RA (2007) *Economic analysis of law*. Wolters Kluwer, Austin
- Schäfer HB, Aksoy HC (2015) Alive and well, the good faith principle in Turkish contract law. *Eur J Law Econ* (in print), <https://doi.org/10.1007/s10657-015-9492-1>
- Schäfer HB, Ott C (2004) *The economic analysis of civil law*. Edward Elgar, Cheltenham
- Schwartz A (2003) The law and economics approach to contract theory. In: Wittman DE (ed) *Economic analysis of the law*. Blackwell, Maiden
- Sepe SM (2010) Good faith and contract interpretation: a law and economics perspective. *Arizona Legal Studies*, discussion paper no. 10-28. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1086323. Accessed 18 Feb 2014
- Summers RS (1968) “Good faith” in general contract law and the sales provisions of the uniform commercial code. *Virginia Law Rev* 54(2):195–267
- Timm LB (2008) Funcao Social Do Direito Contractual No Código Civil Brasileiro Justica Distributiva vs. Eficiência Economia. *Rev Tribunais*, 97, 876:11–43
- Williamson OE (1975) *Markets and hierarchies: analysis and antitrust implications*. Free Press, New York
- Zimmermann R, Whittaker S (2000) *Good faith in European contract law*. Cambridge University Press, Cambridge

Further Reading

- Bridge MG (1984) Does Anglo-Canadian contract law need a doctrine of good faith? *Can Bus Law J* 9(4):385–426

- O'Connor JF (1990) Good faith in English law. *Darmouth, Aldershot*
- Piers M (2011) Good faith in English law – could a rule become a principle? *Tulane Eur Civ Law Forum* 26(1):123–169
- Teubner G (1998) Legal irritants: good faith in British law or how unifying law ends up in new divergences. *Modern Law Rev* 61(1):11–32
- Whittaker S (2013) Good faith, implied terms and commercial contracts. *Law Q Rev* 129(3):451–469

Good Faith and Game Theory

Caspar Rose
Copenhagen Business School, Copenhagen,
Denmark

Abstract

This article shows how game theory can be applied to model good faith mathematically using an example of a classic legal dispute related to *rei vindicatio*. The issue is whether an owner has a legal right to his good if a person has bought it in good faith by using updated probabilities. The article illustrates that a rule of where good faith is irrelevant Pareto dominates a rule where good faith may protect an innocent buyer.

Synonyms

Using updated beliefs in modeling good faith and its impact on Pareto efficiency

Definition

If a potential buyer incurs effort and the new updated beliefs result in an increased probability that the good is stolen compared to the initial beliefs, then a buyer is said to be in *bad faith*; if, however, the probability decreases, he is said to be in *good faith*. The buyer is also in bad faith, if he deliberately chooses to stay in ignorance.

The Advantage of Using Game Theory

One may ask, what are the benefits from using game theory for modeling good faith? Game theory relies on a number of assumptions, in particular, the full rationality by the players, which means that the players' objective is to maximize their expected utility; see, e.g., Rasmussen (2004). This means that players make rational choices and have stable preferences.

Game theory has the advantage that it takes into account the players strategic interdependency meaning that the players utility does not only depend on their own actions but also on the opponents actions and vice versa or mathematically expressed: $U_i(a_i, a_j)$ and $U_j(a_j, a_i)$.

Game theory is well suited for analyzing different legal settings, in particular, the incentive effects of different legal rules. Players know what kind of actions is available for them, which naturally is influenced by the legal status, e.g., the possible choice of remedies of breach in a contractual relationship. The players know what the expected outcomes from a sequence of actions are. This is presupposed to be common knowledge among the players. This does not mean that players have perfect information. However, by applying the proper equilibrium concept, one may be able to predict the player's actions, i.e., the outcome of the game. The basic solution concept is the Nash equilibrium where each player plays a best response to each other's actions. This must be refined depending on whether the game is static/dynamic with perfect or imperfect information. In the following analysis, the applied solution concept is the so-called subgame perfect Nash equilibrium which is used for dynamic games with perfect information, but it can include external uncertainty (moves by Nature). In short, game theory is a powerful tool to study games with strategic interdependency where the legal setup defines the player's action spaces; see, e.g., Landes and Posner (1996) for a study of legal rules and strategic interdependency as well as Baird et al. (1998) for an excellent example of how legal rules can be studied using game theory.

The Usefulness of Using Game Theory in Law and Economics

Good faith has been analyzed using game theory, but it turns out that game theory is very well suited for analyzing the economic effects of legal rules. The basic setup of using game theory in law and economics is to study different games under different legal regimes with the same players. To illustrate, it is well known that car drivers and pedestrians behave differently under a regime of strict liability or negligence. Specifically, a car driver may show less precaution under a negligence rule, whereas the pedestrians may be quite careful walking on the sidewalk. In short, law and economics that rely on game theory may be formulated in the following single equation.

$$\Delta \text{Rules/court rulings} \Rightarrow \Delta \text{actions} \\ \Rightarrow \Delta \text{payoffs} \Rightarrow \Delta \text{incentives} \Rightarrow \Delta \text{efficiency}$$

The equation simply displays causal impact of changing the rules or court rulings on efficiency, where Kaldor-Hicks efficiency may be preferred for practical reasons in most cases. Most game theories assume that players are risk neutral, but this assumption may not hold in all cases. Introducing risk aversion in the good faith analysis may extend the analysis to include optimal risk allocation between the parties, including the relevance of insurance, which may serve as an interesting future research topic.

An Illustration: The Sale of Stolen Goods

A person who is offered a television in the parking lot outside a bar is rarely in doubt that the television is stolen. In other situations, a person may find it difficult to determine the likelihood that the offered good is stolen. This is the case when used goods are offered for sale outside private homes (garage sales) or sold through pawnshops. In fact, most stolen goods are traded in workplaces where many people interact and rarely in sinister places. A consumer buying

goods in a shop may well presume that the shop seller has legal title to sell the good. However, in many other cases, this is not the case, e.g., especially in auctions on the Internet. There have been numerous cases where a buyer in “good faith” had been contacted by the original owner of a good or artifact, claiming the item handed over. This is a classical legal dispute that goes back to ancient times.

The term *rei vindicatio* stems from Roman law and describes an owner’s right to claim his good back from a person who unlawfully has the good in his possession. *Natural law* or *jus naturale* was the dominant legal foundation for the majority of European countries that coincide with Roman law. The rule was an absolute doctrine of vindication, which was regarded as an inherent attribute embedded in the notion of property rights at that time.

Different legal systems in modern times prescribe various solutions to this fundamental problem of property rights; see Levmore (1987) as well as Mattei (1996) for civil law countries. Some legal systems place the entire risk on the buyer, whereas other systems place the risk on the initial owner. In the latter case, this is the case if the buyer is in so-called good faith. Obviously, the choice of legal rules influence the parties incentives, especially how many resources the initial owner is willing to incur to protecting his property and assets as well as a potential buyers costs of verifying the ownership of the offered good; see Cooter and Ulen (2000).

However, there is a strong strategic interdependency in this situation. A potential buyer, who knows that an initial owner has spent much effort to protecting his assets, implies that the likelihood that the offered good is stolen is small, whereas the contrary is the case in the opposition situation. The key issue is what kind of legal system is most efficient from an economic perspective. The analysis of this question can be analyzed using game theory that includes strategic interdependency between a potential buyer and an initial owner of a good. In fact it turns out that the choice of legal rules in this situation influences parties differently, which has consequences for efficiency.

Modeling Good Faith in the Stolen Good Case

The following results rely in Rose (2010) where the full model is outlined. The article also contains a comprehensive description of how the issue of good faith is treated in several different jurisdictions, including what constitutes good faith. The following is a brief outline of the dynamic game theoretical model with perfect information. There are two players, an initial owner and a potential buyer, who must decide how much effort to spend on protecting his good (denoted c), whereas the potential buyer must determine how much to spend on verifying the ownership (denoted e). Even though there is “perfect information,” so the solution concept is subgame perfect Nash equilibrium; there is still randomness which is modeled by introducing Nature as a probability device, denoted $\pi(c)$ which is decreasing in c .

Let $\Pi(c,e)$ denote the potential buyer’s “posterior” probability or “updated beliefs” that the good is stolen, given e , which is specified as follows:

$$\begin{aligned}\Pi(c,e) &= \lambda\pi(c) \text{ if } e = e_H \text{ where } \lambda \in [0, 1/\pi] \text{ and} \\ &= \pi(c) \text{ if } e = 0\end{aligned}$$

These probabilities reflect the idea that a buyer after incurring effort e may revise his initial beliefs that the good is stolen, so that the posterior beliefs that the good is stolen may either increase or decrease. If the buyer incurs effort e and the new updated beliefs result in an increased probability that the good is stolen compared to the initial beliefs, then a buyer is said to be in *bad faith*, i.e., $\lambda\pi(c) > \pi(c)$; if, however, the probability decreases, i.e., $\pi(c) > \lambda\pi(c)$, then he is said to be in *good faith*. The buyer is also in bad faith, if he deliberately chooses to stay in ignorance, i.e., when $e = 0$.

The basic idea is that if it is common knowledge that there is a 20% probability that the good is stolen, but that the posterior beliefs after trying to verify the ownership shows that the good may be equally stolen, then he is said to be in bad faith. The potential buyer is said to be in good faith if the

posterior beliefs have shown a reduced likelihood that the good may be stolen.

Furthermore, it is assumed that if a buyer accepts an offer to buy a good that might be stolen this will stimulate criminal activity, since it becomes easier to sell stolen goods; hence, the market for stolen goods becomes more profitable. This will cause more people to enter the market, and as a consequence, the probability of theft increases in the economy; see, e.g., Benson and Mast (2001) for a study on private security services as well as Ayres and Levitt (1996) who show that such crime reduction may generate externalities.

The probability of theft also depends on a potential buyer’s decision, i.e., $d = \{\text{accept}, \text{reject}\}$ with the following specification:

$$\begin{aligned}\pi(c,d) &= \varphi\pi(c,d) \text{ if an offer is accepted,} \\ &\quad \text{where } \varphi \in [1, 1/\pi] \text{ and} \\ &= \pi(c) \quad \text{if an offer is rejected}\end{aligned}$$

With a rule where good faith is legally irrelevant, the good is given back to the initial owner irrespective of any good faith, and the purchaser’s utility declines to 0, but at the same time, the initial owner’s utility V is restored. It is assumed that the initial owner can always document that she is the rightful owner. In contrast, the good is only handed over to the initial owner under a *good faith rule* if the buyer was in bad faith and if the case is solved by the police. It is assumed that goods cannot be consumed as well as the original owner’s utility declines by the value of the good taken away.

We first start up by analyzing what is the optimal situation which afterward is compared with the two regimes, i.e., with good faith and one without good faith.

It can be shown that a regime that does not take into consideration the good faith of a buyer of a stolen good will induce a buyer to accept any offer. The buyer will avoid spending resources on examining the ownership of the offered good. On the other hand, it can be shown that under a rule of law where good faith plays a decisive role in determining property rights, there is a positive probability that a potential buyer will incur costs

in order to verify ownership. Compared to a rule of law where good faith is irrelevant, an owner will spend more resources on seeking to protect his property from being stolen. However, more importantly, a rule where good faith is irrelevant, Pareto dominates a rule where good faith may protect an innocent buyer, as a potential buyer wastes resources on trying to verify ownership and owners of goods incur higher costs in order to deter burglars from stealing their property. This is the case only under the assumptions especially that parties are assumed to be risk neutral. Most game theoretic models assume risk neutrality; hence, payoff functions are simple to work with as they are not concave, as is the case when players are risk averse.

Good Faith in Other Settings

The abovementioned example focuses on a central aspect where good faith may be vital. However, there are other examples such as within tort law. To illustrate, the business judgment rule protects top management from liability, but the legal doctrine has severe incentive effects that can also be analyzed with game theory. According to this doctrine, CEOs and board members are not liable for a decision that results in huge financial losses, as long as it makes the decision on a well-informed basis. Inferior business skills do not constitute a basis for holding an incompetent management liable. A decision which is based on a business judgment made in good faith, but which results in a substantial financial loss for the company, does not make the management liable to legal proceedings by shareholders or creditors.

One may argue that the business judgment rules increase shareholder returns over a situation where management conduct is subject to a strict negligence standard. The rule allows management to engage in more risky projects without the fear of being held personally liable. It is by no means certain that the courts have sufficient knowledge to determine *ex post facto*, whether a business decision should have been taken or not. In the absence of the business judgment rule, the volume of litigation would increase and transaction costs would rise, not only due to higher fees for lawyers

and court fees but also because of increased managerial opportunity costs. The basic idea is that there is a strategic interdependency between managers and shareholders. Good faith protects managers from legal liability, but it also influences the return of the shareholders and their incentives to sue management, which is analyzed in Rose (2011).

References

- Ayres I, Levitt SD (1996) Measuring positive externalities from unobservable victim precaution: an empirical analysis of Lojack. John M. Olin Center for Law, Economics, and Business, Harvard Law School. Discussion paper no 197
- Baird D, Gertner RH, Picker RC (1998) Game theory and the law. Harvard University Press, Cambridge
- Benson BL, Mast BD (2001) Privately produced general deterrence. *J Law Econ* 44(2), part. 2, pp 725–746
- Cooter R, Ulen T (2000) Law and economics, 3rd edn. Addison-Wesley
- Landes WM, Posner R (1996) The economics of arts. In: Ginsburgh VA, Menger PM (eds) Selected essays. Elsevier
- Levmore S (1987) Variety and uniformity in the treatment of the good faith purchaser. *J Leg Stud* XVI:43–65
- Mattei U (1996) Property rights in civil law. In: Bouckaert B, De Geest G (eds) Encyclopedia of law and economics. pp 162–165, Edwar Elgar
- Rasmussen E (2004) The economics of agency law and contract formation. *Am Law Econ Rev* 6:369–409
- Rose C (2010) The transfer of property rights by theft: an economic analysis. *Eur J Law Econ* 30:247–266
- Rose C (2011) Director's liability and investor protection: a law and economics perspective. *Eur J Law Econ* 31: 287–305

Gordon Tullock: A Maverick Scholar of Law and Economics

Richard E. Wagner

Department of Economics, George Mason University, Fairfax, VA, USA

Abstract

This essay sketches the central features of Gordon Tullock's (1922–2014) contributions to law and economics. My reference to Tullock as a

“maverick scholar” is to indicate that he stood apart from the mainstream of law and economics scholarship as this is represented by Richard Posner’s canonical statement that the common law reflects a relentless pursuit of economic efficiency. In contrast, Tullock denied efficiency claims on behalf of common law. Examination of Tullock’s contrary claim illustrates how any analytical claim necessarily rests on and is derived from some preceding set of conceptual presuppositions because such qualities as “efficiency” are not objects of direct apprehension but rather are inferences derived from particular theoretical frameworks.

Biography

It is worthwhile noting that Tullock’s only academic degree was a JD from the University of Chicago in 1947. He was also an undergraduate at Chicago, enrolling in 1940, and was eligible to receive his baccalaureate degree but elected not to pay the \$5 fee required to obtain it because he was already eligible to enter law school and recognized that the law degree he would soon receive would trump a bachelor’s degree. His legal studies were cut short, however, by his being drafted into the Army in 1943, where he served in the Infantry and entered into France a week after the D-Day landing in June 1944. After the war, Tullock returned to Chicago to finish his legal program. Tullock spent a few months as a practicing attorney upon graduation from Chicago and then spent 9 years with the US Department of State, including assignments in China, Hong Kong, and Korea. While with the Department of State, Tullock published three papers in major economics journals, all on Asian topics. He had also written a manuscript that was finally published in 1965 as *The Politics of Bureaucracy*, in which he sought to systematize some of his experiences and observations while working with the Department of State. While Tullock recognized that neither legal nor diplomatic practice was how he wanted to live his life, he nonetheless carried into his subsequent academic scholarship the practice-

oriented orientation toward his material that his legal and diplomatic activities entailed.

Gordon Brady and Robert Tollison (1991) describe Tullock as a “creative maverick” with respect to public choice theory, and in this designation, they are surely right. Tullock was equally the creative maverick with respect to law and economics. It should be noted that his work in what would generally be considered law and economics is a relatively small portion of his full body of work. This smallness can be seen directly by examining the ten volumes of Tullock’s *Selected Works* that Charles Rowley (2004–2006) collected. *Law and economics* was the title of one of those ten volumes. In addition, about 10 % of the first volume titled *Virginia Political Economy* is classified as pertaining to law and economics. The preponderance of Tullock’s *Selected Works* was distributed across such fields of inquiry as public choice, rent seeking, wealth redistribution, bureaucracy, social cost, social conflict, and bioeconomics.

Yet we may doubt that any exercise in counting pages can give an accurate representation of Tullock’s mind at work. An observer might distribute Tullock’s works across discrete fields to provide a scheme of classification, but for Tullock his scholarly activities reflected a mental unity across these superficially diverse fields of inquiry. There was an animating form to Tullock’s scholarly inquiries and with that form brought to bear on diverse particular objects of inquiry. With respect to Isaiah Berlin’s (1970) essay on Tolstoy, Tullock was surely more the hedgehog than the fox. In his survey of Tullock’s scholarship as of the mid-1980s, Richard Wagner (1987: 34) asserts that “someone writing a survey of Tullock’s works would surely think he was surveying the work of the faculty of a small university.” With respect to his substantive interests, Tullock’s scholarship made significant contact with such university programs as public administration, philosophy, history, biology, sociology, criminology, military science, international relations, and Asiatic studies, in addition to his contributions to law, economics, and political science. Despite this substantive variety within Tullock’s body of work, his varied inquiries sprang from and reflected a

common analytical core, which I shall explore briefly with particular regard to his scholarship in law and economics.

A. How do law and economics fit together? Melvin Reder (1999) reminds us that economics is an intensely contested discipline. So, too, is law. So how is the compound term “law and economics” to be construed when scholars have a menu of options available to them for thinking about economics and also about law? There are several ways that an economist can theorize about his or her object of interest. A legal scholar faces the same situation.

The primary concept of economic theory is the abstract noun designated as a “market economy.” Economists use this idea in explaining how it is that economic activities within a society are generally coherent and coordinated even though there is no person or office who creates that coherence. To be sure, a good number of economists find fault with market outcomes and seek to explain how political power can be deployed to secure what that economist regards as improvement. Nonetheless, the fundamental mystery of economic theory is to explain how societies exhibit coherence and coordination even though each member of that society operates mostly in self-directed fashion. The central answer economists give to explain this mystery relies upon the presence of an institutional framework grounded in the principles of private property, liberty of contract, and freedom of association. This institutional framework brings the legal order directly into the analytical picture, for the efficacy of the market order depends on the quality of the complementary legal framework. Hence, law and economics are two sides of the proverbial coin.

There are, however, different frameworks for economic analysis, and those different frameworks will in turn commend different orientations toward the relationship between law and economics. What Melvin Reder (1982) describes as Chicago economics rests on the twin claims that people maximize given utility functions and that market clear. These claims generate the Pareto efficiency of competitive equilibrium. An economist who thinks that the Pareto efficient character

of competitive equilibrium provides a good analytical window for examining economic activity will require a complementary legal framework should that economist choose to explore legal doctrines and procedures. Richard Posner’s (1973) animating claim is that the array of common law doctrines and procedures can be rendered coherent by the principle that they reflect the promotion of economic efficiency within society. Posner’s claim about law is thus complementary with Reder’s claim about the efficiency of competitive equilibrium, when coupled with the further claim that the competitive model is a good analytical window for viewing economic activity within society.

Equilibrium theory is not the only window through which economic phenomena can be viewed. One problem that arises with this approach to economic inquiry is that change comes as exogenous shocks to equilibrium and not as internally generated features of a dynamic economic process. Someone who thinks in terms of the continual generation of change as a quality of an economic system will likewise and necessarily recognize that conflict and also limited and distributed knowledge is part of that system. This alternative analytical window renders claims on behalf of efficiency dubious, mostly because the dynamic economic process is simultaneously extinguishing and creating profit opportunities, with new commercial plans continually being formed simultaneously with other commercial plans being abandoned. Social processes are naturally turbulent and there is no God’s-eye vantage point from which efficiency can be determined. This view of social economic processes, which was Tullock’s view, will lead in turn to the examination of legal doctrines and procedures through a different analytical window than through which an efficiency-always theorist would use.

James Buchanan (1987) argues that Tullock was a “natural economist.” By this, Buchanan meant that Tullock thought naturally in terms of the universal quality of the logic of choice and economizing action, rendering Tullock a theorist in the style of *homo economicus*. There is no small irony in this description, given that one of Buchanan’s widely recognized claims is that

economics should be about exchange and not choice. For Tullock, however, the logic of choice was only a point of analytical departure. It was never a destination, for Tullock was always a social theorist who never reduced society to a representative agent. Tullock was a theorist of interaction, perhaps more so than Buchanan, and not a theorist of choice, as Wagner (2008) explains in his comparison of Tullock's *Social Dilemma* and Buchanan's *Limits of Liberty*.

For Tullock it is necessary to distinguish the form by which a theory of utility maximization is expressed from the substantive content of particular instances of human action, all of which are capable of being rendered intelligible within that formal shell. The postulate of rational action means simply that people seek to succeed and not fail in their chosen actions. This is a meta-physical ordering principle and not some behavioral hypothesis. Albert Schweitzer and Adolf Hitler were both utility-maximizing creatures who pursued disparate plans of action that other people appraise to strikingly different effect, but with that difference reflecting the importation of substance into form. George Stigler and Gary Becker (1977) argued that the core of economic theory should rest on the presumption that preferences are invariant across time and place. In contrast, Tullock's analytical core was more like what Ross Emmett (2006) expressed in speculating on what Frank Knight would have said in response to Stigler and Becker. It's also similar to the contrast Richard Wagner (2010: 11–16) advances between neo-Mengerian and neo-Walrasian research programs.

The form of Tullock's theorizing was grounded in rational choice, but for Tullock form was only partly biologically or genetically determined. Tullock would not reduce economics to ethology because continuing and creative social interaction also resided within his analytical core. This brought continuing conflict and continual novelty into his social theory. In other words, Tullock's theoretical framework was one of spontaneous ordering and creative evolutions, as Todd Zywicki (2008) recognizes in his examination of Tullock's critique of the efficiency claims often made on behalf of common law.

I don't recall ever seeing or hearing Tullock refers to Carl Schmitt's (1932) treatment of the autonomy of the political in society, which expressed the idea that politics could never be reduced to law through constitutional design. Yet in a 1965 essay on *Constitutional Mythology* which is not included in his *Selected Works*, Tullock embraced the impossibility of eliminating politics through constitutional law. For Tullock, law and politics were inseparable, and both domains were populated by economizing creatures that gave conceptual coherence to Tullock's various particular lines of scholarship.

B. Tullock and Posner and their Contrasting Orientations. In his contribution for a Festschrift honoring Tullock, Charles Goetz (1987) observes that Tullock's scholarship in law and economics has been largely ignored, in contrast to the reception his scholarship has received in other fields. Tullock (1996: 1), moreover, concurred in Goetz's judgment when he noted that "I have been cast into the outer darkness . . . by the law and economics movement in the United States." In making his point, Goetz compares Tullock (1971) with Posner (1973). Goetz attributes Tullock's lukewarm reception among lawyers as largely a product of Tullock's informal and conversational style of presentation. Recognizing that Goetz wrote this from his vantage point as a professor of law in a premiere law school, I am in no position to challenge that claim. All the same, I think there is more to the differences in reception accorded to those works and their theorists. That difference partly reflects the ability of different conceptual formulations to enable other scholars to do the work they choose to do. Posner (1973) accomplished more than Tullock (1971, 1980a) in this respect. But why is that so? What lessons can be gleaned from this comparison?

Consider one of the several examples where Posner asserted a claim on behalf of legal efficiency and which Tullock (1980b) disputed. Consider the common law principle that railroads owed a duty of care to pedestrians only at crossings, whereas they always owed a duty of care to straying cattle. About this principle, Posner offered the gloss that it would be less costly for a pedestrian to choose a different path than it would

be for a railroad to watch continually for pedestrians. Posner further claimed that the reverse relationship would hold for straying cattle, for it would be more costly for a farmer to fence cattle than it would be for a railroad to keep watch. This gloss fits the efficiency rationale. But is that gloss warranted? Tullock claims that it wasn't. Perhaps walking along the railroad path might save a pedestrian only 100 m as compared with following the sidewalk. But what if that alternative distance were 1,000 m, or even longer? Tullock points out that Posner's empirical claim about relative costs is accompanied by no evidence about those costs. Even more, the cost of watching for humans and cattle are joint and not separable costs. Given that a railroad will be watching for cattle, the marginal cost of watching for humans is zero.

At this point we return to the unavoidable situation where the theoretical framework we use influences what we see in the first place. "Efficiency" is not something that is directly apprehensible. It is rather a conclusion that is drawn from a particular analytical model and with different models often leading to different conclusions. For instance, economists describe a competitive equilibrium as entailing the condition that price equals marginal cost. Cost functions are boundaries that separate what is possible from what is impossible. An average cost function, for instance, defines a boundary where it is impossible for cost to be lower but it is possible for cost to be higher. In recognition of this boundary quality, one might reasonably wonder why economists regularly presume that economic outcomes occur along the boundary and not at some interior position. The answer has nothing at all to do with observation, for it is impossible to observe what is truly the least cost manner of doing something. One might observe that some people produce in lower-cost fashion than other people without being able to determine the lowest possible cost of production. Rather what economists do is make the plausible claim that an individual proprietor would prefer to avoid waste in his or her operation because the proprietor owns the residual between revenues and expenses and transform this reasonable claim into a theoretical generalization.

Tullock's claims that deny the efficiency claims on behalf of common law are not evidentiary-based claims that can point to a failure to use lower-cost options. Tullock's claims are rather advanced in recognition that no evidence has been presented one way or another. In the railroad and cattle case, evidence would entail separately generated data pertaining to different ways of fencing cattle relative to different costs to a railroad of trying to avoid accidents. It may very well be the case that the cost of making finer determinations about costliness in such cases may exceed what people are willing to spend to make such a determination. A judgment will have to be rendered in the case all the same, simply because the case must be resolved in some fashion so that life can go on. This situation doesn't render the common law economically efficient by default, but rather renders the question impossible of being answered.

A famous problem in statistical decision theory involves a lady who claims that in tasting a cup of tea, she can tell whether the tea or the milk was put first into the cup (Neyman 1950). A judge must decide whether to accept or reject the lady's claim. To do this an experiment is designed and evidence generated, from which a judgment is made. With respect to law and economics, this setting is equivalent to asking whether or not the lady is an efficient oracle for determining how a cup of tea is made. Yet there is no way of making such a determination with perfect accuracy because the "truth" of the matter cannot be determined independently of the evidence and standard of judgment used to reach a determination. Errors of judgment will be unavoidable. The lady's claim might be granted when she truly can't tell the difference between the methods. It might also be rejected even though she can tell the difference. Furthermore, an increased effort to avoid one type of error will increase the frequency with which the other type of error is made.

In this setting, efficiency or accuracy is not a binary variable of yes or no but is a quantity that has the property of more and less. Many procedures for reaching a judgment are possible, and these procedures will typically differ both in their costliness and in their ability to avoid one or the

other type of error. Tullock's approach to legal efficiency fits fully within the framework of decision theory. Aside from accuracy and cost, there is also question of the standard of judgment and with higher standards of judgment typically involving more costly procedures. For instance, accuracy will increase as the lady is given more cups of tea to taste, but cost will also rise with the number of cups tasted. It is here where standards of judgment come into play. One judge might feel comfortable granting the lady's claim by invoking a 60 % confidence level, perhaps as expressed by a preponderance of evidence. Another judge might insist upon a 95 % confidence level, perhaps as expressed by overwhelmingly strong evidence. In these situations there is no God's-eye platform from which "truth" can be pronounced. Rather what exists are various decision procedures with varying ability to provide useful evidence and which differ both in their costliness and in their error-avoidance properties.

Tullock's approach to the cost and accuracy of judicial proceedings fits within this decision theory motif. For instance, where Posner analogizes the common law to a competitive market system, Tullock analogizes it to a socialist bureaucracy. There is something to be said in support of each analogy, while there are also things to be said against each of them. The common law system features competition among attorneys for custom and between attorneys at trials. Tullock argues that this system resembles an arms race with both cheap and expensive points of equilibrium and with the process tending toward the expensive equilibrium. On this basis Tullock favors the civil law procedure where judges and not lawyers are the dominant players. But judges are bureaucrats who are paid through tax revenues and not from clients who are seeking their services. Furthermore, law and politics are deeply entangled, as all legal systems are replete with public ordering where the private law principles of property and contract are hemmed in through various politically articulated requirements. Even arbitration, which entails far more private ordering than either common or civil law, requires the willingness of public officials to enforce the judgments that are reached in those proceedings.

Impact and Legacy

A scholar's legacy is, of course, something that will be determined with the passing of time. At the present time, the Goetz-Tullock appraisal of Tullock's reception within the Anglo-Saxon world remains reasonable. That appraisal reflects in good measure the continued predominance of static equilibrium as the core of economic theory. There are, however, signs that various forms of ecological and evolutionary theorizing are gaining momentum within economics. Should that momentum continue to build in the coming years, it is certainly possible that the process-oriented and substantive questions that Tullock addressed will rise within the attention spaces of law and economics scholars relative to formulations that address questions that pertain to models of static equilibrium.

References

- Berlin I (1970) The hedgehog and the fox: an essay on Tolstoy's view of history. Simon and Schuster, New York
- Brady G, Tollison RD (1991) Gordon Tullock: creative maverick of public choice. *Public Choice* 71:141–148
- Buchanan JM (1987) The qualities of a natural economist. In: Rowley CK (ed) *Democracy and public choice: essays in honor of Gordon Tullock*. Basil Blackwell, Oxford, pp 9–19
- Emmett RB (2006) Die gustibus est disputandum: Frank H. Knight's response to George Stigler and Gary Becker's 'die gustibus non est disputandum'. *J Econ Methodol* 13:97–111
- Goetz CJ (1987) Public choice and the law: the paradox of Tullock. In: Rowley CK (ed) *Democracy and public choice: essays in honor of Gordon Tullock*. Basil Blackwell, Oxford, pp 171–180
- Neyman J (1950) *First course in probability and statistics*. Henry Holt, New York
- Posner RA (1973) *Economic analysis of law*. Little, Brown, Boston. (8th edn, 2011)
- Reder M (1982) Chicago economics: permanence and change. *J Econ Lit* 20:1–38
- Reder MW (1999) *Economics: the culture of a controversial science*. University of Chicago Press, Chicago
- Rowley CK (ed) (2004–2006) *The selected works of Gordon Tullock*. Liberty Fund, Indianapolis
- Schmitt C (1996 [1932]) *The concept of the political*. University of Chicago Press, Chicago
- Stigler GJ, Becker GS (1977) De gustibus non est disputandum. *Am Econ Rev* 67:76–90

- Tullock G (1965) Constitutional mythology. *New Individualist Rev* 3:13–17
- Tullock G (1971) *The logic of the law*. Basic Books, New York
- Tullock G (1980a) *Trials on trial: the pure theory of legal procedure*. Columbia University Press, New York
- Tullock G (1980b) Two kinds of legal efficiency. *Hofstra Law Rev* 9:659–669
- Tullock G (1996) Legal heresy. *Econ Inq* 34:1–9
- Wagner RE (1987) Gordon Tullock as rhetorical economist. In: Rowley CK (ed) *Democracy and public choice: essays in honor of Gordon Tullock*. Basil Blackwell, Oxford, pp 27–38
- Wagner RE (2008) Finding social dilemma: West of Babel, not east of Eden. *Public Choice* 135:55–66
- Wagner RE (2010) *Mind, society, and human action: time and knowledge in a theory of social economy*. Routledge, London
- Zywicki TJ (2008) Spontaneous order and the common law: Gordon Tullock's critique. *Public Choice* 135:35–53

Governance

Paul Dragos Aligica
 Mercatus Center, George Mason University,
 Arlington, VA, USA

Definition

“Governance” is a concept that has emerged to salience at the interdisciplinary interface of several fields relevant for law and economics. A typical governance situation involves not just formal institutions but informal ones, social norms and rules, law, legislation and economic incentives, regulations and enforcement, arbitration and conflict resolution mechanisms, preference aggregation procedures, values and perceptions, and path dependent elements. In the diverse literatures dealing with these phenomena, several major modes of using the notion of “governance” have emerged: governance as describing a recently emerging phenomenon transcending the traditional regulatory, administrative, and public/private choice interface functions of modern states; governance as a functional domain – structures and functions operating in relationship to commons, natural resources and specific collective goods production,

management and consumption processes; governance as a theoretical apparatus, a set of concepts, models, and theories aiming at analyzing regulatory, administrative, and public/private choice interface situations, including those depicted by the abovementioned literatures. The concept also applies to the related phenomena associated with the modern corporation and its (legal, political, and social) operational environment. The notion of “corporate governance” is, for instance, a typical example of how this concept is blurring the lines between the “public” and the “private,” between regulation and self-regulation, between markets, hierarchies, and networks.

Governance: A Multifaceted Motion

The concept of “governance” has emerged in the last several decades as an alternative mode of theoretically framing the complex political, economic, legal, and social mechanisms and processes associated to the collective decision making, regulation, and administration of political and economic systems, transcending the classical “markets” versus “states” and “public” versus “private” approach. At the same time, it is a way of describing all political and public administration processes and cases taking place in addition to, beyond, or outside of the state/government-centered structures and methods. That covers both the cases of complex modern large-scale advanced postindustrial democracies systems and the cases of groups, small communities, and societies that do not share the scale and complexity of the modern systems. The concept also applies to the related phenomena associated with the modern corporation and its legal, political, and social operational environment. The notion of “corporate governance,” is, for instance, a typical example of how this concept is blurring the lines between the “public” and the “private,” between regulation and self-regulation and between markets, hierarchies, and networks.

Thus, “governance” is by its very nature, a concept emerging at an interdisciplinary interface. A typical governance situation involves not just formal institutions but informal ones,

social norms and rules, law, legislation and economic incentives, regulations and enforcement, arbitration and conflict resolution mechanisms, preference aggregation procedures, values and perceptions, and path-dependent elements. A large literature (both positive-analytical and normative-prescriptive) has grown under this label with roots in political economy, law and economics, political philosophy, public administration, management studies, anthropology, sociology, and history. At the core of this exploding literature stands a large variety of “governance” forms and domains and their multiple facets, as framed using diverse theoretical lenses. This includes governance as a political theory concept, as a structure (architecture of institutions), as a process (the dynamics of functions), as a mechanism (procedures and decision rules), as a public choice intuitionism theory, as a strategy (actions related to control and institutional design shaping preferences and decisions), and as a normative ideal.

Due to its multifaceted nature and its versatility and traveling capacity from one discipline and intellectual tradition to another, the notion of “governance” is also increasingly assuming all the features of a typically “essentially contested” concept, gradually joining in this respect other key terms that form our political vocabulary such as “democracy,” “liberalism,” “government,” “politics.” “Essentially contested” concepts do not display a mere ambiguity about usage, about a core meaning, which could be solved using standard analytical methods, but reflect a deeper tension or disagreement about different configurations of theoretical, descriptive, and normative elements that are implicit in the usage of the concept by different schools of thought, disciplines, and intellectual traditions. These systems or configurations of ideas, which embed the concept in case, are sharing some elements – sometimes at all three levels, descriptive, theoretical, or normative – but not all of them, a fact which makes the disentanglement of the meaning of the concept in case always extremely difficult and contestable. When it comes to the concept of “governance” that means that the usual caveats applied to any encyclopedia entry

dealing with this special class of concepts, should apply to it as well.

In what follows, this entry will illustrate the potential of the family of perspectives gravitating around the concept, by focusing on three modes of using it. All are at the general and encompassing level, all are backed by solid literatures: (1) governance as describing a recently emerging phenomenon transcending the traditional regulatory, administrative and public/private choice interface functions of modern states; (2) governance as a functional domain – structures and functions operating in relationship to commons, natural resources and specific collective goods production, management and consumption processes; (3) governance as a theoretical apparatus, a set of concepts, models, and theories aiming at analyzing regulatory, administrative, and public/private choice interface situations, including those depicted by the above mentioned literatures.

The New Realities of “Governance”

The starting point of the first is the observation that a “new reality” has been created at both the national and the global levels: Globalization, competition, technological change, financial pressure, and increased demands due to changes in perceptions, preferences, values, and expectations put an increasing pressure of governments. A shift has taken place. The new conditions have generated a pressure to adjust the regimes centered on nation-states and on the hierarchical command and control arrangements based on Weberian bureaucracy and Wilsonian public administration. The typical national governmental structures are overextended, lacking capabilities and unable to administer, intervene and control as directly and as effectively as before. Because the state capacity of control diminishes, it has to adjust its modus operandi. A change of approach has to take place focusing now on indirect, negotiated influence, “outsourcing” or building partnership with the private sector and civil society.

Private actors and hybrid forms of organization become players in traditionally public domains, blurring thus the public-sector/private-sector

distinction. The retreat of government's control has encouraged an expanding role of civil society, a diminished view of the role of elected officials, an emphasis on political entrepreneurship, the pooling of public and private resources, and reliance on market or quasi-market discipline. Decentralization and de-monopolization has encouraged a variety of formal and informal hybrid public-private organizational forms, pluralist, multilevel, and polycentric arrangements. These changes have led to a move from politics, hierarchies, and communities to markets; from provision to regulation; from public authority to private authority; from big government to small government; and from regulation based on command and control to flexible and diverse forms of regulation in which self-regulation is an important element. The overall result has been the growth of "governance without government." Given these developments, "governance" has come to indicate an entire range of public and semi-public mechanisms and institutions for implementing policy, working either as alternatives to "government" or in conjunction with it. The boundaries between the public, private, and voluntary sectors have changed and with that the very role of the state (Peters and Pierre 1998).

A very telling metaphor recurrently used to portray these phenomena is based on the distinction between "rowing" functions (service provision, taxation, redistribution) and "steering" functions (rulemaking, monitoring and enforcement). Expressed in these terms, the point is that a shift has taken place. Initially, states were bent on command and direct provision, insisting on doing both the "rowing" and "steering." Now states are increasingly focused on "steering," while markets, the voluntary sector, and nonprofit civil society do the "rowing." The literature is not displaying a unitary interpretation; there are alternative modes of understanding the "new reality." For instance, on the one hand, there is the "hollowing out the state" thesis whereby authority is shifting either "up" to markets and transnational institutions or "down" to local governments, business communities, and nongovernmental organizations. On the other hand, there is the thesis that the state has in fact reasserted its authority by

shifting its focus at the meta-level towards "regulating the mix of governing structures such as markets and networks and deploying indirect instruments of control," the notion of meta-governance (Levi-Faur 2012, p. 36).

As noted, the developments outlined above have generated a robust literature. One could appreciate the publication of volumes like *The Oxford Handbook of Governance* (Levi-Faur 2012) and *The Oxford Handbook of Public Management* (Ferlie et al. 2007) as one of the most significant indicators of the status and impact of that literature. The first volume charts the "new reality" created at both the national and the global levels as the role of the state and the boundaries between the public, private, and voluntary sectors have changed. It illustrates how these themes are at the core of both the academic and applied agenda, and it points out to some ways of theorizing, describing, and analyzing them. The second volume takes the applied angle and documents the emergence to salience of themes such as proactive administration, administrative responsiveness and responsibility, street-level bureaucracy, citizen participation, public-private partnerships, decentralization, hybridity, contracting, and NGOs.

Governance as a Functional Domain

The second mode of illustrating the "governance" perspective is as the study of institutional arrangements dealing with specific social dilemmas of collective action, in particular circumstances defined in functional, local, and contextual terms. Its most salient application is regarding the commons, natural resources, and specific collective goods production, management, and consumption processes. The contribution of 2009 co-recipient of Nobel Prize in Economics, Elinor Ostrom, is illustrative, her "governing the commons" work being a source of inspiration and considered as paradigmatic in this respect (Ostrom 1990).

From this perspective, the notion of "governance" is a term associated to a set of institutional and procedural solutions to collective action situations generated by the collective production, financing, and consumption of goods and services

that have particular properties that diverge from the private goods ideal (in terms of excludability and jointness of consumption) (Ostrom and Ostrom 2004). A “commons” is a good that is shared by multiple users and has two properties: rivalrous and nonexcludability. The use of a commons by one user changes the state of the commons and has implications for the preferences and behavior of other users, who change their behavior and overuse or deplete the resource, leading, without intention, on the long run, to a decrease of the welfare for the entire group: This is the so-called “tragedy of the commons.” It is by definition a collective-action problem: The incentives defining a situation induce a divergence between the individual and collective logic, generating a rush to enjoy/use/consume the resource. Thus, the commons is predicted to deteriorate, affecting the welfare of all the commons users.

E. Ostrom and her team of researchers not only built an extensive database and a theoretical and methodological apparatus to study how people deal with such situations, but also engaged in extensive fieldwork examining particular case-studies of governance systems of fisheries, irrigation systems, pastures, forests, etc., in both developed and developing countries, at various levels and scales. At the core of the agenda was a demonstration that the “tragedy of the commons” could be and has been avoided in many cases, without making use of state-based mechanisms and enforcement. Ingenious institutional arrangements and social norms were created and enforced, leading to viable and resilient governance structures. The “governing the commons” research program has thus shown that “governance” is possible outside the state-market dualism that dominated mainstream policy analysis. It illustrates empirically and analytically not only that community-based governance may be a viable alternative, but also that governance functions may be operationalized through diverse institutional forms. Hence the notion of “governance” has come to be associated to the notion of “institutional diversity.” In addition to that, a significant part of the work was focused on the conditions and mechanisms motivating the commons users to behave cooperatively. E. Ostrom identified a set

design principles, noting that they characterize robust institutions used by communities managing common-pool resources such as forests or fisheries. With that, the notion of “governance” has also come to be associated to the notion of “institutional design.”

Last but not least, the “design principles” demonstrated the possibilities for self-governance of human communities and the ability of people to bottom up design and manage solutions to social dilemmas. The Ostroms’ research program exploring the governance of the commons has thus to be seen as part of a larger program having an assumed applied and normative dimension: “What is missing from the policy analyst’s tool kit – and from the set of accepted, well developed theories of human organization – is an adequately specified theory of collective action whereby a group of principals can organize themselves voluntarily to retain residuals of their own efforts” (Ostrom 1990, p. 24).

Governance as a Theoretical Apparatus

The third mode of using the concept of “governance” is precisely in connection to this theoretical toolkit. “Governance theory” is in this respect the range of conceptual and theoretical instruments available for the study of the phenomena identified and described above. From this perspective, “governance” refers to a description of “governance phenomena” (as social coordination, cooperation, and institutionalization) as seen through the lenses of this theoretical apparatus. It is a dual relationship. Theoretical lenses lead to a specific description and diagnosis of governance problems. While the problems, thus defined, entail for their analysis and solutions, the logic and social mechanisms implied by the toolkit. Most of the components of the toolkit have origins in economics: public choice, industrial organization, institutional theory, applied game theory, etc.

The theory of private and public goods, as already suggested above, offers a very effective illustration in this respect: Because the nature of goods determines the practicality of the various institutional structures meant to produce, finance, distribute, and consume these goods, a typology

of goods is a first step in understanding institutional diversity. Hence, a governance theorist approaches institutional arrangements as driven by a functional structure shaped by the nature of goods or services that those institutions are supposed to satisfy. Thus, a typology of goods (based on the “exclusion” and “jointness of use or consumption” dimensions, sometimes further elaborated in game theoretical forms) helps identify typologies of institutional forms. At their turn, the typologies of institutional forms help chart and analyze governance situations. This is the strategy advanced, for instance, by E. Ostrom and V. Ostrom (2004) or by 1986 Nobel Prize in Economics, James Buchanan (1967).

In a similar way, one may use transaction cost analysis, on the lines pioneered by Ronald Coase (1991 Nobel Prize in Economics) and further elaborated by O. Williamson (2009 Nobel Prize in Economics co-recipient). The logic is similar. The diverse nature of the transactions people have to engage into and especially their associated costs (search, monitoring, enforcement etc.) determine the practicality of the various institutional structures meant to deal with those transactions. (Coase 1992; Williamson 1979) Hence, a typology and analysis of transactions and of the costs associated to them is a first step in understanding institutional structure, function, and diversity. At its turn, the typology of institutions helps chart and analyze governance situations.

The list may go on: externalities, agent-principal, rent seeking, asymmetric information are all part of the catalogue or toolbox of the governance theorist or analyst. It is a modular toolkit which has a double function: On the one hand, it offers a vocabulary and thus the lenses to identify and describe governance situations, and at the same time, it offers analytical instruments to explain the social mechanisms and processes in place or to suggest such mechanisms and structures as governance solutions to the problems thus diagnosed. In this respect, it has strong connections with mechanism design theory (2007 Nobel Prize in Economics for Leonid Hurwicz, Eric Maskin and Roger B. Myerson) – an approach that uses “reverse game theory,” which starts with a desired function or objective, and tries to

identify and reconstruct the governance and institutional mechanism which may be able to achieve that specific objective or realize that functional solution (Maskin 2008). As such the “governance” approach assumes strong and unequivocal normative and applied level dimensions.

Cross-References

- [Bloomington School](#)
- [Constitutional Political Economy](#)
- [Public Choice: The Virginia School](#)

References

- Buchanan JM (1967) Public goods in theory and practice: a note on the Minasian-Samuelson discussion. *J Law Econ* 10:193–197
- Coase RH (1992) The institutional structure of production. *Am Econ Rev* 82(4):713–719
- Ferlie E, Lynn LE, Pollitt C (2007) *The Oxford handbook of public management*. Oxford University Press, Oxford
- Levi-Faur D (ed) (2012) *The Oxford handbook of governance*. Oxford University Press, New York
- Maskin ES (2008) Mechanism design: how to implement social goals. *Am Econ Rev* 98(3):567–576
- Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK
- Ostrom E, Ostrom V (2004) The quest for meaning in public choice. *Am J Econ Sociol* 63(1):105–147
- Peters BG, Pierre J (1998) Governance without government? Rethinking public administration. *J Public Adm Res Theory* 8(2):223–243
- Williamson OE (1979) Transaction-cost economics: the governance of contractual relations. *J Law Econ* 22(2):233–261

Government

Walter E. Block

Joseph A. Butt, S.J. College of Business, Loyola University New Orleans, New Orleans, LA, USA

Abstract

The institution of “government” is perhaps more debated among philosophers, political scientists, economists, historians, and other

social scientists than any other. In order to give a comprehensive assessment of this concept, I will offer several perspectives on it.

Definition

Government is a criminal gang with a lot of good public relations, so much so that most people think it has legitimacy

The Theocratic

In this vision, the king is the head of government and is also the leading religious figure. Sometimes, this type of emperor is actually God himself. As owner of the kingdom, literally, he is the unquestioned master of not only every stick, blade of grass, and other property in his realm but also the people in it. They are in effect his slaves; he is their master. His authority is unquestioned, both from the religious and secular perspective. One need not spend too much time criticizing this form of government. How can it be proven that any given individual is actually God or his chosen spokesman? Why is it justified that it is from this particular family that kings, queens, and princes emanate? Typically, this is based on success in war, but any justification for this process must rest on the unproven contention that “might makes right.”

The Democratic

Here, at least in pure form, the government is elected by the governed, and all decisions are based upon majority rule. The franchise might be broad or very narrow (e.g., limited to white, male property owners). The inclusive version of this system enjoys wide support. Most countries in the so-called civilized world engage in this political form. According to Winston Churchill, “democracy is the worst form of government except for all the others that have ever been tried.” But there are problems here. For one thing, Hitler, no popular figure, came to power

through a completely democratic process, not via a coup d'état. (Of course, it cannot be denied that he remained in power by abrogating the democratic system.) One is tempted to conclude from this one fact, “so much for democracy.” But this is only the tip of the iceberg, for there is such a thing as “the tyranny of the majority.” Just because a majority of the electorate prefers one candidate or policy over another does not guarantee he or it is justifiable. In many southern towns in the United States in the nineteenth century, a majority of the people might well have voted to lynch all black people.

The Republic

This form of polity allows for democracy but is subject to a constitution. That is, the society can vote alright, but it cannot cast a ballot for anything and everything. For example, some things cannot be determined by majority vote, such as should a certain minority of people be eliminated. This seems like an improvement over pure democracy, but it all depends upon how good the constitution is and whether or not it is followed. For instance, the US constitution allowed for slavery; the majority need not have voted to enslave black people: the constitution already allowed for that. Another example was the constitution of the Soviet Union. It protected all sorts of minority rights. But this relatively splendid document was all but ignored.

What have libertarians had to say about government? I pursue this question since this philosophical movement has some interesting and important things to say about this institution. Four different versions of libertarianism can be identified in terms of their vision of proper government. They are as follows:

1. Anarcho-capitalism
2. Minarchism
3. Classical liberals, or Constitutionalists
4. Free market supporter

Libertarianism is predicated upon two basic building blocks: the nonaggression principle

(NAP) and private property rights based upon homesteading. The first maintains it is improper to threaten or initiate violence against an innocent person or his property. The second is a theory as to how just property titles come into being. They start with self-ownership, continue with the mixing of one's labor with land and other natural resources, and conclude with any voluntary means of title transfer, such as buying and selling, gifts, and gambling.

Anarcho-Capitalism

The anarcho-capitalist version of libertarianism cleaves most closely to these principles. It brooks no compromise, none whatsoever, with private property rights and the NAP. That is, it concludes that government is a per se violation of rights. All governments do two things: they tax their subjects and forcefully prevent other entities from competing with them, within their geographical territory. But taxation is theft. People did not agree to be taxed. Any "social contract" to the contrary is a figment of the statist's imagination. Taxes are not akin to club dues. The latter rises from contract, not the former. Thus, for the anarcho-capitalist version of libertarianism, government is not merely akin to a robber gang, but indeed takes on that exact description. The only difference between it and a bunch of outright thugs is that the state has legitimacy in the minds of most people and that gangsters do not. So similar are the two, apart from the far better public relations of the government, that when a criminal engages in a crime such as murder, rape, and theft, he can be said to have undertaken a governmental act (apart from the seeming legitimacy of it).

The scholars most closely associated with this perspective are Murray N. Rothbard, Lysander Spooner, and Hans-Hermann Hoppe.

Minarchism

The libertarian perspective consistent with the principles of this philosophy to the second greatest degree advocates minimal state government and is called minarchism. Here, the government is indeed a legitimate institution, but the sole function of the state is to protect the persons and property rights of its citizens. To that end there are

three and only three legitimate governmental institutions: One, armies to protect domestic inhabitants against foreign aggression. (Not to become policeman to the world; not to engage in imperialistic ventures here, there, and everywhere; solely for *defense*.) Think in terms of a gigantic, ferocious, and very powerful Coast Guard. The second justified institution is police, to protect us from local violators of the NAP: to stop murderers, rapists, thieves, etc. (not to arrest victimless criminals for "crimes" involving consenting adult behavior such as pertaining to drugs, pornography, prostitution, gambling). Third, courts to determine guilt or innocence of those accused of real crimes and to address contract disputes. Minarchists strictly limit government to these three institutions.

Key contributors to this perspective are Robert Nozick, Ayn Rand, and Ludwig von Mises.

Classical Liberals, or Constitutionalists

Next in the libertarian pantheon in terms of consistency to principle are the Classical Liberals or Constitutionalists. To the three functions of government, advocates of this version of libertarianism would add a few more: dealing with communicable diseases and storms and floods and providing "public goods" such as roads and utilities like gas, electric, water, sewage removal, etc. Some in this category would include anything mentioned in the US constitution, such as post offices and coinage. Ron Paul is perhaps the most famous person connected with this viewpoint.

Free Market Supporters

Here, there are lots of compromises: health, education, welfare, externalities, antitrust policy, and monetary crankism (e.g., rejection of the gold standard). So much so that there are questions as to whether or not this viewpoint can be considered libertarian at all. On the other hand, people associated with this perspective are staunch advocates of the free enterprise system and adamantly oppose government incursions such as rent control, minimum wages, tariffs, corporate welfare, crony capitalism, and entry restrictions for business. They favor the profit and loss system. Those

associated with this point of view include Milton Friedman, Friedrich Hayek, and Rand Paul.

I now move to a discussion of anarcho-capitalism, the purest form of libertarianism, since its indictment of the government is the most accurate. That is, the state really is an organized means of NAP violation. It is the only institution in all of society which may legally initiate violence against innocent people. As such, it deserves intensive analysis, in any effort to understand what government is all about.

Let us start off with Rothbard (1977) who avers that it is important that any libertarian

... hates the existing American State or the State per se, hates it deep in his belly as a predatory gang of robbers, enslavers, and murderers... the State – *any* State – is a predatory gang of criminals... the radical (libertarian) regards the State as our mortal enemy, which must be hacked away at wherever and whenever we can. To the radical libertarian, we must take any and every opportunity to chop away at the State, whether it's to reduce or abolish a tax, a budget appropriation, or a regulatory power. And the radical libertarian is insatiable in this appetite until the State has been abolished...

There are those who would quarrel with such an analysis and maintain that the government is really a voluntary organization, like a club. And, just as in the case of the golf, or tennis, or chess club, every member must pay his dues. In this case that would be taxation. But Schumpeter (1942, p. 198) puts paid to this excuse for brutality: “The theory which construes taxes on the analogy of club dues or of the purchase of the services of, say, a doctor only proves how far removed this part of the social science is from scientific habits of mind.” The point is there is simply no contract, “social,” or otherwise that binds people together. Only a very few people signed the declaration of independence. How can it be binding on an entire populace?

But did not three-quarters of the states ratify the national government in the case of the United States, by a majority in each case? This cannot be denied. But how does that bind the people in the states that did not ratify this document? How does it obligate those who did not vote or voted against the proposal? No rational court would hold a person responsible for a commercial contract

to which he did not agree, just because others, even many others, did so.

Suppose a slave master allows his property to vote on overseer Goodie, who will only beat them once per week, and overseer Baddie who will do so three times per day. The slaves opt for the former. Does this mean these poor unfortunates have agreed to be slaves and that they agree to have Goodie whip them? Of course not. Spooner (1966[1870]) is justly famous in libertarian circles for making just this point. He also disposed of the argument that tax compliance demonstrates agreement with government overseers. We also pay the highwayman when he holds a gun on us. This hardly provides evidence that our “transaction” with him is a voluntary one, and unless it is the ethical case for the state has no foundation.

What about the following argument: “If you don’t like it here, under this government, you are free to leave. The fact that you stay indicates voluntary agreement with these institutional arrangements.” Stuff and nonsense. One difficulty with this is that it argues in a circle. It presupposes the truth of that very thing that is under debate, namely, that the government has a right to kick people out of “its” territory if they do not comply with its onerous regulations. If we jettison this basic unproven premise, then there is no case for inviting people to “leave” who object to statist depredations. Another difficulty is that there were people living on the territory taken over by the government before it was set up. This is necessarily so. There had to be people in existence to set up a government before it was operational. Thus, these people were there *before* the advent of the government. They could with as much justification say to those in the process of forming a government: “If you don’t like it here, under our present state of free market anarchism, you are free to leave.” Why must those who object to this obnoxious institution be the ones who must leave, on logical grounds?

Let us attempt to further “hack away” at this truly evil type of organization. It is said that the reason we need a government is because two individuals, or groups of people, will fight each other without such an institution, and, if allowed, this would create chaos. Or bullies would prey on

the weak. There are great difficulties with this view. First of all, as a matter of historical fact (Rummel 1994), governments have murdered more than 150 millions of their own citizens in the last century, and this is apart from the wars they have fomented with each other. And this statistic does not even include some 40,000 people slaughtered on US government highways (Block 2009). Gangsters and bullies are simply not in the same category. Secondly, a logical implication of national government is world government. If we need a national government because people in that territory will fight with each other, then we need a world government since national governments will (and most certainly have) fight with each other. The problem with world government is of course that there will be nowhere to run when and if such an institution turns totalitarian. But if world government is rejected, so, then, must the same fate await national government, since there is a perfect analogy between the two cases.

Here is yet another attempt to justify the unjustifiable: the government gives us goods and services in return for the taxation it imposes on us. This, too, cannot be denied. The state provides a myriad of goods and services, ranging from battleships to postal service, from welfare to food stamps, from bailouts to Detroit, to Wall Street, and to numerous points in between. But suppose a philosophical mugger who is willing to engage in discourse were to hold up an innocent man. The latter objects on the ground that the robber is not providing any goods or services, whereupon the thief agrees with his victim, amends his evil ways, and offers him the services of a smile or a snarl, and, also, a paper clip or a rubber band. The point is, it is entirely irrelevant whether or not the private or governmental burglar offers his prey anything in return for his depredations or not. The only relevant issue is if the citizen did *agree* to the interaction or not, and, as we have seen, not only is there no evidence that everyone unanimously agreed, but also there *could not* have been any such occurrence. For, if it did, if, that is, all the citizens welcomed the government, that would be a logical contradiction. For government is defined as that institution about

which everyone most patently *did not* agree. If there were an organization to which all participants voluntarily subscribed, it would no longer be a government! Rather, it would be some sort of private institution. A very large one, but similar in all essential regards to a business firm. For the latter embodies “capitalists acts between consenting adults” in the felicitous phraseology of Nozick (1974, p. 163).

Cross-References

► [Organization](#)

References

- Block WE (2009) The privatization of roads and highways: human and economic factors. The Mises Institute, Auburn. http://www.amazon.com/Privatization-Roads-And-Highways-Factors/dp/1279887303/ref=sr_1_1?s=books&ie=UTF8&qid=1336605800&sr=1-1
- Nozick R (1974) Anarchy, state and utopia. Basic Books, New York
- Rothbard MN (1977) Do you hate the state?. In: The libertarian forum, Joseph Peden, New York vol 10, no 7. <http://www.lewrockwell.com/rothbard/rothbard75.html>
- Rummel RJ (1994) Death by government. Transaction, New Brunswick. <http://www.hawaii.edu/powerkills/NOTE1.HTM>; <http://www.hawaii.edu/powerkills/20TH.HTM>
- Schumpeter JA (1942) Capitalism, socialism and democracy. Harper, New York
- Spooner L (1966[1870]) No treason: the constitution of no authority and a letter to Thomas F. Bayard. Rampart College, Larkspur. <http://jim.com/treason.htm>

Government Failure

Giuseppe Di Vita
Department of Economics and Business,
University of Catania, Catania, Italy

Abstract

The analysis begins from the origin of the concept of market failure and continues with an explanation of the reasons why a

measure of political economy could worsen market allocation and the level of social welfare. Mention is made of the differences between Europe and the United States regarding the weight of the State in the economy and of the efforts made by research to give a quantitative measure to government failure.

Definition

Government failure occurs when a measure of economic policy or the inactivity of the government worsens the market allocation of resources reducing economic welfare.

Government Failure

From a reading of the first few chapters of textbooks of *Principles of Economics*, we have learned that under the assumption of a perfect competitive market, for each good (or bad) exchanged, the economy achieves its maximum level of welfare as a result of the Smithian theory of the “invisible hand” (Smith 1776). This means that in such a Pareto-optimal situation, there is no room for economic policy measures that, if adopted, may only constitute an obstacle to the proper functioning of the market.

Despite this unlimited confidence in the market virtues, the same Adam Smith was forced to recognize the limits of the market, regarding national defense, justice, and public services (i.e., waterworks, hospitals, schools, roads, etc.), for which there is no profit in producing at an individual level, because the costs are always greater than the revenues (Smith 1776). Although one of the fathers of the modern economic theory recognized the limits of the market, the dominant view, up until the 1940s, was that the government should not interfere with the market to modify or change market allocation or income distribution.

The Great Depression started in the late 1930s and the subsequent Keynesian theory (Keynes 1936) was further instrumental in destroying the myth of the free market. After the Second World War, the economic theory started to consider than

before the existence and causes of market failure more seriously (Bator 1958) and saw the need for some government intervention, the so-called visible hand (Chandler 1977), to improve market allocation and raise the level of social welfare.

During the second half of the last century, the broad consensus regarding the positive consequences of adopting economic policy to correct the effects of market failures on economic welfare Coase (1964) introduced into the economic literature the expression “government failure,” as opposed to “market failure,” to express some concern about the continual assumption of Pareto-improving consequences of economic policy measures (Posner 1969, 1974). (Ronald Coase (1964) agreed with the approach which evaluated the effects of economic policy measures comparing regulated industries with industries not subject to regulation, although this approach cannot be followed because it is difficult to find industries that are comparable regardless of the degree of regulation.)

Coase drew inspiration from the talk of Professor Roger C. Cramton, a law scholar, held in Boston during the Seventy-Sixth Annual Meeting of the American Economic Association: Cramton stated that lawyers “. . . focus on the fact that public officials and tribunals are going to be fallible at best and incompetent or abusive at worst. . . .” Coase (1964) remarks that this statement is completely opposed to the assumption that the economists make about the government. In the economic theory, the government is seen as a benevolent planner who wants to correct the defects of the invisible hand, to improve market allocation, and to raise the level of welfare. The opinion of the government expressed by Professor Crampton would probably have been considered provocative, but for the economist, it was an alarm bell underlining the necessity to rethink the assumption about the government bodies and the impact of economic policy measures on the level of welfare. (In general, about the measure of economic policy, we can report the words of Milton Friedman (1975): “. . . The government solution to a problem is usually as bad as the problem and very often makes the problem worse. . . .”)

Moreover, Coase (1964) emphasizes the inability of the government to supply an immediate answer to the changes in economic conditions and the crucial role of “detailed knowledge” (or information) of the economic phenomenon addressed by a measure of economic policy, as an ineluctable condition of government intervention. (Winston (2006) has recently affirmed that government failure “. . . arises when government has created inefficiencies because it should not have intervened [in the market] in the first place, or when it could have solved a given problem or set of problems more efficiently, that is, by generating greater net benefits. . .”; such a definition emphasizes the inability of the government to improve market allocation.)

It is possible to affirm that a government policy is worth adopting in the presence of a market failure that is a source of not negligible social costs (see Datta-Chaudhuri (1990), Krueger (1990), Wallis and Dollery (1999, 2001), and Wiesner (1998) about the policy measures adopted in developing countries, since the 1960s, as sources of economic inefficiency which worsen the welfare level) where the government policy is at least improving market performance and efficiently correcting the market failure and optimizing the economic welfare (for normative economic theories of government failure, see Dollery and Worthington 1996).

The first possible source of government failure is the dynamic nature of an economic process such that the government in charge cannot commit itself for future measures which will be adopted by the subsequent government. (The dynamic nature of the economic process makes it necessary to consider the discount rate in the analysis: this is a task that we leave to future and more detailed analysis.) Stiglitz (1998) offers the example of hydroelectricity plants to produce electricity which need to be subsidized for a period of time longer than the duration of the legislature, while the hydroelectricity industries have no guarantee that successive governments will leave unaltered the subsidies and the legislation. (The inability to make commitments causes another set of inefficiencies: the cost of creating next-best credibility-enhancing mechanisms. While those in the

government at one date cannot commit future governments, they can affect the transaction costs of reversing policies (Stiglitz 1998, p. 10).)

The economic policy measures worsen market allocation whenever there is imperfect information between the policy maker and the private parties involved. Imperfect information may represent an obstacle to improving Pareto efficiency, if the policy maker does not have access to all the relevant information to adopt the correct policy measure or establish the magnitude of the measure (e.g., the amount of the rate of taxation).

The third source of government failure is the so-called destructive competition, a process that occurs in the presence of imperfect competition. Firms can get ahead not just by producing a better product at lower costs but also by raising the costs of their rivals (Salop and Scheffman 1983). Destructive competition is a zero-sum game where the profit of one is at the expense of another (see, e.g., the political games, with positions to be won or lost). This means that under the imperfect competition market structure, the liberalization of the market or the inactivity of the policy maker (Orbach 2013), as a result of a rational choice, determines the achievement of a suboptimal level of social welfare.

Another source of government failure is uncertainty about the consequences of policies: this may be due, as in the case of economic policy measures taken under asymmetric information, to the inability to predict the future impact of economic policies adopted now. A lesson we learned is that to be effective, economic policy measures should have well-defined objectives to achieve; otherwise, the measures taken by the government without a clear and precise statement of the objectives to be pursued and of the resources to meet them are just a way to leave unchanged the allocation of resources and the distribution of income (Cramton 1964). This means that some cases that may be attributable to the concept of government failure are just from the point of view of social welfare, because the government policies worsen market allocation but satisfy the real will of the policy maker to protect some established interests, without displeasing public opinion.

Finally, the last point we address is that government policies may be constrained in their action and worsen welfare due to the complexity of the economic measures that are hard for public opinion to understand, so Stiglitz (1998) empathizes the “simplicity constraint” in economic policies. Simple policies are easily explained to, and approved by, public opinion.

In a more interdisciplinary environment, considering laws, institutions, and economics, the theory of government failure has recently been enriched, affirming that it may also be due to the inefficient designing of rules for the economy that may be excessively specific (i.e., standard), too broad, conflicting, and unfair (Dolfsma 2011). (In general, on inefficient regulation as a source of government failure, see Posner (1974).)

The theories of government failure have played a different role in the United States and in the Old Continent because in the former they have been used to justify the limited adoption of economic policy measures and a reduced role of the state within the economy. In Europe, the limits of government action have been used to justify a reduction of the weight of the public sector in the economy (Vickers and Yarrow 1991).

Despite the differences between Europe and the United States in the role of the state in the economy that can be measured by the tax burden (Romer and Romer 2010), the government follows the business cycle in its policies because during crises, it is forced by public opinion to adopt stronger measures to curb the crisis (Rajan 2009), while during periods of prosperity, it is more prone to pander to the desire of firms for freedom.

Winston (2006), basing his theories on an empirical research limited to antitrust, monopsony policy, and economic regulation to curb market power, so-called social regulatory policies to correct imperfect information and externalities, and public production to provide socially desirable services, reached the conclusion that government intervention in markets has either been unnecessary or has missed significant opportunities to improve performance. (For recent development of the theory of “government failure,” see Mitchell and Simmons (1994) and Winston (2006).)

The economic policy measures to correct the limitations of markets have played a significant role in economic history (Datta-Chaudhuri 1990) and will probably continue to be useful in the future, but to improve market allocation, the policy maker should be aware of the limits of the “visible hand” (Winston 2012).

Cross-References

- [Government](#)
- [Market Failure: Analysis](#)
- [Market Failure: History](#)
- [Public Goods](#)

References

- Bator FM (1958) The anatomy of market failure. *Q J Econ* 72(3):351–379
- Chandler A (1977) *The visible hand*. Belknap, Cambridge, MA
- Coase R (1964) The regulated industries: discussion. *Am Econ Rev* 54(2):194–197
- Cramton RC (1964) The effectiveness of economic regulation: a legal view. *Am Econ Rev* 54(3):182–1191
- Datta-Chaudhuri M (1990) Market failure and government failure. *J Econ Perspect* 4(3):25–39
- Dolfsma W (2011) Government failure -four types. *J Econ Issues* XLV(3):593–604
- Dollery B, Worthington A (1996) The evaluation of public policy: normative economic theories of government failure. *J Interdiscip Econ* 7(1):27–39
- Friedman M (1975) *An economist's protest*, 2nd edn. Thomas Horton & Daughters, Sun Lakes
- Keynes JM (1936) *The general theory of employment, interest and money*. Macmillan, London
- Krueger AO (1990) Government failure in development. *J Econ Perspect* 4(3):9–23
- Mitchell WC, Simmons RT (1994) *Beyond politics: markets, welfare, and the failure of bureaucracy*. Westview Press, Boulder
- Orbach B (2013) What is government failure? *Yale J Regul* 30:44–56
- Posner RA (1969) Natural monopoly and its regulation. *Stanf Law Rev* 21:548–643
- Posner RA (1974) Theories of economic regulation. *Bell J Econ* 5(2):335–358
- Rajan R (2009) The credit crisis and cycle-proof regulation. *Fed Res Bank St Louis Rev* 91(5):397–402
- Romer CD, Romer DH (2010) The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. *Am Econ Rev* 100:763–801

- Salop SC, Scheffman DT (1983) Raising rivals' costs. *Am Econ Rev* 73(2):267–271
- Smith A (1776) *An inquiry into the nature and causes of the wealth of nations*. W. Strahan and T Cadell, London
- Stiglitz JE (1998) Distinguished lecture on economics in government: the private uses of public interests: incentives and institutions. *J Econ Perspect* 12:3–22
- Vickers J, Yarrow G (1991) Economic perspectives on privatization. *J Econ Perspect* 5(2):111–132
- Wallis J, Dollery B (1999) *Market failure, government failure, leadership and public policy*. St. Martin's Press, New York
- Wallis J, Dollery B (2001) Government failure, social capital and the appropriateness of the New Zealand model for public sector reform in developing countries. *World Dev* 29(2):245–263
- Wiesner E (1998) Transaction cost economics and public sector rent-seeking in developing countries: toward a theory of government failure. In: Picciotto R, Durán EW (eds) *Evaluation and development: the institutional dimension*, World Bank Editions. Transaction Publishers, Washington, DC
- Winston C (2006) Government failure versus market failure. *Microeconomics policy research and government performance*. R. R Donnelley, Harrisonburg
- Winston C (2012) Government policy for a partially deregulated industry: deregulate it fully. *Am Econ Rev* 102(3):391–395

heritage, and constitutional design. The lion's share of empirical work has employed measures of government quality based on perceptions. A debate exists about the convenience of conflating government quality with institutional quality. To measure institutional quality, it may be more advisable to generate measures or indicators which capture the extent to which institutions actually constrain governments.

Definition

Government quality is a multidimensional concept. Government quality is said to improve if the public sector does not distort the proper functioning of the private sector (respects property rights, does not overregulate), is an efficient administrator (low corruption, less bureaucracy, high tax compliance), provides public goods (education, health, infrastructures, etc.), and, finally, allows for and protects political freedom.

Government Quality

Over the years, economists have come to identify government or institutional quality as a fundamental cause of economic development. Early on, North and Thomas (1973) argued that the establishment of private property rights was instrumental for the rise of the West since the middle ages. Inspired by this insight, a great deal of scholarship has empirically explored the link between institutional quality and development (most notably Hall and Jones 1999; Acemoglu et al. 2001; 2002; Rodrik et al. 2004).

Growth-promoting institutions have at least two features: they must enforce property rights for a broad cross section of society so that all individuals have an incentive to invest, innovate, and take part in economic activity. And they must furnish some degree of equality of opportunity in society, including equality before the law (so that those with good investment opportunities can take advantage of them), and equal access to health and education (to promote human capital and to account for the fact that entrepreneurial initiative

Government Quality

Andreas P. Kyriacou
Universitat de Girona, Girona, Spain

Abstract

A growing body of work in economics has pointed towards the crucial importance of government quality for economic development. A major issue emerging in related empirical work has been the need to account for the confounding influence of other factors as well as the presence of reverse causality or the likelihood that development itself may facilitate governance. The potentially key role of government quality in explaining international differences in economic development has led many scholars to try and identify those factors that determine it. This research effort has brought forth the role of numerous variables including social heterogeneity, cultural

may be randomly distributed in the population and as such be independent of the distribution of property rights to existing resources) (Acemoglu et al. 2005).

A major concern which emerges when attempting to calibrate the impact of government quality on development is the issue of reverse causality or, in other words, the expectation that development facilitates institutional quality. In this vein, it has been argued that economic development makes better quality institutions more affordable (Islam and Montenegro 2002) and will tend to create a demand for better government (La Porta et al. 1999), perhaps because of income's positive effect on education, literacy, and depersonalized relationships (Treisman 2000). To deal with reverse causality, scholars have employed various instruments of government quality. Arguably the most notable has been settler mortality in the New World proposed by Acemoglu et al. (2001, 2002). They argue that in those colonies with high indigenous population densities where the disease environment was inimical to mass European settlement, extractive institutions, with few limits on the capacity of the state to expropriate individuals, were set up. And these institutions persist over time, not least because of the interest of the extractive elite to maintain their privileges. Conversely, in those colonies which were sparsely populated where, moreover, the disease environment was more benign, mass European settlement occurred leading to demands for good quality government (c.f. Sokoloff and Engerman 2000). Cross-country empirical work based on settler mortality suggests that government quality is a key factor explaining international differences in economic development, even trumping the direct effect of geography and trade (Rodrik et al. 2004).

The robustness of these finding has not gone unchallenged. On the one hand, the quality of the mortality data has been the subject of some debate among scholars (Albouy 2012; Acemoglu et al. 2012). On the other, it has been argued that the pattern of European settlement may have influenced growth through other, more fundamental, channels, namely, human capital. Thus, European colonizers took with them their human

capital, and in those colonies which were more extensively settled, these human capital endowments enhanced both institutional and productive capacities leading to stronger economic development (Glaeser et al. 2004).

Notwithstanding these qualifications, and based on theoretical and empirical work pointing towards a key role of institutions for the wealth of nations, scholars have tried to identify those factors that determine institutional quality, beyond the level of economic development itself and initial conditions framed by climate and factor endowments. In particular it has been variously argued that government quality may be determined by a range of variables which include inter-personal income inequalities (You and Khagram 2005; Glaeser et al. 2003; Sonin 2003), ethnic heterogeneity (Alesina et al. 1999; Glaeser and Saks 2006; Alesina and Zhuravskaya 2011), and ethnic group inequalities (Baldwin and Huber 2010; Kyriacou 2013). In addition, international differences in government quality may be due to cross-country cultural differences (Fisman and Miguel 2007; Licht et al. 2007) which may be partly the result of different religious heritages (La Porta et al. 1999; North et al. 2013). Government quality may also be responsive to constitutional design including electoral rules and government systems (Persson et al. 2003; Kunicová and Rose-Ackerman 2005) and the degree of fiscal and political decentralization (Fisman and Gatti 2002; Enikolopov and Zhuravskaya 2007).

Quantitative work in law, economics, political science, and business administration has employed government quality indicators from a range of sources. These include the World Bank's Worldwide Governance Indicators, Transparency International's Corruption Perception Index, the World Economic Forum's Executive Opinion Survey, the International Country Risk Guide (ICRG) published by the Political Risk Services Group, and several indicators elaborated by Freedom House and the Fraser Institute. Because of their country coverage and comparability over time, many experts have relied on indicators from the ICRG – available since 1984 – as well as the Worldwide Governance

Indicators, available since 1996. The ICRG data are based on in-house expert assessments of different dimensions of government quality including government stability, contract viability or expropriation, corruption, law and order, democratic accountability, and bureaucratic quality.

Unlike the ICRG indicators which are derived from a single source, the Worldwide Governance Indicators are obtained from a host of perception-based sources, including the ICRG and the other sources listed above. The Worldwide Governance Indicators measure government quality across six dimensions: (1) voice and accountability, (2) political stability and the absence of violence, (3) government effectiveness, (4) regulatory quality, (5) the rule of law, and (6) control of corruption. A key feature of the indicators is that all country scores are accompanied by standard errors which reflect the number of sources available for a country and the extent to which these sources agree with each other (Kaufmann et al. 2010). It has been pointed out that the indicators measuring the last four dimensions are very strongly correlated and, as such, they may be measuring the same thing (Langbein and Knack 2010). As a result, it makes sense to combine these four into an aggregate indicator in empirical work. Moreover, it may be useful to view the first two dimensions as contributing towards the latter four.

There is some debate about conflating institutional quality with government quality. According to Douglass North (1991, p. 97) “institutions are the humanly devised constraints that structure political, economic and social interaction.” From this perspective, institutions are constitutions and other laws, and the test of good institutions would imply the examination of how such constraints improve social welfare. Because empirical work has tended to equate institutional quality with government quality, scholars have measured the former by way of the perception-based indicators reviewed above. But government quality indicators typically measure outcomes rather than institutional constraints per se, and objective measures of institutional constraints are typically uncorrelated with measures of government quality (Glaeser et al. 2004). However, empirical efforts relating de jure institutional constraints with

social outcomes are likely to be undermined by the possibility that the measures or indicators used may not capture the extent to which constraints are actually binding (Voigt 2013).

References

- Acemoglu D, Johnson S, Robinson J (2001) The colonial origins of comparative development: an empirical investigation. *Am Econ Rev* 91:1369–1401
- Acemoglu D, Johnson S, Robinson J (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Q J Econ* 117:1231–1294
- Acemoglu D, Johnson S, Robinson J (2005) Institutions as a fundamental cause of long-run growth. In: Aghion P, Durlauf S (eds) *Handbook of economic growth*. Elsevier, Amsterdam, pp 385–472
- Acemoglu D, Johnson S, Robinson J (2012) The colonial origins of comparative development: an empirical investigation: reply. *Am Econ Rev* 102(6): 3077–3110
- Albouy D (2012) The colonial origins of comparative development: an empirical investigation: comment. *Am Econ Rev* 102(6):3059–3076
- Alesina A, Zhuravskaya E (2011) Segregation and the quality of government in a cross-section of countries. *Am Econ Rev* 101:1872–1911
- Alesina A, Baqir R, Easterly W (1999) Public goods and ethnic divisions. *Q J Econ* 114(4):1243–1284
- Baldwin K, Huber J (2010) Economic versus cultural differences: forms of ethnic diversity and public good provision. *Am Polit Sci Rev* 104(4):644–662
- Enikolopov R, Zhuravskaya E (2007) Decentralization and political institutions. *J Public Econ* 91:2261–2290
- Fisman R, Gatti R (2002) Decentralization and corruption: evidence across countries. *J Public Econ* 83: 325–345
- Fisman R, Miguel E (2007) Corruption, norms and legal enforcement. Evidence from diplomatic parking tickets. *J Polit Econ* 115(6):1020–1048
- Glaeser E, Saks R (2006) Corruption in America. *J Public Econ* 90:1053–1072
- Glaeser E, Scheinkman J, Shleifer A (2003) The injustice of inequality. *J Monet Econ* 50:199–222
- Glaeser E, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) Do institutions cause growth? *J Econ Growth* 9:271–303
- Hall R, Jones C (1999) Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114(1):83–116
- Islam R, Montenegro C (2002) What determines the quality of institutions? Background paper for the World Development Report 2002, *Building Institutions for Markets*, Washington DC
- Kaufmann D, Kraay A, Mastruzzi M (2010) The worldwide governance indicators: methodology and

analytical issues. World Bank Policy research paper no 5430, Washington DC

- Kunicová J, Rose-Ackerman S (2005) Electoral rules and constitutional structures as constraints on corruption. *Br J Polit Sci* 35(4):573–606
- Kyriacou A (2013) Ethnic group inequalities and governance: evidence from developing countries. *Kyklos* 66(1):78–101
- La Porta R, Lopez-de-Silanes F, Shleifer A, Vishny R (1999) The quality of government. *J Law Econ Organ* 15(1):222–279
- Langbein L, Knack S (2010) The worldwide World Governance Indicators: Six, one, or none? *J Dev Stud* 46(2):350–370
- Licht A, Goldsmith C, Schwartz S (2007) Culture rules: the foundations of the rule of law and other norms of governance. *J Comp Econ* 35:659–688
- North D (1991) Institutions. *J Econ Perspect* 5:97–112
- North D, Thomas R (1973) The rise of the western world. A new economic history. W.W. Norton, New York
- North C, Orman W, Gwin C (2013) Religion, corruption, and the rule of law. *J Money Credit Bank* 45(5):757–779
- Persson T, Tabellini G, Trebbi F (2003) Electoral rules and corruption. *J Eur Econ Assoc* 1(4):958–989
- Rodrik D, Subramanian A, Trebbi F (2004) Institutions rule: the primacy of institutions over geography and integration in economic development. *J Econ Growth* 9:131–165
- Sokoloff K, Engerman S (2000) History lessons: institutions, factor endowments, and paths of development in the new world. *J Econ Perspect* 14:217–232
- Sonin K (2003) Why the rich may prefer poor protection of property rights. *J Comp Econ* 31:715–731
- Treisman D (2000) The causes of corruption: a cross-national study. *J Public Econ* 76:399–457
- Voigt S (2013) How (not) to measure institutions. *J Inst Econ* 9(1):1–26
- You J, Khagram S (2005) A comparative study of inequality and corruption. *Am Sociol Rev* 70:136–157

Greece: Ancient Greece

Michel S. Zouboulakis
Department of Economics, University of
Thessaly, Volos, Greece

Abstract

The Ancient Greek miracle is the synthesis of institutionally constrained free citizens seeking for their personal improvement within a democratically evolving law structure which guarantees public interest and social justice.

Definition

What we call ancient Greece was actually a large number of dispersed and independent city-states with a common language, religion, and social customs and yet with quite distinct political, social, and economic institutions. Chronologically, what we call Ancient Greek world extends from the eighth century BC to the second century AD. Geographically, the Ancient Greek world covered parts of Southern Italy and Sicily in the west, as well as the entire coast on the other side of the Aegean and the Black Sea, and many other parts of the Mediterranean, including North Africa, Southern France, and Spain. Our purpose here is to focus on a restricted part of Ancient Greek social and economic institutions, saying the minimum about the political institutions and nothing about the informal ones, meaning the values and social norms which shaped this glorious era of the western civilization.

Population and Citizenship

Cities were organized around a well-defined geographical center, and their population size varied from 250,000 in fifth-century Athens to 3,000 in the city-state of the island of Aegina. A common characteristic among all Greek cities was that not all the inhabitants were recognized as citizens with full political rights and civil duties. In general, citizens were free males above the age of 18, having completed their military duties. But not all adult free citizens, even if they were born in the city, had full citizenship. Noncitizens who were born free or were freed in their lifetime were often called and had military and tax obligations. An important part of every Greek city consisted of slaves, in an analogy (very different from city to city) one citizen for every two or three slaves. Social stratification differed from one city-state to another, although there were many common characteristics. In Athens, Solon introduced his institutional reforms in 594 BC, based on the division of citizens according to their income. Citizens, and therefore their obligations toward their city, were defined according to their

production: the *pentakosiomedimnoi* producing more than 500 *medimnos* (approximately 260,000 lit.); the *ippeis*, producing more than 300 *medimnos* or 156,000 lit.; the *zeugitai*, with more than 200 *medimnos* or 104,000 lit.; and finally the *thetes*, with less than 200 per year. Among these four property classes, the poorest had limited public rights; the *thetes* could vote in the Assembly but were excluded from public office. The richer classes provided the main military effort, as cavaliers (*ippeis*) or as heavily armed foot soldiers (*hoplites*). The situation in Sparta was different. Descendants of Spartan citizens got full citizenship at the age of 18, having received a proper military education called *agoge*. Other inhabitants were the *perioikoi*, who were free to live in Spartan territory but were noncitizens, and the *helots*, the state-owned serfs (Glötz 1928; Finley 1964; Cartledge 1993; Sakellariou 1999).

Political Institutions

Politically, the majority of Greek cities were ruled by aristocracies, oligarchies, and tyrannies in the eighth, seventh, and sixth centuries and oligarchies and democracies in the fifth and fourth. In Corinth, Argos, Sparta, Thebes, and many other cities, oligarchy was based on property. Aristotle described four types of oligarchy from the most autocratic cities of Thessaly in central Greece to the more open cities of Thebes and Orchomenos in Boeotia; participation in the privileged ruling few (*oligoi*) was only a matter of middle-range personal revenue among full citizens. Oligarchic cities such as Croton, Region, Acragas, and Colophon had very large Assemblies (*ecclesiae*) consisting of 1,000 wealthy citizens who legislated by open vote. Sparta had a smaller “executive” body of 30 elders (*gerousia*) advising the hereditary two kings (Glötz 1928; Cartledge 1993).

Athenian direct democracy of the fifth century was an exception that followed well-established formal rules and had two collective bodies of governance: the *Boulé*, a body of 500 citizens selected by draw for a year term from the ten

counties of Attica, met every day, except on holy festivals, to administrate the city. All the citizens (i.e., all 42,000 in 431 BC) had the right to participate in the general assembly (*Ecclesiae*) 10–40 times in a year at the *Pnyx*, a hill opposite to the Acropolis, to decide freely on all the important public matters of the city and to legislate. The farther one lived, the less often he was present to vote. So, distant communes of Attica tended to be underrepresented, this being one of the main problems of Athenian democracy, together with the lack of material motives to participate. Poor citizens would be less willing to participate in the *Ecclesiae*, not affording to lose one day’s wages. After the reforms of Pericles in 460 BC, citizens were paid a small remuneration, little less than a day’s wage for their participation in the Assembly (Glötz 1928; Mossé 1995). But, there were clear advantages as well. Since the voting rule was simple majority by show of hands, it was quite difficult for organized groups to carry the day if they could not convince the majority of the citizens, many of whom changed their minds after hearing the orators’ arguments. There were no formal political parties in the modern sense, but there certainly were political groups representing interests, gathered around some charismatic leaders, such as Cleisthenes, Themistocles, Ephialtes, and Pericles for the democratic “party” and Aristides, Kimon, and Nikias for the aristocratic or rather conservative “party.” The latter was mainly supported by the wealthier landowners and also by medium holders whose revenues came from agriculture, while the farmers and breeders, as well as the traders, artisans, and sailors, supported the democratic side (Finley 1964).

Private and Public Finance and Institutionalized Benevolence

In the great majority of archaic and classical Greek, city-state inhabitants lived mainly of agriculture. Even the rich Athens, Corinth, and Chalkis were mainly inward-looking city-states, but they all had a significant export trade and a commercial fleet as well. Archeological finds in

Italy and Asia Minor prove the flourishing trade of other Greek naval city-states such as Miletus, Ephesus, Colophon, Sybaris, Syracuse, Acragas, and even Massalia (Marseilles). Since the seventh century, golden, silver, and electron coins contributed to the rapid monetization of trade and economic transactions in general. Economic development rapidly changed the life of these naval city-states, while mainland agricultural cities in Boeotia and the Peloponnese declined slowly in power and wealth (Glotz 1928). Individual banks (*trapezes*) appeared early enough to assist commerce with credit and change facilities. Although the institution of the commercial bank appeared only in the Renaissance, ancient capital owners actually acted as bankers lending money with interest to farmers and naval traders (Bitros and Karayiannis 2010). The fact that bankers differentiated interest rates according to the purpose of the loan, offering loans with a nominal rate of 12 % to landowners and up to 30 % to shipowners, proves that they had a clear notion of risk.

Classical Athens had a highly developed system of property rights and duties. Despite property-based distinctions of social rank, inherent to all hierarchical societies, tax burdens and public and military obligations were specified according to each one's income situation. Hence, the city-state of Athens was able to establish an exact contract with its richest citizens in order to ensure the production of both private and public goods and to secure a permanent source of state revenue. The city's public revenues came from rents of public land, custom duties, fines, and war booty. In many Greek cities, rich citizens contributed in a semi-voluntary basis to the public expenditure according to the social institution of *liturgy*. The most important liturgies were the *choregia*, the organization and finance of theatrical productions; the *gymnasiarchia*, the financial support and supervision of athletic training; and the *estiasis*, the organization of public religious feasts (Davies 1981; Gabrielsen 1994; Mossé 1995). Hence, when Athens decided to build 100 ships by means of the silver discovered at Lavrion, to defend the city from the second Persian invasion, an institutional reform introduced a new liturgy, called *trierarchy*. It consisted of

entrusting to the 100 richest citizens the charge of building 100 more new ships so that the Athenian fleet was comprised of 200 triremes, equivalent to the two thirds of the fleet that defeated the Persians in the naval battle of Salamis in 480 BC (Kyriazis and Zouboulakis 2004).

After Salamis, the *trierarchy* system was further developed and refined. Under its new form, the city itself financed the construction of warships but entrusted their maintenance to rich citizens, elected each year. Eligibility was based on the formal rule of property qualification (Glotz 1928; Davies 1981). If an Athenian citizen was able to prove before the Court that there was a richer citizen than himself, then he could avoid the duty. If the supposedly richer man disagreed with the terms, he had to choose to refuse the obligation and exchange properties with the challenger or to undertake the liturgy. This procedure was called *antidosis* (Gabrielsen 1994; Carmichael 1997). A refusal to discharge any liturgy entailed punishment. While the city was safeguarded against fraud by the liturgists by mortgaging their land assets and houses, rich Athenians considered it a duty and an honor to be chosen to finance and manage the production of public goods and services (Lyttkens 1997). To avoid that liturgies do not fall on the same persons for consecutive years, after 378 BC taxpayers were organized in 20 groups called *symmoriae* to share the burden (Mossé 1995).

The Athenian tax system was comprised also of market dues, a tax on sales concluded before state officials, an annual charge of 12 drachmae on *metics* and of 0.5 drachmae on slaves and freedmen, a tax on those who pursued certain callings requiring special supervision (such as oracle mongers, jugglers, and prostitutes), as well as import and export duties. Direct taxation levied on land during the sixth century BC was abolished because it was regarded as an infringement on civil liberty except in times of crises, like the Peloponnesian War, when the Athenians imposed upon themselves a special property war tax *eisphora*. Still, the liturgies functioned as a kind of progressive direct wealth taxation. Redistribution followed indirectly, since the taking over of public goods by the richest citizens liberated

public means to be spent on other purposes, such as a “poor-relief” system, granting assistance to the incapacitated and maintaining the children of those who fell at war, until they came of age.

As it was argued (Ackroyd 1992), lessons from Classical Greece demonstrate that private interests are not incompatible with social justice and public interest insofar as these values are not imposed by a central public power but are enforced by the Law decided by the citizens themselves (ΝΟΜΟΣ ΒΑΣΙΛΕΥΣ).

Cross-References

- ▶ [Constitutional Evolution in Ancient Athens](#)
- ▶ [Development and Property Rights](#)
- ▶ [Fiscal System](#)
- ▶ [Institution](#)
- ▶ [Public Enforcement](#)

References

- Ackroyd P (1992) Greek lessons for property right arrangements: justice and nature protection. *Am J Econ Sociol* 51:19–26
- Bitros G, Karayiannis AD (2010) Morality, institutions and the wealth of nations. Some lessons from Ancient Greece. *Eur J Polit Econ* 26:68–81
- Carmichael C (1997) Public munificence for private benefit: liturgies in Classical Athens. *Econ Inq* 35: 261–270
- Cartledge P (1993) *The Greeks*. Oxford University Press, Oxford
- Davies JK (1981) *Wealth and the power of wealth in Classical Athens*. Arno Press, New York
- Finley M (1964) *The Ancient Greeks*. La Découverte, Paris, (French translation), 1984
- Gabrielsen V (1994) *Financing the Athenian fleet: public taxation and social relations*. Johns Hopkins University Press, Baltimore
- Glötz G (1928) *La cité grecque*. Albin Michel, Paris, 1953
- Kyriazis N, Zouboulakis M (2004) Democracy, Sea power and institutional change: an economic analysis of the Athenian Naval Law. *Eur J Law Econ* 17: 117–132
- Lyttkens CH (1997) A rational actor perspective on the origins of liturgies in Ancient Greece. *J Inst Theor Econ* 153:462–484
- Mossé C (1995) *Politique et société en Grèce ancienne. Le “modèle” athénien*. Aubier, Paris
- Sakellariou MB (1999) *The Athenian democracy*. University of Crete Press, Heraklion

Greece: Modern Greece 1821–2018, A political History of

Aristides Hatzis

Department of History and Philosophy of Science, National and Kapodistrian University of Athens, Athens, Greece

Definition

Modern Greece has a history of almost two centuries. During these centuries, the country managed to move from the backwaters of Europe to a prosperous liberal democracy before economic crisis hit the country hard in 2010. Greece was founded after a War of Independence from the Ottoman Empire that was based on liberal and democratic principles. This left a political legacy which led to universal male suffrage as early as 1844 and one of the longest parliamentary histories in Europe, despite the tumultuous political life and brief periods of authoritarian regimes. The nineteenth century was a period of a slow modernization of the country (in infrastructure and institutions) but it was also suffocated by “Megali Idea,” the irredentist dream of the enlargement of the Greek state to include all lands, under Ottoman rule, inhabited by large Greek-speaking populations. A great part of Megali Idea was realized in early twentieth century but the triumphs ended with a devastating catastrophe in 1922. Greek political elites were often incompetent and corrupt, but several reformist statesmen managed gradually to achieve convergence with other western European countries. Most importantly, they were very effective in steering Greece on the right (i.e., winning) side of history during every major European or Global conflict (Balkan Wars, World Wars, Cold War). Greece, after World War II and a ferocious Civil War, enjoyed one of the strongest, almost uninterrupted growth on a global level. This led to the accession to the European Communities in 1981 and later the Eurozone. Today, after 10 years of economic crisis and painful austerity,

Greece must meet one of the most difficult challenges: to achieve growth by adopting inclusive institutions.

Greeks in the Ottoman Empire

Greece became an independent state in 1830. Its independence was the result of a national uprising against the Ottoman Empire in the early nineteenth century. After the end of the classical era, Greece was controlled by empires, mostly by the Roman, the Eastern Roman or Byzantine and the Ottoman empires. The Roman and the Byzantine empires were strongly influenced by the ancient Greek civilization and from the seventh century on, Byzantine Empire was linguistically Hellenized.

After the fall of Constantinople, in 1453, the Ottoman Empire recognized the Christian Orthodox Church and the Patriarch of Constantinople as the spiritual but also the political leader of the Greek Christian Orthodox (*Rum*) community. The Christian Orthodox Church, which was dominated by a Greek speaking clergy, was granted several privileges, including some autonomy and judicial jurisdiction in certain private law disputes among Orthodox Christians.

As a result of the power and the great influence of the Ecumenical Patriarchate of Constantinople, at the beginning of the nineteenth century, the Greek Christian Orthodox Commonwealth was so prominent as to capture even high-ranking political stations in the Ottoman administration. Since the late seventeenth century, Greek Christian Orthodox laypeople, “Phanariotes,” who inhabited Phanare, the area around the seat of the Patriarch, had a significant political power. Phanariotes dominated the influential position of Grand Dragoman, the official interpreter for the Imperial Council, but also the position of the Prince (*Voivode*) of the semi-autonomous Danubian Principalities of Moldavia and Wallachia, i.e., modern-day Romania. Phanariotes, many rich merchants, several local communities, and the Church created, from seventeenth century on, a big number of local

but also influential small schools which became, until the nineteenth century, the standard educational units in the Balkan Christian Orthodox communities.

There was also a local Christian Orthodox landed class in Southern Greece, mostly in Peloponnese. The *Kodjabashis* in Turkish or *Proestoi* in Greek were the leaders of their communities. They were part of the Ottoman administration since they were usually responsible for tax collection. These local notables were supported by private armed groups, and they had direct connections to the Ottoman administration and the Patriarchate. They had to deal with dangerous bands of brigands (*klephts*) who had managed to control the mountainous areas of Peloponnese and Central Greece (*Roumeli*) defying Ottoman power and Christian Orthodox notables. In some areas of Central Greece, their power and influence were so significant as to be recruited by the Ottomans as militiamen (*Armatoloi*), supposedly to protect the area from their alter egos, the klephts. Nonetheless, they kept reversing their allegiances according to their interests by alternating between the two roles.

During the eighteenth century, Greek merchants operating in the Balkans and the Asia Minor but also in Central and Northern Europe, including Russia, and Greek shipowners living in small islands of the Aegean and operating in the Mediterranean and the Atlantic, managed to amass great wealth since they controlled a great part of the trade in the Ottoman empire. In the eighteenth and early nineteenth century, they exploited the vacuums in French and English maritime trade during the Second Hundred Years’ War (1714–1815), especially the Napoleonic Wars (1803–1815). One of the consequences of the end in European Wars in 1815 was idle capital and labor in the Greek peninsula, mostly in the islands.

The intellectual elite of the Orthodox Christians who spoke Greek was influenced by the eighteenth century enlightenment, earlier from the other elites in the Ottoman Empire. By the middle of the eighteenth century and in the first half of nineteenth century, major commercial, professional, and academic

(predominantly student) communities of Greek-speaking Orthodox Christians were formed in Vienna, Italian cities, Odessa, and other European metropolitan and commercial centers. Some clergymen and sons of wealthy merchants educated abroad were among the leading proponents of the enlightenment ideas, major works were translated into Greek, and schools teaching Greek were proliferated in key commercial centers. The backlash from the conservative ecclesiastical hierarchy after the French Revolution failed to stop the dissemination of ideas, even in mainland Greece. The half-baked Greek enlightenment undermined the authority of the Church, reconnected the Christian Orthodox elites with Ancient Greece and Western Europe, and created the fertile ground for a revolution that was not only nationalist but also democratic and liberal.

The Greek War of Independence (1821–1832)

The Greek War of Independence was not the first insurgency against the Ottoman Empire in the Balkans. Not even the first by Greek speaking people. The last major uprising by Greeks was the Orlov revolt (1770–1771), incited by Russia's Catherine the Great to distract the Ottomans during the Russo-Turkish War of 1768–1774 and it was rather easily suppressed by the Ottomans.

The French Revolution and the Napoleonic Wars were the events that triggered the Greek Revolution. Rigas Feraios, an intellectual and early revolutionary, tried to secure Napoleon's support for an insurgency in the Balkans but he was betrayed and he was arrested by the Austrians who delivered him to the Ottomans who tortured and executed him. He became the first national hero of modern Greece and he inspired similar revolutionary activities, especially in the Greek diaspora, between merchants, students, and some Phanariotes.

The Greek War of Independence was organized by unlikely revolutionaries: a group of small-time businessmen, with ties to freemasonry and Carbonarism, who founded *Filiki*

Etairia (Society of Friends) as a secret organization, on September 1814 in Odessa. By 1821, Filiki Etairia had hundreds of members in the Greek peninsula, mostly in the Peloponnese, the islands and the Constantinople. The plan was to incite a revolution in the Balkans (from Bucharest to Crete) but the two major points for the break out was first the Danubian Principalities and then the Peloponnese. The leader of the revolution was Prince Alexandros Ypsilantis, a major general of the Russian army and aide-de-camp of Tsar Alexander I.

Still, leading Greek intellectuals, like the liberal Adamantios Korais and statesmen, like Count Ioannis Kapodistrias, foreign minister of Russia from 1816 to 1822, were negative to the idea. They, respectively, believed that Greeks were not ready or that the international situation was extremely hostile to revolutions after the Congress of Vienna (1814–1815).

The Ypsilantis Campaign in Moldavia and Wallachia started on February of 1821 and failed after four excruciating months. Ypsilantis was arrested and imprisoned by the Austrian authorities. He failed to obtain the support of other Balkan peoples, the uprising was renounced by Russia and the revolutionaries were excommunicated by the Patriarchate in Constantinople.

Nonetheless, Greeks took advantage of the fact that Ottomans were distracted by the uprising of Ali Pasha of Ioannina, a powerful warlord who challenged the Ottoman Empire. The suppression of his uprising occupied formidable Ottoman armies for almost a year before and a year after the breakout of the Greek revolution. These forces, thus, were unavailable to quash the insurgents in the Greek peninsula, offering to the revolutionaries the necessary breathing space. Ali Pasha's insurgency was instrumental for the decision by Filiki Etairia leadership to start the revolution during the Spring of 1821.

Even though the Greek revolution was dangerous for the *status quo* that the Congress of Vienna had established in 1815, it generated conflicting sentiments in Europe. Klemens von Metternich, the powerful Chancellor of the Austrian Empire, was very hostile and suspicious

of the Greek rebels. He thought that this was not a purely local insurgency of aggrieved Christians against the cruel and corrupted Ottomans but the sperm of a broader revolution in the Balkans, a threat to multiethnic empires, a radical uprising. He was right. The Greek War of Independence stroke a chord with liberal thinkers and activists but also with romantic poets and writers. These were the Greeks after all, fighting to liberate their “sacred” land from the “uncivilized” Muslim conquerors. The first victories of the Greeks in Peloponnese, Central Greece and the sea (several islands offered a strong makeshift fleet from converted merchant vessels), the Ottoman atrocities against Christian populations (in retaliation for the revolution but also for Greek atrocities) and the intelligent way the revolutionaries presented themselves to Europe had two impressive results: the local insurgency became, almost instantaneously, an international event and scores of influential Europeans, poets (like Lord Byron), and intellectuals (like Jeremy Bentham) decided to support the cause financially or politically. This led to an impressive Philhellenic movement in Europe which was instrumental for the final liberation of Greeks.

The protagonists of the Greek War of Independence came under broad but not necessarily mutually exclusive, categories. Two important groups were:

- (a) The former klephts and *armatoloi* who were brave and shrewd enough to defeat the superior military forces that Ottomans sent to the south. Theodoros Kolokotronis, an ex-klepht, was the most important military leader who also tried to capitalize politically from his well-deserved fame and popularity. However, his view of the world was parochial and his behavior quite opportunistic. He failed to become the “Greek George Washington” because he lacked the American founder’s humility and the political sense to understand what the stakes were.
- (b) The westernized intellectuals, most of them vehicles of liberal and democratic ideas, were the ones who managed to take

over the Revolution. They were former Phanariotes or children of wealthy merchants who studied in Europe and came back to join the Revolution with a clear agenda: to transform the new country into a European constitutional democracy. Their leader was Alexandros Mavrokordatos. They passed, during the Revolution, three liberal and democratic constitutions, approved by national assemblies.

These two factions came to an eventual conflict which was far more multifaceted than a clash between liberals and traditionalists. After the initial military success of the Greeks, there was a ruthless struggle for domination which led to a civil war in two stages. The Ottomans found the opportunity to regroup, asking also the help of Muhammad Ali of Egypt who sent his son Ibrahim to Greece with cavalry and infantry of 20,000 well-trained, by European officers, Egyptian troops. Ibrahim, with the help of other Ottoman forces, managed to quench the revolution in the greatest part of central and southern Greece but his cruel tactics, the massacre of Missolonghi (April 1826) and geopolitical considerations, led to the intervention of three major European powers. The warships of United Kingdom, France, and Russia destroyed Ibrahim’s armada at Navarino bay in October 1827. It took another year for his forces to evacuate Peloponnese, under the pressure of a French expeditionary force. The independence of Greece was declared in February 1830 and with the London Protocol of August 1832, the Kingdom of Greece was established.

Instrumental to this result was the competition among the three Great Powers to influence the direction and the alliances of the new small state but mostly the smart policy of the British foreign minister, George Canning. He was the first to realize that Greece could be a natural and strategic ally in Eastern Mediterranean and the western-oriented Greek faction managed, in 1825, to persuade the revolutionaries to officially ask for the protection of Great Britain. From the mid-1820 (esp. after 1854) to 1947, Greece remained in the orbit of the overbearing British

Empire, being its staunchest ally but also benefiting from the strong political and diplomatic ties with the major power of the era.

Even though the intention of most Revolutionaries, as well as of the Great Powers, was for Greece to become a monarchy, Greeks elected, in 1827, as a President of the new independent state, Ioannis Kapodistrias. Kapodistrias arrived in Greece in early 1828 and realized that the only way to govern over the different factions was with a strong hand. He abolished the Constitution and he assumed dictatorial powers with the initial consent of most representatives in the revolutionary assembly. This was the end of the first Greek republic (1822–1828). However, Kapodistrias was the only Greek who could do the job (statecraft) adequately. His authoritarian reformism, especially his attempt to establish a centralized rule over local notables and warlords, led to his assassination in late September 1831. In less than 4 years, he managed to transform Greece from something resembling a state, with warlords and local notables reigning over poor farmers to a semblance of a European-oriented state. His political agenda was modernization. He organized the administration, he fought highwaymen and pirates, he organized tax authorities and the judiciary, and he tried to secure international loans. His most important legacy was land reform and the establishment of a rudimentary national education system.

Growing Pains: The New Kingdom (1833–1911)

Kapodistrias' assassination is a traumatic event in Greek history. His death led to 16 months of civil unrest and anarchy. So, when the Great Powers elected the Bavarian Prince Otto from the House of Wittelsbach for the throne of the new Kingdom in 1832, almost everyone in Greece received their decision with a relief. The 17-year-old Otto arrived in Greece in early 1833 with a Regency council who governed Greece for almost 3 years, until Otto reached majority. Both the Regency Council and Otto were autocrats. Despite the popularity of the young King, his

absolutism and the fact that he was a Roman Catholic, married to the Protestant and childless Amalia, were among the causes of resentment together with austerity measures he had to adopt, including suspension of benefits to war veterans, due to Greece's insolvency. This led to a bloodless insurgency on September 1843. The uprising was supported by the military garrison of Athens and several politicians. Otto had to yield to their demands. He promised to grant a constitution and to terminate the involvement of Bavarians in the administration. The new Constitution was proclaimed in March 1844, transforming Greece into a constitutional monarchy with a bicameral parliament. With the electoral law of March 18, 1844, Greece was the first country in the world to introduce universal male suffrage – nine out of ten male adult citizens obtained voting rights. However, Otto's destabilizing constitutional transgressions and his direct involvement in politics led to more bitterness and criticism. In October 1862, he was forced to abdicate and leave Greece.

The Bavarian legacy is mixed. The three decades were a period of establishing hierarchy and bureaucracy in the public administration and the army, of organizing education and judiciary, of modernization and Europeanization. One of the regents, the law professor Georg Ludwig von Maurer, was instrumental in the process of the adoption of modern European institutions and legal codes by the new state but also of the unpopular "nationalization" of the Greek Orthodox Church, even though he stayed in office for only 18 months. Otto had moved the capital from Nafplion to Athens for obvious symbolic reasons. During his reign, he adopted the irredentist policy of the "Great Idea" (*Megali Idea*), i.e., the enlargement of the Greek state to include all lands under Ottoman rule inhabited by large Greek-speaking populations, including Constantinople, southern Balkans, the Aegean Sea islands, Crete, Cyprus, and the western part of Asia Minor. From 1844 to 1922, irredentist nationalism became the dominant policy for Greece leading to impressive territorial expansions but also to a major catastrophe. One of Otto's major failures was to take advantage of

the Crimean War (1853–1856) in order to expand against the Ottoman Empire.

Otto's abdication was the result of an uprising against his reign. His successor was a Danish Prince from the House of Glücksburg who became King George I and reigned for half a century, from 1863 to 1913. He was a solid anglophile and willing to accept a democratic constitution. With the new constitution of 1864, Greece became one of the first parliamentary democracies in the world, especially when the democratic principle was reinforced in 1875 when popular sovereignty was guaranteed by the introduction of the constitutional principle that the government should enjoy the confidence of the Parliament.

The leading reformist political leader of the late nineteenth century was the liberal Charilaos Trikoupis. From 1875 to 1895, he dominated Greek politics. He became prime minister seven times and he governed more than a decade in total. He began his career fighting the constitutional transgressions of King George I. During his tenures, the infrastructure which modernized Greece was built, but it was funded by foreign loans, leading the country to bankruptcy in 1893. A few years later, in the summer of 1896, Athens hosted the first international Olympic Games in modern history but Greece had also to face an "unfortunate" war with Turkey in 1897.

From 1864 to 1881, Greek territory was enlarged peacefully. Britain ceded the Ionian islands to Greece in 1864 as a gift and rewarded Greece with Thessaly and part of Epirus, for not siding with Russia, despite the nationalistic temptation, in the Russo-Turkish War of 1877–1878. However, these minor territorial gains couldn't satisfy Megali Idea, in an era where the Eastern Question preoccupied the minds of governments and people in the Balkans. The evolving Cretan question (Cretans revolted almost uninterruptedly against the Ottoman authorities during the nineteenth century) pressured enormously the Greek governments and kindled nationalistic feelings. Nevertheless, British protection again ensured that Greek territory remained intact, even after the Greco-Turkish War of 1897 while Crete became autonomous in 1898 under Prince

George of Greece. But Greece's military defeat led to a national humiliation and the imposition of international control of Greek finances which led to an impressive fiscal consolidation.

Belle Époque was not so superb for Greece. The bitter military defeat and the bankruptcy led to the disillusionment of Greeks and resentment against the Palace. The Cretan issue was not considered resolved since Cretans themselves were not satisfied with autonomy, they wanted to be united with the "motherland." At the same time, a new antagonism begun in Ottoman Macedonia, mostly between Greeks and Bulgarians. From 1893 to 1908 (when the Young Turks Revolution took place), the two countries used armed propaganda, cultural and religious influence to attract and subdue the mixed-ethnically people living in Macedonia. Additional chronic problems, during the nineteenth century, were political corruption, rigged elections, a powerful clientelist system, and a dysfunctional bureaucracy.

The turmoil in Macedonia but mostly in Crete and the initial success of the Young Turks led to the first military coup in the twentieth century Greece. The pronunciamiento was bloodless, organized by young and politically inexperienced military officers without a clear political agenda, in mid-August 1909. The government capitulated to their demands, but a political stalemate was the inevitable outcome of their indecisiveness. Finally, they offered the political leadership to Eleftherios Venizelos, a middle-aged Cretan politician and former revolutionary. Venizelos arrived in Greece and decided, reluctantly at first, to play the political game by levelling first the play field. He was a political genius and the greatest statesman in modern Greek history. He dominated Greek politics for 25 years with his impressive successes but also failures and his polarizing figure. From 1910 to 1915, he managed to gain the trust of King George I, he was appointed Prime Minister, he won election after election, he founded the first modern Greek political party (the Liberal Party), he amended the constitution (in 1911), he enforced many structural reforms in almost every area: from the reorganization and training of the army to education.

The Decade of Triumph and Tragedy (1912–1922)

The disappointment that followed the promise of the Young Turks Revolution for equality and democratic representation of the minorities and the ripening of the Eastern Question were two of the reasons behind the Balkan Wars of 1912–1913. A military alliance of Bulgaria, Serbia, Montenegro, and Greece attacked the Ottoman Empire in October 1912. Bulgarians and Serbians, with their strong armies, never believed that the relatively weak Greek army would be so successful and fast as to reach the finishing line first by capturing Thessaloniki, the bone of contention between the Balkan allies. Bulgaria was so disillusioned by the unforeseen Greek success as to attack Greece and Serbia after the first stage of the Balkan Wars. The Greek army was able to defeat the strong Bulgarian army and gain some additional territory. The greatest part (51%) of the apple of discord, Macedonia, became a part of the Greek state. This was the result of Venizelos' political and diplomatic genius. But he shared the credit with Prince Constantine, the son and heir of King George I who was also the military commander in chief. Venizelos and Constantine's relationship at the end of the Balkan wars was overcast by their disagreement as to the strategic objectives of the Greek expansion. King George managed to control his insolent son and to satisfy his successful prime minister, minimizing the cost of their divergence. He moved temporarily to Thessaloniki to symbolize the accession of the new territory to Greece. A few months before his golden jubilee in 1913, he was assassinated by a drunkard with anarchist sympathies. The box of Pandora was now open, as his son, became King Constantine I.

Greece had managed to double its territory and population by acquiring southern Macedonia, southern Epirus, most Aegean islands, and Crete. The integration of these new territories with sizable ethnic and religious minorities and the efficient administration was not an easy task for the small Kingdom. Nevertheless, for almost a century, Greeks waited impatiently for the Megali Idea to be realized and now, after all these years and many disappointments, everything signified that they

were on a roll. They saw Venizelos as the one who made it happen politically, but Constantine was also extremely popular as a successful military leader and a symbol: he was the first King born in Greece and raised as a Greek-Orthodox, the one destined to restore Greece to greatness.

The confrontation between the two seems inevitable with hindsight. They both were strong-minded, but their views differed. Venizelos was a liberal, a staunch anglophile and he represented the interests of the most dynamic part of the new bourgeoisie. Constantine was a believer in the divine rights of monarchy, with populist impulses and rather unrelenting in his prejudices. But worst of all, Constantine was strongly pro-German. When the first World War I broke out, their conflict evolved into a National Schism, a traumatic experience with devastating consequences. From 1915 to 1936, Greece was bitterly divided between anti-Venizelists and Venizelists and eventually (after 1924) to royalists and republicans.

It was a conflict of foreign policy. But it was also a constitutional crisis. Venizelos insisted that Greece should enter the Great War on the side of Entente but Constantine was adamantly against, he favored neutrality but he also schemed in the back rooms for the benefit of the Central Powers. Venizelos resigned but even though he triumphed at the national elections, Constantine didn't relent. He saw foreign policy as his royal prerogative. Venizelos resigned for the second time. Entente pressured ruthlessly the royal governments that followed to submit to its demands and, when German-escorted Bulgarian troops seized part of the Greek Macedonia, Venizelos decided to set up a provisional government in Thessaloniki with the support of the French army. Greece was torn. There was even an armed confrontation in the streets of Athens between the army of the royalist government and French forces in November 1916, which led to an even deeper wedge between royalists and Venizelists. A naval blockade by the Allies made Constantine leave Athens (he didn't abdicate) together with his first-born son and heir, George, in June 1917. Venizelos returned to Athens and assumed power. Alexander, the second-born son of Constantine, became a kind of "interim" King. He was ideal for the job. A malleable individual, he was easily handled by

Venizelos. Greece entered the war on the side of the Allies at the last stage of it helping them in their offensive in May 1918 that led to the capitulation of Bulgaria.

The fact that Bulgaria but also the Ottoman Empire sided with the Central Powers who lost the war gave Greece an enormous diplomatic advantage. Venizelos reinforced this advantage by sending Greek troops, in 1919, to fight the Red Army together with the White Forces at the multinational military expedition in Ukraine. Venizelos proved himself the most credible ally in the Balkans and he was splendidly rewarded in the Treaty of Sèvres of 1920. Western and Eastern Thrace were delivered to Greece. The Greek army reached the outskirts of Constantinople and landed in western Asia Minor, Smyrna (today Izmir) and a large surrounding area with the objective to annex it to Greece, with the approval of the Allies, after a referendum.

Venizelos could not enjoy his well-deserved triumph. King Alexander died unexpectedly and himself was nearly killed in an assassination attempt by two royalist officers. The elections his government held in November of 1920 to capitalize politically on his successes led to a defeat. A disheartened Venizelos left Greece for France. The exultant royalists organized a plebiscite to bring back Constantine, despite stern warnings by the Allies.

An ailing and less confident Constantine and his incompetent governments ruled Greece for the next 2 years taking revenge against the Venizelists and trying to cajole the British. Their foolish blunder was a futile attempt to destroy the Turkish forces led by Mustafa Kemal (later Atatürk) by reaching and capturing Ankara. This was a strategic mistake comparable to the French and German invasion of Russia in the nineteenth and twentieth century. After 2 years of pyrrhic victories, the Greek army was defeated and retreated, leaving Smyrna to the advancing Turkish army. Smyrna was destroyed, and its Greek inhabitants were killed, captured, fled, or deported. The, more than 2,500 years, Greek presence in Asia Minor ended with the greatest catastrophe in Modern Greek history. It was also the end of Megali Idea. Greece had to abandon eastern Thrace, Smyrna, and two small Aegean

islands. Greeks left their ancestral homelands in Pontus and Cappadocia.

Bitter Divide: The Interwar Period (1922–1940)

After the catastrophe, a military coup, organized by mostly Venizelist officers, made Constantine abdicate in favor of his firstborn son who became King George II, but only temporarily. The process of transforming Greece into a republic had started. The radical Venizelist officers organized a mock martial court trial of the leadership of the royalists. Six of them were sentenced to death and they were speedily executed. Their execution led to political scars that for several decades plagued Greece. The royalists abstained in the 1923 national elections and the result was the political dominance of republican liberals for the next decade. On March 25, 1924, the second Greek Republic was proclaimed and a new constitution was adopted in 1927, but most royalists didn't accept the new regime and the National Schism deepened. To make matters worse, one radical republican, General Theodoros Pangalos, assumed dictatorial powers for a year. One of the challenges was to accommodate 1.3 million refugees from Asia Minor who integrated themselves completely only after decades.

Venizelos returned to power after a landslide electoral victory in 1928. His tenure from 1928 to 1932 was one of the most fruitful in almost all areas. From structural and institutional reforms to international relations. Even though the refugees were a great part of his electorate, he dared to visit Ankara and sign a friendship agreement with Atatürk, ending a century of conflict. However, the Great Depression led Greece to economic destabilization and to the loss of majority support to Venizelists. The anti-Venizelists ("People's Party") came back to power with a vengeance in the 1933 elections, ending the dominance of the Liberal Party and eventually capturing the bureaucracy and the military. Venizelos (dismayed after a second attempt against his life) and the republican officers tried to prevent the restoration of the King with an ill-executed coup in 1935 which led Venizelos to an

exile in Paris where he died the following year. George II returned with a rigged plebiscite. This was the end of the second Greek Republic (1924–1935). The Liberals accepted the regime change, but it was not enough. The rise of the Greek Communist Party and a wave of labor strikes was used as pretext by George II to accept a dictatorship under Ioannis Metaxas, an experienced and competent but ruthless retired officer and royalist politician. The Greek people tired by almost three decades of wars and political turmoil didn't resist to the new regime which was developed into a conservative authoritarian government fashioned after Fascist Italy.

The Decade of Wars (1940–1949)

Despite the rise of Nazi Germany, George II and Metaxas had learned their lesson from World War I. They had decided to keep Greece neutral, if possible, but if there was no choice, to remain faithful to Britain. When Mussolini, in October 1940, invaded Greece, the King and Metaxas decided to fight back, the Greek people totally agreeing with their decision and fighting bravely. The Greek army not only fended Italian forces off but also it managed to humiliate Mussolini by advancing in the under Italian control Albania capturing one city after another. This was the first military success against the Axis, so Hitler had to intervene while he prepared Operation Barbarossa against the Soviet Union. Germans invaded Greece in April 1941. The exhausted Greek army and a small British expeditionary force was not able to protect the country. Nevertheless, it took Germans almost 2 months to occupy the whole of Greece due to the fierce resistance by Greek and British forces in Crete.

King George (Metaxas died unexpectedly in early 1941) fled to Egypt where he appointed a new prime minister, the former Venizelist, Ioannis Tsouderos. In Athens, a puppet regime was established by collaborationists. Greece was separated in three different zones, occupied respectively by German, Italian and Bulgarian forces. Several resistance groups were established from the very beginning in mountainous areas with the

help of the British. The most well organized was EAM, that was controlled by the Communist Party with the same agenda as similar partisan organizations in occupied Eastern Europe, i.e., to distract German forces in their invasion of Soviet Union and to seize power after the end of the war.

EAM and its military branch, ELAS, managed to annihilate almost every rival resistance group and dominate the field. When the defeat of Germany was more than obvious, the puppet regime tried to organize paramilitary groups to prevent the capture of power by the communists after the German retreat. EAM formed a provisional government in order to impose its participation to the government in exile. A compromise was achieved with the backing of the British and a coalition government was formed under a seasoned liberal politician and a former associate of Venizelos, George Papandreou. Papandreou returned to Athens after the departure of German forces in October 1944 but EAM/ELAS controlled the rest of the country. The communists could have grabbed power rather easily. However, Winston Churchill had already secured Greece for the West in his negotiation with Stalin and the latter didn't encourage Greek communists who were baffled. The behavior of EAM ministers in the coalition government was not cooperative and the noncommunist members didn't trust them. EAM didn't want to disarm its army and relinquish the control of the periphery.

After the bitter resignation of its ministers, EAM organized a strong demonstration in early December 1944. The demonstration had a wretched ending after police killed several demonstrators. This led to a bloody confrontation in the streets of Athens: The numerically superior EAM/ELAS tried to seize control of the capital against the British forces stationed in Athens which were supported by a coalition faithful to the government: veterans from the Greco-Italian war and the North Africa Campaign, right-wing guerilla groups, policemen, even some former collaborationists. The communist forces were far superior, but they couldn't defeat this unlikely pro-government coalition. Winston Churchill ordered troops from the Italian front to join the small contingent in Athens and quell the

rebellion. Under considerable domestic pressure, the British prime minister deemed the situation critical enough for him to spend Christmas day in Athens, in an inconclusive mediation effort. Overpowered, the communists capitulated and accepted the disarmament of ELAS.

However, atrocities from both sides (Red/White terror), the decision of the Communist Party to abstain from the first postwar national elections in March 1946 (leading to a triumph for the royalist Right) and the return of King George II after a referendum polarized even more the Greeks and led to the final stage of the Civil War (1946–1949). This time the Communist Party had decided to fight to the end. The prospects were good: the communists controlled many parts of Greece, the kindred governments of Albania, Yugoslavia, and Bulgaria already provided material support and safe haven for the communist army, Soviet aid was anticipated while the exhausted British Empire relinquished its traditional role in Greece. But, it was replaced immediately by the United States who decided to assert its role as the postwar superpower, providing abundant economic and military support to the Greek government, under a veteran leader of the minority Liberals, Themistoklis Sofoulis. The failure of the communists to capture and hold any significant town was exacerbated by the Tito-Stalin split which eventually led to Yugoslavia closing its borders. After two inconclusive years of fighting, in summer 1949, the government forces launched their final assault against the communist stronghold on the mountains of northwestern Greece. By the end of August, the remnants of the defeated communist forces fled to Albania.

From Illiberal Democracy to Dictatorship (1949–1974)

Greece remained part of the western democratic world (joining NATO in 1952), but the price was an illiberal (tutinary) democracy which treated the defeated communists (and fellow travelers in general) harshly, deepening the rift instead of investing in reconciliation. The United States

exerted an enormous influence on Greek politics while the King (Paul, the third son of Constantine I succeeded George II in 1947 and Paul's son Constantine II succeeded his father in 1964) antagonized elected governments, especially during the mid-1960s. Nevertheless, Greece remained a democracy. The Communist Party was banned but it was represented by a leftist party, a political front, which, 9 years after the end of the Civil War, managed to become, for a brief period, the second strongest party in parliament. The leading politician of the period was Konstantinos Karamanlis, a conservative reformist who in 8 years (1955–1963) transformed Greece. He successfully pursued Greece's association with the European Communities and he was responsible for the most spectacular economic growth due to rapid industrialization and investment in infrastructure and tourism. Karamanlis also managed to reach a compromise solution for the Cyprus problem which since the early 1950s had plagued Greece's foreign relations and domestic politics. Greek Cypriots were the majority (78%) in the, under British administration, island. Nonetheless, there was a sizable Turkish Cypriot minority (18%) who felt threatened by the prospect of *Enosis* (the union of the island with Greece). Rather than insisting on *Enosis*, Greece agreed with Turkey and Great Britain for Cyprus to become an independent Republic in 1960. After a quarrel with King Paul, Karamanlis resigned and then lost the elections of 1963 to a centrist party led by George Papandreou. Papandreou himself had to resign in the summer of 1965 when the young King Constantine denied him his constitutional prerogative to discharge the defense minister of his government.

A period of political turmoil led to a military coup, organized by colonels against the political system in general in April 1967. Constantine II fled the country 8 months later, after a failed counter-coup, and all political activity ceased for 7 years. The self-confident leader of the military junta, George Papadopoulos, proclaimed Greece a "republic" in 1973, appointed a docile civilian government, and announced (controlled) elections for the following year. His plans went awry after student riots in November 1973 gave to the

regime hardliners an excuse to sack Papadopoulos and forestall the supposed liberalization process. Their foolish mistake was the attempt to assassinate the President of the Republic of Cyprus, Makarios, in order to annex Cyprus to Greece in mid-July of 1974. Turkey invaded Cyprus (the north of the island is still occupied by Turkish forces) and the chain of events led to the fall of the humiliated military junta. Konstantinos Karamanlis returned on July 24, 1974, to Greece after a self-imposed exile of 11 years.

The Third Greek Republic: From Metapolitefsi to the Economic Crisis (1974–)

Karamanlis, as the prime minister in a coalition government, managed to restore democracy (*metapolitefsi*) in a paradigmatic way, he legalized the Communist Party, and he passed several necessary structural reforms. Greece became a Republic with the constitution of 1975 (currently in force) after the 1974 referendum which ended Greek monarchy. It was the beginning of the current period of the Third Greek Republic which is the less tumultuous in modern Greek history. Karamanlis governed from 1974 to 1980 when he was elected President of the Republic. In January 1981, Greece became a full member of the European Communities.

The 1980s were dominated by the socialist party PASOK under its charismatic leader Andreas Papandreou (son of George) whose welfare populism triumphed. He governed Greece from 1981 to 1996 with a short interval (1989–1993) when the conservative party of Karamanlis (New Democracy) returned to power with a liberal reformist leader, Constantine Mitsotakis. PASOK's dominance continued after the death of Papandreou with the reformist social democrat, Kostas Simitis, as prime minister. Simitis' 8-years tenure was marked by high growth rates and Greece's accession to the Eurozone but with no major structural reforms. He was succeeded by Kostas Karamanlis (the nephew of the former Prime Minister and President) whose free-spending policies precipitated the Greek debt crisis of 2009–2010.

Karamanlis lost the 2009 elections to PASOK's George Papandreou (son of Andreas) who had to deal with the sovereign debt crisis. In May 2010, Greece entered a bailout agreement with the Eurozone countries and the IMF. Two more bailouts followed since, reaching loans to Greece to a total of over €250 billion. The adopted austerity measures agreed between Greece and its creditors (three economic adjustment programs) led to a remarkable fiscal consolidation but they stagnated the economy. This was so, because mostly fiscal measures were adopted and only some half-baked institutional reforms were enforced. The first bailout agreements were enforced by PASOK and then two coalition governments formed by New Democracy and PASOK. At the end of 2014, Greece was in a process of fragile recuperation which was upset by the victory of the radical leftist and populist SYRIZA. SYRIZA tried to renegotiate the bailout agreement by following a self-defeating brinkmanship strategy which led to its bitter capitulation in July 2015. The third bailout agreement included austerity measures, harsher than the preceding ones. Since then SYRIZA and its leader Alexis Tsipras managed to secure their power by cooperating with the creditors in a government coalition with a minor ultra-right-wing party. Despite its obvious failure to deliver on its electoral promises to get rid of austerity, this coalition outlived all preceding bailout governments.

Greece in the Twenty-First Century

Modern Greece has a history of almost two centuries. During these centuries, the country managed to move from the backwaters of Europe to a prosperous liberal democracy. From 1929 to 1980, Greece had an average annual rate of growth of income per capita of 5.2% (during the same period, Japan had an average of 4.9% and Germany 3.0%). However, this development was based on extractive institutions. The membership in the European Union and the Eurozone helped Greece put its extractive institutions under the rug of EU convergence funds, cheap international borrowing and fudged statistics. The crisis of

2008 was the triggering effect of the perfect storm that hit Greece in early 2010. Greece must replace its extractive institutions with inclusive institutions suitable for economic growth. This should be the new “Megali Idea” for the Greek people.

Acknowledgments The author is grateful to Konstantina Botsiou, Maria Efthymiou, Yulie Foka-Kavaliaraki, Basil Gounaris, Evanthis Hatzivassiliou, Akritas Kaidatzis, Dimitris Keridis, Nikos Marantzidis, Iakovos Michailidis, Ioanna Sapfo Pepelasis, Ioannis Stefanidis, Thanos Veremis, Spyros Vlachopoulos and Elpida Vogli for valuable comments to a previous draft. The usual caveat applies.

References

- Brewer D (2001) *The flame of freedom: the Greek War of Independence, 1821–1833*. John Murray, London
- Brewer D (2010) *Greece, the hidden centuries: Turkish rule from the fall of Constantinople to Greek Independence*. I.B. Tauris, London
- Brewer D (2016) *Greece, the decade of war: occupation, resistance and civil war*. I.B. Tauris, London
- Clogg R (2013) *A concise history of Greece*, 3rd edn. Cambridge University Press, Cambridge
- Diamandourous N (1972) *Political modernization, social conflict and cultural cleavage in the formation of the modern Greek state: 1821–1828*. Columbia University PhD thesis
- Featherstone K (ed) (2014) *Europe in modern Greek history*. Hurst, London
- Featherstone K, Papadimitriou D (2015) *Prime ministers in Greece: the paradox of power*. Oxford University Press, New York
- Gallant TW (2015) *The Edinburgh history of the Greeks, 1768 to 1913 (the long nineteenth century)*. Edinburgh University Press, Edinburgh
- Gallant TW (2016) *Modern Greece: from the war of independence to the present*, 2nd edn. Bloomsbury, London
- Gounaris BG (1996) Social cleavages and national ‘awakening’ in Ottoman Macedonia. *East Eur Q* 29:409–426
- Greene M (2015) *The Edinburgh history of the Greeks, 1453 to 1768 (the Ottoman Empire)*. Edinburgh University Press, Edinburgh
- Hatzis AN (2002) The short-lived influence of the Napoleonic Civil Code in 19th century Greece. *Eur J Law Econ* 14:253–263
- Hatzis AN (2018) Greece’s institutional trap. *Manag Decis Econ* 39. (forthcoming)
- Hatzivassiliou E (2011) *Greece and the Cold War: front line state, 1952–1967*. Routledge, London
- Hering G (1992) *Die politischen Parteien in Griechenland 1821–1936*, (2 vols.). R. Oldenbourg Verlag, Munich
- Iatrides JO (1972) *Revolt in Athens: the Greek Communist Second Round, 1944–1945*. Princeton University Press, Princeton
- Kaltchas N (1940) *Introduction to the constitutional history of modern Greece*. Columbia University Press, New York
- Kalyvas SN (2015) *Modern Greece: what everyone needs to know*. Oxford, New York
- Keridis D (2009) *Historical dictionary of modern Greece*. Scarecrow Press, Lanham
- Kitromilides PM (ed) (2006) *Eleftherios Venizelos: the trials of statesmanship*. Edinburgh University Press, Edinburgh
- Kitromilides PM (2013) *Enlightenment and revolution: the making of modern Greece*. Oxford University Press, Oxford
- Koliopoulos JS, Veremis TM (2007) *Greece: the modern sequel (from 1821 to the present)*, 2nd edn. Hurst, London
- Koliopoulos JS, Veremis TM (2009) *Modern Greece: a history since 1821*. Wiley-Blackwell, Malden
- Kostis K (2018) *History’s spoiled children: the story of modern Greece*. Oxford University Press, Oxford
- Mavrogordatos GT (1983) *Stillborn republic: social coalitions and party strategies in Greece, 1922–1936*. University of California Press, Berkeley
- Mazower M (1993) *Inside Hitler’s Greece: the experience of occupation, 1941–44*. Yale University Press, New Haven
- Mazower M (ed) (2000) *After the war was over: reconstructing the family, nation, and state in Greece, 1943–1960*. Princeton University Press, Princeton
- Mazower M (2004) *Salonica, city of ghosts: Christians, Muslims and Jews 1430–1950*. Vintage Books, New York
- McNeil WH (1978) *The metamorphosis of Greece since World War II*. University of Chicago Press, Chicago
- Pappas TS (1998) *Making Party Democracy in Greece*. Palgrave
- Petropoulos AJ (1968) *Politics and statecraft in the kingdom of Greece: 1833–1843*. Princeton University Press, Princeton
- Ricks D, Beaton R (eds) (2016) *The making of modern Greece: nationalism, romanticism, and the uses of the past (1797–1896)*. Routledge, London
- Stefanidis ID (2007) *Stirring the Greek nation: political culture, irredentism and anti-Americanism in post-war Greece, 1945–1967*. Ashgate, Aldershot
- Woodhouse CM (1976) *The struggle for Greece, 1941–1949*. Ivan R. Dee, Chicago
- Woodhouse CM (2000) *Modern Greece: a short history*. Faber & Faber, London

Greenhouse Effect

► [Global Warming](#)

Hard Law

Katarina Zajc

Laws School, University of Ljubljana, Ljubljana, Slovenia

Abstract

Hard law represents rules that are binding and precise and delegate the power either to explain or adjudicate to third parties. Soft law, on the other hand, does not have a status of a binding rule but nonetheless influences the behavior of public. The literature still debates whether the hard and soft law are either complementary or antagonistic. Especially with development of EU, we can see that soft law has gained momentum. We will see more soft law in the areas of uncertainty or in areas where the changes are day-to-day occurrence or where soft law usually paves the way for the hard law.

Definition

Hard law refers to legally binding obligations that are precise (or can be made precise through adjudication or the issuance of detailed regulations) and that delegate authority for interpreting and implementing the law. (Abbot and Snidal 2000)

Soft law, on the other hand: "... consists of rules issued by lawmaking bodies that do not comply with procedural formalities necessary to give the rules

legal status yet nonetheless influence the behavior of other lawmaking bodies and of the public." (Gersen and Posner 2008)

The desire to gain and distribute rights among people and the disputes about rights are as old as human race. Therefore, usually governments/states take on the role of enacting the rules in order to distribute the rights and solving disputes. Even though Coase (1960) claimed that rules are not necessary if the rights are distributed and transaction costs are zero. However, we do not live in the world of zero transaction costs, and rules might be therefore beneficial.

Broadly speaking, the governments/states have either hard law or soft law at their disposal, depending on the goal that they try to reach with enactment of the rules and the advantages and disadvantages associated with each.

Even though definitions of hard law vary to a certain degree, the literature agrees that hard law represents rules that are binding and precise and delegate the power either to explain or adjudicate to third parties (Abbot et al. 2000). Hard law rules therefore have to be precise in such a way that it is unambiguous what kind of conduct they prescribe or require even though some hard law rules are and can be somehow relaxed along its dimension. If they are too relaxed, they might fall under the realm of soft law, even though they are binding and delegate the power to adjudicate to some third party. Hard law puts obligations on the entities within the realm of the jurisdiction of the hard law,

and if the entities breach their obligations, the sanctions are prescribed *ex ante*. For example, almost all legal regimes include a maxim “*neminem laedere*,” the obligation of not hurting anybody. If somebody commits a tort and is found liable in the court of law, it usually has to pay damages, which range is prescribed in law *ex ante*. Committing a crime usually means times in prison, which is also predetermined in hard law *ex ante*. Delegation means that special institutions, for example, courts, arbitrations, and similar bodies, have the authority to interpret the legal rules to a certain degree and to adjudicate the dispute.

Hard law is enacted in the parliamentary bodies of countries, and very rigorous rules are used in order to pass the hard law in the parliament. Usually, the rudimentary rules about the enactment of hard law are legislated in the Constitution and further elaborated in certain legislative acts. For example, in a parliamentary system with bicameralism, only the government, each parliament member, a higher parliamentary chamber as a whole, or a certain number of voters can propose an act. (Article I, section 7 of the US Constitution requires that a bill be approved by both houses of Congress and signed by the President.) The proposal of an act has to have certain elements, for example, the reasons for enactment of the act, goals of the enactment, financial consequences for the country’s budget, confirmation that there are enough assets in the budget to support the act once, and if, enacted, comparative study. There are usually certain prescribed phases through which the act should go before the final vote is taken in one of the chambers of the parliament. Once the lower chamber passes the law (using different majorities as prescribed either by the Constitution or other acts), the higher chamber has to pass it too. They can veto the act and then the lower chamber has to pass the law with higher majority than the first time, for example, with absolute instead of the relative majority. The president usually executes the act and the act is published in the official gazette.

Advantages and Disadvantages of Using Hard Law as Opposed to the Soft Law

One of the main advantages of hard law is that it reduces the transaction costs of subsequent transactions, since the rules, once enacted, regulate all future behavior regulated by the act. (see Abbot and Snidal 2000.) For example, if the parties are in a contracting relationship, they do not have to either negotiate over the provisions (even though not all provisions are defaults in a sense that the parties can contract around, one of the provisions are mandatory, meaning that the parties cannot change them with the contract) or do not negotiate at all about the provision enacted in the Law on Obligation. The transaction costs of contracting are therefore decreased. Also, hard law, in domestic law and in international relations, strengthens the credibility of commitments and decreases the costs associated with the problems of incomplete contracts. Parties of the contract, for example, know that if one of the parties breaches the contract or obligation, they could file a claim at predetermined institutions which must adjudicate according to rules and impose sanctions determined by law and can therefore rely on the promises made by the parties in the contract. In other words, adjudication is guaranteed and costs of it are fixed and determined *ex ante*. Credible commitments are crucial in situations when one party performs its obligations before the other party, when there are asset-specific investments made upfront, basically in any situation in which one party is vulnerable to the other. Without such credible commitments, the number of transactions would drop or transactions would be made among parties that know and trust each other, which diminished the number of potential partners and consequently the advantages of contracting.

Disadvantages of Hard Law

A major disadvantage of hard law is the cost of its enactment. As already described, the procedures

are rigorous, and the agreement among the parties deciding about enactment of the hard law must be pretty high. Also, once the act is passed and in the case that is very precise, it cannot adjust to the changed circumstances which might not be advantageous in certain circumstances, for example, in the area of finance where changes are occurring daily. However, it should also be pointed out that even some procedures to enact soft law can be very costly.

Interaction of Hard and Soft Law

The literature still debates whether the hard and soft law are either complementary or antagonistic (see Shaffer and Pollack 2010.) Complementarity is seen in two ways. Soft law can lead to adoption of hard law with the same content, and hard law can be further elaborated and explained through soft law provisions. However, especially in the international relationships, the hard and soft law can be antagonistic, especially in the situations of significant distributive conflicts and pluralistic regime complexes. The antagonistic interaction of hard and soft law can in the mentioned situations lead to “hardening” of the soft law and “softening” of the hard law.

Conclusion

Hard law might be advantageous to reach some goals of the government. However, especially with the development of the EU (Trubeck et al. 2006), we can see that soft law gained momentum. As literature claims, we will see more soft law when the formalities of the hard law rise relative to the costs of the soft law. It could be also said that we will see more soft law in the areas of uncertainty or in areas where the changes are day-to-day occurrence as, for example, on the financial markets where rules should be very flexible. Also, as we see predominantly in the EU law, soft law usually paves the way or tests the waters for the hard law.

References

- Abbot WK, Snidal D (2000) Hard and soft law in international governance. *Int Organ* 54:421–456
- Abbot K, Moravcsik A, Slauther A-M, Snidal D (2000) The concept of legalization. *Int Organ* 54:17–35
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Gersen JE, Posner EA (2008) Soft law: lessons from congressional practice. *Stanford Law Rev* 61:573–627
- Shaffer GC, Pollack MA (2010) Hard vs. soft law: alternatives, complements and antagonists in international governance. *Minn Law Rev* 94:712–799
- Trubeck DM, Cottrell P, Nance M (2006) Soft law, hard law and European integration: toward a theory of hybridity. In: Scott J, de Búrca G (eds) *New governance and institutionalism in Europe and in the US*. Hart Publishing, Oxford, pp 65–94

H

Hardship

► [Impracticability](#)

Harmonization of Tort Law in Europe

Mauro Bussani^{1,2} and Marta Infantino¹

¹University of Trieste, Trieste, Italy

²University of Macau, S.A.R. of the P.R.C., Macau, China

Abstract

The aim of the entry is to provide an overview of the projects regarding harmonization of European tort laws and the costs and benefits associated with them. To this purpose, the entry critically examines the arguments for and against tort law harmonization in Europe and investigates the assumptions and approaches underlying the many institutional and academic initiatives in the field. The entry tentatively draws some conclusions about the short- and long-term prospects of European tort law harmonization.

Definition

Processes aiming to align European tort law rules and practices, either through hard or soft law measures.

Introduction

In the past decades, the idea of harmonizing European private laws has been high on many agendas – especially in the EU institutions and among European legal scholars. While at the beginning institutional and doctrinal efforts mostly focused on contract law, tort law has recently been added to the picture, and its harmonization at the European level has now become a primary goal for many EU institutions and research groups.

The aim of this entry is to offer an overview of the various projects regarding harmonization of European tort laws and the costs and benefits associated with them. To this purpose, the entry will start with a description of what European tort laws have and do not have in common (section “[The Variety of European Tort Laws](#)”). Against this framework, we will be able to appreciate the arguments that have been put forward for and against tort law harmonization in Europe (section “[Harmonization’s Pros and Cons](#)”), as well as the assumptions and approaches underlying the many institutional and academic initiatives in the field (section “[The EU’s Piecemeal Approach](#)” and “[Scholarly Endeavors](#)”). We will first deal with the steps taken by the EU institutions in harmonizing substantive and conflicts of law rules in tort law matters (section “[The EU’s Piecemeal Approach](#)”). We will then investigate the main features of the research groups that have recently been established to support – although through different means – the Europeanization path (section “[Scholarly Endeavors](#)”). Some of these initiatives, like the *European Group on Tort Law* and the *Study Group on a European Civil Code*, seek to draft a “soft” European tort law, e.g., a law the binding force of which is not derived from the official authority conferred by its source but from the high reputation enjoyed or

displayed by its compilers (section “[Striving for Harmonization: The European Group on Tort Law and the Study Group on a European Civil Code](#)”). Other projects refuse to devise solutions and engage instead in deepening the dialogue between European legal cultures. This is particularly the case with two projects – the *Ius Commune Casebook for a Common Law of Europe* and the *Common Core of European Private Law*. Despite some divergence, both are devoted to developing a better knowledge of European legal landscapes (section “[Building Knowledge: The Ius Commune Casebooks for the Common Law of Europe and the Common Core of European Private Law Project](#)”). Such an overview will allow us to draw some conclusions about the short- and long-term prospects of European tort law harmonization (section “[Conclusions](#)”).

The Variety of European Tort Laws

European tort laws are far from being harmonized. Judicial opinions in tort disputes (as in any private law conflict) range from being concise and, in principle, self-contained, such as the French Cour de Cassation’s rulings, to being lengthy and full of references to academic literature as is the case in Germany (Quézel-Ambrunaz 2012, 108; Markesinis and Unberath 2002, pp. 9–12). They may look like unanimous and anonymous decisions of the court (civil law), or, as in common law, they may show the individualized opinion of individual judges, be they concurring or dissenting (van Dam 2013, pp. 53–5, 75–7, 95–7). Legal scholarship plays a role in (tort) law-making processes which is strong and evident throughout the continent except England (Bussani 2007a, p. 378). In continental Europe, tort law reasoning usually starts with the black-letter words that legislators use in codes and statutes, while in common law countries the same role is played by judicial doctrines hardened into precedents (Bussani and Palmer 2003, pp. 120–159).

Fault-based rules are presented everywhere as the foundation of the tort law system (Werro and Palmer 2004, p. 13), but they may either be open ended, attaching liability for any conduct which

causes damage to another (e.g., France, Italy, Poland), or may provide for liability only in certain situations, thereby restricting recovery to the breach of specific duties (typically England) or to the infringement/violation of a list of specific interests (such as Germany, Austria, Scandinavian countries) (von Bar 2009, pp. 229–33). General rules on fault liability may require the plaintiff to prove the defendant's negligence (Western Europe) or may state that once the damage is proved, the defendant's negligence is presumed (post-socialist countries) (Menyhárd 2009, pp. 350–2). Resort to strict liability rules may be made routinely (as in France) or as an exception (as is the case in England), with a range of intermediate positions (Werro and Palmer 2004, pp. 29–31). Vicarious liability for acts of a minor or an employee, which in German tort law is grounded in the responsible person's fault in supervision, is in the majority of the other jurisdictions imputed to him/her in an objective manner (Werro and Palmer 2004, pp. 393–6). Different from other European legal systems, in France (and to a certain extent in Belgium), the plaintiff is barred from suing the defendant(s) in tort if there is a contract between them (although the plaintiff can combine, in a single proceeding, actions in contract and in tort against different defendants) (Giliker 2010, p. 44; von Bar and Drobnig 2004, pp. 198–9). In Scandinavian countries, the impact of insurance on tort law mechanisms is more pervasive than anywhere else, effectively turning tort law into a residual remedy in many areas (Palmer and Bussani 2008, pp. 65–6; Bussani and Palmer 2003, pp. 156–8; Ussing 1952). Punitive damages and jury trials are available in England in an extremely limited number of cases but are almost unknown throughout the continent (Magnus 2010, pp. 106–7). European systems also diverge as to the (limited) extent to which they allow the procedural aggregation of victims' claims and contingency fee agreements between lawyers and their clients (see, respectively, Oliphant 2010, pp. 120–124, 163–164, 204–209, 287–288; Reimann 2012, pp. 3, 45).

Similarities and differences of course do not end there, but these illustrations suffice to make

clear the great variety of European tort laws. It is a variety that at first sight – but at first sight only – may resemble the scenario in the United States, where increasing divergence between the federal and states' (tort) laws pushed the *American Law Institute* to launch, in the late 1920s of the twentieth century, the idea of the *Restatements of the Law*. Such a first impression presented by the analogy would, however, be wrong. There are considerable differences between the European case and that of the United States. We will not point out how punitive damages, jury trials, aggregation of claims, and contingency fee agreements are largely unknown to European tort laws, while they shape the basic vocabulary of tort law in the United States (Magnus 2010, pp. 102–24). There is more to it than this. Although the United States display 51 tort law jurisdictions, including the federal one, these regimes largely share a common language and employ the same reservoir of notions and technicalities. Most US legislatures are affected by the same pressures coming from power groups acting across state boundaries, like the insurance industry and the American Trial Lawyers Association. Moreover, the US Supreme Court, in matters it can intervene in, operates as a driving force for uniform outcomes (Palmer and Bussani 2008, pp. 52–3; Stapleton 2007, p. 25). Europe, by contrast, lacks one Supreme Court for private law matters. The insurance market is still not homogeneous, and lawyers are associated at the national level only. There is no common language, but 24 distinct ones, and every legal system relies upon its own set of notions and technicalities – a set that, as we just saw, may significantly overlap but may also greatly diverge among different jurisdictions.

Harmonization's Pros and Cons

As said above, the idea of harmonizing the variety of European tort laws for a long time did not attract the attention of institutions and scholars. Up until the 2000s it was monopolized by promoting legal convergence of contract laws, thought as the area that impacted business activities directly (Bussani 2007c; for the observation

that convergence would be easier in contract law rather than in tort law, insofar as contract laws are all inspired by the aim of facilitating trade, while tort laws are closely linked to domestic preferences as to what is to be protected (see Ogus 1999, pp. 412–7).

From the 2000s onward, things have changed. Although the EU has plunged into the field of liability insurance since the 1970s, the majority of statutory interventions on tort law have been adopted after the turn of the century (see below, section “The EU’s Piecemeal Approach”). In the same period, books on comparative law of European tort law started to multiply (Infantino 2012; Bussani and Werro 2009; Bussani 2007b; van Dam 2006; Brüggemeier 2004; Bussani and Palmer 2003; Zimmermann 2003b; van Gerven et al. 2000; von Bar 1998, 2000), as did studies on the costs and benefits of tort law harmonization (van Boom 2009; Bussani 2007a; van den Bergh and Visscher 2006; Wagner 2005; Faure 2003; Hartlief 2002; Banakas 2002). Two scholarly projects aiming at drafting a text for a future European tort law – the *European Group on Tort Law*, established under the auspices of the German reinsurance company Munich Re, and the *Study Group on a European Civil Code*, generously financed by the EU – published their results in 2005 and 2006, respectively (see below, section “Striving for Harmonization: The European Group on Tort Law and the Study Group on a European Civil Code”). In 2010, the *European Centre of Tort and Insurance Law*, closely affiliated to the *European Group on Tort Law*, distributed the first issue of the *Journal of European Tort Law* (de Gruyter).

Despite the different perspectives taken by the initiatives, the EU institutions and the scholars who participate in striving-for-harmonization projects agree that simplifying the current diversity of national tort laws is an aim worth pursuing. The reasons put forward to support such a view are many. First, a single European tort law regime is thought to contribute toward achieving the wider goal of a common area for free movement of goods, services, capital, and people. Simplifying the current European tort law patchwork would avoid the risk of inhibiting the mobility of persons

and goods by the different conditions for liability and amount of compensation (Magnus 2002, pp. 206–207). It would minimize the risk of European businesses’ forum shopping in search for the jurisdiction with the lowest quality and liability standard, thus fending off pressures on states to engage in a race to the bottom (Faure 2003, pp. 47–51). It would also increase the attractiveness of the European market to non-European economic operators: Europe would be more business-friendly if economic actors only had to tackle one unified regime, instead of 29 (1 supranational and 24 national) (van Boom 2009, p. 438; Reg. n. 864/2007, whereas 16, 20). Additionally, it has been emphasized by the EU institutions that a harmonized scenario would facilitate courts’ handling of trans-boundary torts, decrease the length and complexity of transnational litigation, and guarantee more uniformity between judicial outcomes (Reg. n. 864/2007, whereas 16, 20). Last but not the least, the EU Commission has underlined that a single tort law framework would allow insurers to better operate throughout the European Union and to establish and provide services in a freer manner (European Commission Directorate-General for Internal Market and Services 2007, pp. 47–8) – an observation echoed by scholars, who have stressed that the fragmentation of European tort law has a number of detrimental consequences for insurance enterprises operating within the single market (Wagner 2005, 1274).

This institutional and academic enthusiasm has not been supported by everyone. Many have criticized the claims upon which the very idea of harmonizing tort law is founded. They have pointed toward the lack of empirical evidence supporting the allegation that fragmentation of tort laws affects the free circulation of people and goods (van den Bergh and Visscher 2006, pp. 513–6; Faure 2003, pp. 44–7) and the establishment and movement of business in Europe (Wagner 2005, p. 1272; Hartlief 2002, p. 228). In this light, the true beneficiaries of the harmonization process would not be people and business but insurers and scholars themselves. On the one hand, tort law harmonization would reconcile the many legal surroundings to which European insurance companies currently adjust to and

would therefore lower insurers' barriers to entering and/or operating in multiple jurisdictions (Wagner 2005, p. 1274). On the other hand, the pro-harmonization movement would provide many scholars with a suitable avenue to enhance their own prestige, enabling them to promote their career, collect funding, and attract social, legal, and economic attention to their work (Infantino 2010, pp. 48–9; with regard to harmonization of laws in general, see Schepel 2007, pp. 187–8; Hesselink 2004, pp. 688–9).

Many have reminded that the benefits that harmonizing tort law may provide should be weighed against its possible costs, such as the difficulty in changing laws, the suppression of local preferences, and the abolition of any regulatory competition (van den Bergh and Visscher 2006, pp. 513, 516–8; Faure 2003, pp. 36–7, 59–66). There is indeed little doubt that for any harmonization plan to be effective, the new uniform tort law rules would have to be subject to the unifying control of one Supreme Court (Bussani 2007a, p. 377) and would need to be coordinated with the notions and solutions offered in other fields of law, including civil procedure, criminal matters, administrative remedies, and constitutional provisions (Bussani 2007a, pp. 365–8, 377; Faure 2003, pp. 45–6). Even if judicial review and coordination with the broader legal framework could be guaranteed, any top-down harmonization effort would put into circulation rules foreign to the tradition and heritage of some, if not all, legal traditions involved. Lack of familiarity with the new rules and their underpinning rationales, as well as the possible path dependency on deep-rooted local traditions, could lead to the defeat of any harmonization project (Bussani 2007a, pp. 373–7). Moreover, it has been noted by many that the idea of harmonizing European tort law would have to face the same issues that are raised by any harmonization endeavor at the European level. Such issues include policy and linguistic choices, the (lack of) competence of the EU institutions to harmonize private law in general, the political legitimacy of the drafters, and the feasibility and desirability of a European code in general (Giliker 2009, pp. 268–72; Bussani 2007a, pp. 371–3; Magnus 2002, pp. 208–212).

This is not the place to align with either one or the other side of the debate. The following pages will simply try to review how the EU institutions and scholars have contributed to the tort law harmonization discourse. Let us start with the EU's role.

The EU's Piecemeal Approach

According to the EU treaties, the EU does not have the general and comprehensive power to intervene in the field of tort law. The only competence assigned to the EU by the treaties with regard to tort law concerns the responsibility of Member States and of the EU itself in cases where they breach their obligations under the treaties (see Arts. 260(1), (2), and (3) and Art. 340(2) of the Treaty on the Functioning of the European Union). The task of judging the compliance of Member States and the EU institutions with the treaties is entrusted to the Court of Justice of the European Union, whose case law on this point has played an important role in shaping the states' liability across Europe (van Gerven 2009, pp. 32–3).

Yet the EU has, through time, carved out new competencies in the realm of tort law. It has done so through the adoption of statutory laws – mostly directives – aimed at harmonizing those segments of tort law that were deemed to most often cross national boundaries and/or affect the development of the internal market the most (von Bar 1998, pp. 401–7). The trend started in the 1970s with the directive on liability insurance, the objective of which was to establish a harmonized insurance system to facilitate people's free movement and guarantee compensation to persons injured in a Member State different from their own (Council Directive 73/239/EEC of 24 July 1973 on the coordination of laws, regulations, and administrative provisions relating to the taking-up and pursuit of the business of direct insurance other than life assurance, now replaced by the Commission Directive 2009/139/EU of 25 November 2009 on the taking-up and pursuit of the business of insurance and reinsurance). In 1985, a directive on products liability pursued consumer safety

through the adoption of a strict liability regime for producers of defective products (Council Directive 85/374 of 25 July 1985 on the approximation of the laws, regulations, and administrative provisions of Member States concerning liability for defective products). In 2004, two other directives introduced a common framework, respectively, for compensating crime victims (Council Directive 2004/80 of 29 April 2004 relating to compensation to crime victims) and for protecting the environment on the basis of the “polluter pays” principle (European Parliament and Council Directive 2004/35/CE of 21 April 2004 on environmental liability with regard to the prevention and remedying of environmental damage). On the assumption that cross-border externalities may be countered by harmonized rules of private international law, the EU enacted a regulation in 2007 enabling conflict of law rules to designate the law applicable to transnational tort claims (Regulation (EC) No. 864/2007 of the European Parliament and of the Council of 11 July 2007 on the law applicable to non-contractual obligations). In 2013, the European Commission issued a recommendation on a set of common, nonbinding principles for collective redress mechanisms (Commission Recommendation of 11 June 2013 on common principles for injunctive and compensatory collective redress mechanisms in Member States concerning violations of rights granted under Union Law). The last act of the EU statutory series on tort law was a directive adopted (not yet formally) in 2014 and aimed at removing practical obstacles to compensation for victims of infringement of the EU antitrust law (Directive of the European Parliament and of the Council on certain rules governing actions for damages under national law for infringements of the competition law provisions of Member States and the European Union).

Some of the abovementioned reforms have apparently enjoyed a remarkable success, both within and outside the EU borders. Compulsory vehicle insurance, for instance, is now a reality throughout the EU. The directive on products liability has inspired many legislators around the world who preferred to follow the newly

established EU model rather than the US model (Reimann 2003). Yet the actual effectiveness of the abovementioned EU laws is overall much debated. To illustrate, studies carried out on the insurance sector have pointed out that the liability regimes underlying compulsory insurance for traffic accidents are still diverging remarkably among European jurisdictions (van Dam 2013, p. 459). In a similar vein, it has been noted that, despite the external success of the products liability directive, the convergence created by the act has been minimal, and the rate of litigation grounded on the EU-branded products liability regime has remained incredibly low (Reimann 2014; Howells 2008).

Many reasons have been put forward to explain the limited effect of the EU’s harmonizing strategy in the field of tort law.

A first explanation points to the piecemeal approach taken by the EU. The EU institutions have up to now kept themselves far from any intervention in the general architecture of substantive tort law. They implicitly assume that tort law can be divided into a core of general rules to be left to national jurisdictions and “special” rules where the EU legislation can effectively intervene (von Bar 1998, p. 408). Yet the “special” rules can only be applied within and through the framework of the general rules. This is why planting the seeds of a special EU discipline in national tort law frameworks risks being a bad strategy for achieving the goal of minimizing the differences between national tort laws (van Gerven et al. 2000, p. 10; von Bar 1998, p. 408).

Such a risk is increased by both the contradictory character of the EU patchwork of tort law statutory provisions and by the absence of a court entrusted with a general competence over the interpretation of those provisions (Koziol 2007, pp. 5–6; Koch 2007, p. 109; van den Bergh and Visscher 2006, p. 11; Faure 2003, pp. 56–7). Contradictions often arise within the EU legal framework itself, because the EU institutions lack, among other features, a common vocabulary and a standard terminology capable of summarizing the different notions arising from the European linguistic and legal plurality. Contradictions may

also appear within national legal contexts where the EU rules are to be implemented, due to the lack of homogeneity between the EU and national languages, concepts, and rules (von Bar 1998, pp. 387–9).

All the above flaws are said to be worsened by the very use of the directives as the EU's main legislative tool (Koziol 2007, pp. 5–6; Koch 2007, p. 109; von Bar 1998, p. 410). Directives have the virtue of flexibility, i.e., of guaranteeing that each state can adapt the EU acts into its national categories, but the flip side is that frequently the outcomes of the process of implementation diverge greatly due to the tendency of Member States to replicate the traditional features of their legal system into the implemented rules. Thus, directives may result in intensifying legal differences as opposed to supporting uniformity (van Gerven et al. 2000, pp. 9–10; for some concrete examples, see Infantino 2010, p. 58).

Resorting to regulations rather than directives is only a limited cure. With regard to the only regulation so far adopted in the tort law field – Reg. n. 864/2007 on the law applicable to non-contractual obligations – it has been noted that many factors might hinder its uniform application. Divergence, for instance, may stem from the lack of agreement on the meaning of notions, such as “tort claim,” “injury,” “direct,” and “indirect” consequences (Hay 2007, pp. 139–40, 144, 149). Differences may also arise from the well-known tendency of national jurists to interpret foreign law in light of their national notions, categories, and rules of law or to even apply to transnational cases their own national law, no matter what the conflict of law criteria says (Fauvarque-Cosson 2001, p. 412).

Scholarly Endeavors

The above considerations led many scholars to pave their own way toward a truly common European tort law.

Aims and methods of these endeavors are very dissimilar from one another. There are associations, such as the *Pan-European Organisation of*

Personal Injury Lawyers (PEOPIL), whose purpose is the promotion of judicial cooperation between the European jurisdictions in personal injury litigation (see peopil.com). There are scholars who collect, translate, and make available cases from different European jurisdictions, on the assumption that the ability of European courts to engage in exercises of legal comparison when deciding cases might be the right path to achieve a truly, long-term harmonizing effect (Markesinis 2003, pp. 156–176; see also the database at utexas.edu/law/academics/centers/transnational/work_new/; a similar enterprise is carried out by the *Institute for European Tort Law* with its EURO TORT database, atctil.org/ectil/EuroTort.aspx). Two closely connected Austrian-based institutions – the *European Centre of Tort and Insurance Law* and the *Institute for European Tort Law* – provide a forum for research on comparative tort law and a venue for publication of up-to-date information and commentary about European tort law. The two institutions (which also support the *European Group on Tort Law*) organize an Annual Conference on European Tort Law every year, update a database of European case law on tort (see the EURO TORT database mentioned above), and publish many series on tort law and a peer-reviewed journal (the *Journal of European Tort Law*) (for more information, see ectil.org and etl.oecaw.ac.at).

Some projects, such as the *Ius Commune Casebooks* and the *Common Core of European Private Law*, are focused on the need for improving knowledge about the EU's legal systems. Other groups, such as the *European Group on Tort Law* and the *Study Group on a European Civil Code*, have engaged in ascertaining solutions that may best regulate certain legal problems and in codifying those solutions in the text of a would-be European code.

Given the breadth and significance of their work, it is the last four initiatives we mentioned that we will focus our attention on. We will begin with the striving-for-harmonization enterprises and then move on to those whose pivotal aim is building and developing better knowledge of European private laws.

Striving for Harmonization: The European Group on Tort Law and the Study Group on a European Civil Code

As anticipated, two scholarly groups that have thus far attempted to draft a text for a would-be codification on European tort law: the *European Group on Tort Law* and the *Study Group on a European Civil Code*.

The former group was established in 1992 within the Viennese *European Centre of Tort and Insurance Law*, whose main sponsor is the German reinsurance company Munich Re (for more information, see egtl.org and ectil.org). From 1992 to 2005, the group accomplished many studies, the results of which have been collected in a series called *Principles of European Tort Law* (PETL), published by Kluwer Law International (Widmer 2005; Rogers 2004; Magnus and Martín Casals 2004; Spier 1998, 2000a, b, 2003; Koch and Koziol 2002; Magnus 2001; Koziol 1998). Each volume gathered national reports and comparative results of an inquiry carried out on a specific tort law topic (e.g., causation, fault, wrongfulness, strict liability, etc.). Contributors were asked to describe the legal treatment of the assigned topic in their country by responding to some theoretical issues and by solving concrete cases; the editors then summarized the results (European Group of Tort Law 2005, pp. 14–16). The outcomes of the research were used as a starting point for drafting the PETL, which were published in 2005. The PETL are divided into ten chapters: Basic Norm, Damage, Causation, Liability Based on Fault, Strict Liability, Liability for Others, Defences in General, Contributory Conduct or Activity, Multiple Tortfeasors, and Damages (for a general overview of the contents of the PETL, see Oliphant 2009; Koch 2007, 2009; van Boom and Pinna 2008; Koziol 2007; van den Bergh and Visscher 2006; Zimmermann 2003). The group is currently working on a new edition.

The other project committed to shaping a would-be legislative text was the *Study Group on a European Civil Code*, founded in 1998 by Professor Christian von Bar. The *Study Group* has the more general purpose of drafting a European code on the whole of private economic law

(von Bar 2001). In the *Study Group's* view, the preparatory work on a European code had to be performed by scholars, who are the only ones endowed with the necessary expertise to conduct the essential basic research and to set up rules unaffected by the particularities of national interests; the legislator's role could begin only once the academic work of selecting the *Principles of European Law* (PEL) was completed (von Bar 1999). In 2006, the drafting of the tort law book for the would-be European code was completed. The *Principles of European Law on Non-Contractual Liability Arising out of Damage Caused to Another* were first issued online on the group's website at sgecc.net and then included, with minimal modifications, in the *Draft Common Frame of Reference* prepared by the *Study Group* for the European Commission in 2008 and officially published by Sellier in 2009. The PEL are composed of seven chapters: Fundamental Provisions, Particular Instances of Legally Relevant Damage, Accountability, Causation, Defences, Remedies, and Ancillary Rules (von Bar 2009; for a general overview of the PEL contents, see Oliphant 2009; Blackie 2009; Blackie 2007; Blackie 2003 p. 133).

The intention of both the above groups was to supply the European (and national) legislator (s) with a possible basis for a new codification. To accomplish this goal, the groups neither looked for the rules most widely accepted across European countries, nor did they select, among the existing rules, the ones deemed most suitable for Europe. Instead, the groups sought what was the “best” solution to tort law problems, regardless of whether this solution reflected principles already established within any European jurisdiction (von Bar 2009, pp. 229–38; European Group of Tort Law 2005, p. 15).

On many issues, the drafters of the PETL and of the PEL clearly agreed about what the “best” solution for Europe would be. Both the texts take fault as the general basis for liability (cp. Arts. 4:101 PETL and 1:101(1) PEL) and complement it with vicarious liability rules for employers (Arts. 6:102 PETL and 3:201 PEL) and parents/supervisors of minors and mentally disabled persons (Arts 6:101 PETL and 3:104 PEL). In the

PEL as well as in the PETL, not all interests of the plaintiffs are worthy of the same legal protection. While there is no doubt about the abstract recoverability of losses deriving from the infringement of life, bodily and mental integrity, human dignity, liberty, and property (cp. Arts. 2:102 PEL and 2:201–3, 2:206 PETL), in both the texts compensation of pure economic loss is restricted (cp. Arts. 2:101(2)–(3) PETL with 2:204–5, 2:207–8, 2:210–1 PEL). The PETL as well as the PEL stress that compensation is the primary function of tort liability (cp. Arts. 10:101 PETL and 6:101 PEL). Moreover, both the PETL and the PEL emphasize the need to take into consideration special features of tort law conflicts that can actually be brought under examination, calling for an application of their rules carefully tailored to all the relevant circumstances of the case (cp., for instance, Arts. 2:105, 3:103, 3:201, 4:102, 4:201, 10:301(2) PEL and 2:101, 3:103(3), 5:301, 6:101(4), 6:103, 6:202 PEL).

Yet in many other cases the PETL and the PEL consistently diverge from one another as to what the “best” solution is. For instance, under the PETL, minors and persons with mental disability are not exonerated from personal liability (Art. 4:102 PETL), while special rules are designed by the PEL for personal liability of minors and persons with diminished mental capacity (Arts. 3:103 and 5:301 PETL). There is no uniformity between the texts as far as no-fault liability is concerned. The PETL set forth a presumption of fault for evaluating harmful entrepreneurial activities (Art. 4:202(1) PETL) and impose strict liability on whoever engages in an abnormally dangerous activity “for damage characteristic to the risk presented by the activity and resulting from it” (Art. 5:101 PETL). By contrast, the PEL provide no presumption of fault but rather establish a set of strict liability regimes, specifically tailored to compensate the damage caused by the dangerous state of an immovable property (Art. 3:202 PEL), animals (Art. 3:203 PEL), defective products (Art. 3:204 PEL), motor vehicles (Art. 3:205 PEL), and dangerous substances or emissions (Art. 3:206 PEL).

As far as damage is concerned, the PETL give minimal guidance regarding the criteria for the

(un)recoverability of pure economic losses (Art. 2:102(4) PETL), but are more detailed as to the conditions under which redress for pain and suffering may be justified (Art. 10:301 PETL). By contrast, the PEL devote six articles to the requirements for compensation of pure economic losses (Arts. 2:204–5, 2:207–8, 2:210–1 PEL; the circumstance led many scholars to observe that the PEL open the liability floodgates for pure economic loss to an extent that goes far beyond the accepted standards in most Member States: see Zimmermann 2009, p. 496; Wagner 2009, pp. 234–7; Eidenmüller et al. 2008) and do not restrict compensation for pain and suffering and impairment of the quality of life (Art. 2:101(4) (b) PEL).

According to the PETL, even when an interest is deemed worthy of protection, the judge has to take into account all the relevant circumstances of the case in order to determine whether or not the recovery is undesirable in light of other public interests or the necessary “liberty of action” of the defendant (Art. 2:102(6) PETL). To instill a similar pro-defendant flexibility, the PEL adopt a wider approach, allowing the judge to not only verify whether, in light of all the circumstances of the case, granting compensation would be “fair and reasonable” (Art. 2:101(2)–(3) PEL) but also to relieve the faulty defendant from liability when “liability in full would be disproportionate to the accountability of the person causing the damage or the extent of the damage or the means to prevent it” (Art. 6:202 PEL).

The list of divergence between the PETL and the PEL could go further. For instance, the PEL, but not the PETL, authorize the plaintiff to take precautionary measures so as to reduce the risk of damages (Art. 1:102 PEL) and to claim that the defendant disgorge the profits she has obtained from the wrongdoing (Art. 6:101(4) PEL). The PETL, differently from the PEL, allow the judge to disregard trivial damage, regardless of whether another remedy is available to the victim (Art. 6:102 PEL).

Such amount of disagreement between the two texts is indeed not surprising, considering the uncertainty that exists, even among law and economics scholars, as to what tort law model

(s) would best fit the European legal framework (van den Bergh and Visscher 2006, pp. 513–4, 521–40). Rather than the foreseeable differences of opinion between the PEL and the PEL on these issues, what is more interesting to note, in our perspective, is a critique of the methodology and policy agenda that both the projects have embraced. While both the PETL and the PEL, as scholarly by-products, have the merit of offering a concrete basis for the debate on the prospects of European tort law, their future implications are much less straightforward. Leaving aside the doubts about the legitimacy of the jurists involved in these initiatives to act as legislators (Wagner 2005, pp. 1283–4), as well as any concern on the feasibility and desirability of a European codification of tort law (on which see above, section “The Variety of European Tort Laws”), what should be stressed here is that the adoption of either set of rules would require practitioners, judges, and scholars in national jurisdictions to become accustomed to a completely new tort law pattern and to develop methods and techniques that would, more or less, be foreign to their legal tradition. At least in the short run, the lack of such traditional accumulation of methods and techniques would likely drive jurists not accustomed to that pattern to keep relying on their own legal culture and on their traditional repertoire of solutions and technicalities. Thus, a fundamental objection to the top-down imposition of, the not-so-common, principles is the risk of perpetuating the divergence that the harmonization process is precisely aimed to reduce (Bussani 2007a, p. 369).

Building Knowledge: The *Ius Commune* Casebooks for the Common Law of Europe and the Common Core of European Private Law Projects

Aware that solutions to the above issues are critical, other scholars pursued another path. “Knowledge-building” enterprises share the common view that top-down harmonization cannot be undertaken without the collateral support of bottom-up initiatives. Therefore, the real instrument and target for those who are seeking the establishment of a truly European tort law should

be the development of a common legal culture, based on as much knowledge as possible of the legal experience of each European jurisdiction.

Irrespective of the uses to which knowledge may be applied, which may or may not include the pursuit of legal harmonization, knowledge building is both the starting point and the final aim of two projects whose scope is broader than the ones we just examined, insofar as their focus goes beyond tort law only. These two projects are the *Ius Commune Casebooks for the Common Law of Europe* and the *Common Core of European Private Law*.

The *Ius Commune Casebooks* initiative was launched in 1994 by Professor Walter van Gerven with the aim of producing a collection of casebooks covering each of the main fields of European law (on the aims and methods of such a project, see van Gerven 2002, 2007; Larouche 2000, 2001; van Gerven 1996). The long-term purpose of the *Ius Commune* project is to “uncover common general principles which are already present in the living law of the European countries” for the benefit of European students (van Gerven et al. 2000, p. 68). In this view, the casebooks are primarily conceived as teaching materials to be used in the curricula of law schools in order to promote a common European education.

As of now, seven volumes have been published by Hart (van Erp and Akkermans 2012; Micklitz et al. 2010; Beale et al. 2010; Schiek et al. 2007; Beatson and Schrage 2003; van Gerven et al. 1998, 2000), and many are forthcoming, on issues as diverse as labor law, law and art, constitutional law, judicial review of administrative action, conflicts of law, and legal history (see casebooks.eu). Two volumes concerning tort law have already been published (van Gerven et al. 1998, 2000); one of them (van Gerven et al. 2000) is under review for the second edition. Every casebook, whose table of contents and materials are partly accessible on the project’s website, collates legislation, excerpts from books, articles, and, above all, cases from various jurisdictions – mostly from France, England, and Germany, which are considered as representative of the main European legal families. Materials

from other legal systems are included in the casebooks only if they present an original solution when compared to the above legal systems. These materials are accompanied by introductory and explicatory notes, stressing the similarities among European legal systems, and the impact of the EU law “as a driving force towards the emergence of a new *ius commune*” (see casebooks.eu/research.php). The casebooks are written by a task force composed of academics representing what are deemed to be the “main” European legal families (van Gerven et al. 2000, vi-vii). A distinctive feature of the project is that each task force member, instead of dealing solely with his/her national legal system, is charged with writing an entire thematic chapter, even if it refers to legal systems different from his/her country of origin or of education. This distribution of work guarantees that the final outcome is not a patchwork of national reports but rather the genuine result of a truly comparative effort.

The *Common Core of European Private Law* project, led since 1994 by Professors Mauro Bussani and Ugo Mattei, has a different target audience, methodology, and primary goal. The initiative aims to unearth what is common, and what is not, between the EU Member States’ private laws, in order to provide a reliable description of the actual state of the art of the European multi-legal framework (for a general overview of the project, see Bussani et al. 2009; Bussani and Mattei 1998, 2000, 2003, 2007; Kasirer 2002, p. 417; Bussani 1998).

Unlike the *Ius Commune Casebook* project, which emphasizes the solutions given by the legal systems considered to be leading or paradigmatic, the *Common Core* project focuses equally on all the EU national legal systems. The research carried out under the *Common Core* initiatives is published as volumes in a dedicated series by Cambridge University Press (although some books have also been published by Stämpfli and Carolina Academic Press). Of the fifteen volumes published so far (Hondius and Grigoleit 2014; van der Merwe and Verbeke 2012; Brüggemeier et al. 2010; Hinteregger 2008; Cartwright and Hesselink 2008; Möllers and Heinemann 2008; Bussani and Mattei 2007; Pozzo 2007; Graziadei

et al. 2005; Sefton-Green 2005; Werro and Palmer 2004; Kieninger 2004; Bussani and Palmer 2003; Gordley 2001; Zimmermann and Whittaker 2000), five deal with civil liability issues, such as recoverability of pure economic losses (Bussani and Palmer 2003), protection of personality rights (Brüggemeier et al. 2010), boundaries of strict liability (Werro and Palmer 2004), ecological damage (Hinteregger 2008), and pre-contractual liability (Cartwright and Hesselink 2008). Two other volumes on tort law, causation, and products liability, respectively (see common-core.org), are under preparation.

On the shoulders of Rudolf B. Schlesinger’s and Rodolfo Sacco’s path-breaking research (Schlesinger 1968, 1995; Sacco 1991), the *Common Core* adopts a method based on the so-called factual approach as underpinned by the dissociation of legal formants. At the very foundation of the *Common Core*’s efforts, there is the following set of assumptions. Legal systems are not always made up of a coherent set of principles and rules, as domestic jurists tend to assume. What the rules say often does not relate to what the very same rules translate to in practice. Legal actors at work in the same legal system – the bar, the bench, scholars, lawmakers, insurers, bureaucrats, and so on – do not always offer the same account and interpretation of what legal rules are. Further, circumstances that are officially ignored and considered to be irrelevant in the application of the law often operate in secrecy, slipping in silently between what law is said to be and how it is actually applied (Bussani and Mattei 1998, pp. 343–5).

The need to delve into the depths of legal systems, and to obtain comparable answers from jurists from different jurisdictions, led the *Common Core*’s General Editors to adopt questionnaires as their key methodological tool. The research work develops as follows: when a topic is chosen, the person charged with the task of editing a volume on that particular topic drafts a factual questionnaire and distributes it among the national rapporteurs participating in the project. The questionnaire has a sufficient degree of specificity as to require respondents to address all the factors that have a practical impact on the legal

system they are analyzing. This method also guarantees that national responses surface how apparently identical black-letter rules may in fact produce different applications. The use of fact-based questionnaires also allows understanding whether and to what extent solutions depend on legal rules outside the private law field, such as procedural mechanisms, administrative schemes, or constitutional provisions, or on other factors, e.g., policy considerations, economic reasons, and social values, affecting the law-in-action level (Bussani and Mattei 1998, pp. 351–4). Responses to the questionnaires are publicly analyzed, discussed, compared, and subsequently collected by the topic editor of a volume in a book, which is published in the *Common Core* series.

There are many data that attest to the success of these two initiatives. The *Ius Commune* casebooks have been adopted in the curricula of European law faculties and have been quoted by national courts in their domestic tort law opinions (see some illustrations on casebooks.eu/project/aim/). *Common Core* volumes are widely cited by judges and scholars across and beyond European jurisdictions, and some of them have even been translated in languages other than English (see, e.g., Bussani and Palmer 2005). Yet, besides the number of books published and the amount of quotations these books get, the most important outcomes of the *Ius Commune* and the *Common Core* projects lie in the acculturation processes they trigger and in the transformations they help to bring about in European legal systems. These processes and transformations run deep in European legal cultures – so deep that their effects, although not easy to quantify in the short run, promise to be the most long-lasting ones in the long run.

Conclusions

Any assessment of the harmonization perspective in a given field must indeed take into account the time factor.

There can be no doubt that European tort laws, as defined by statutory texts, judicial precedents, and/or scholarly writings, and ultimately applied by courts to the disputes brought before them,

display many lines of convergence. Yet tort law does not live in parliaments, law firms, courts, and law books only. It also lives “in the shadow” of the official systems of adjudication. It lives in the offices of insurance companies, which provide coverage for damages caused by the insured or third parties. It lives in people’s notions about injury and risk, responsibility, and justice, determining people’s conducts in day-to-day activities and their litigation/non-litigation choices once a wrong has occurred. Tort law lives in languages, concepts, and images associated with law in mass-generated popular culture – newspapers, television, movies, and novels – as well as in public debates about what values should be protected and promoted, at what costs, and at whose expense (Bussani and Infantino 2015; Oliphant 2012). This stratified set of individual and mass responses to injury, risk, and responsibility does not stand still. Its different constituents may change over time, moving in the same place at different speeds and in different directions (Bussani 2007a, pp. 369–70).

When dealing with the prospects of a soon-to-be harmonized European tort law framework, one should take time seriously. Harmonization efforts – whether hard or soft, legislative or scholarly driven, or aimed at immediate uniformity or at knowledge building – usually trigger a variety of transformations in the layers of a legal system. The time, speed, and direction of these transformations can widely diverge and are usually very difficult to predict, as is the way in which the transformations affecting one layer will interact with other layers of the system. The only dynamic that can be easily foreseen is that the more abrupt the requested change and the larger its departure from the legal culture where it should take place, the greater would be the time and cost needed by that legal system to adjust to the new framework – no matter how efficient it would be.

References

- Banakas S (2002) European tort law: is it possible? *Eur Rev Private Law* 3:363–375
- Beale H, Kötz H, Hartkamp A, Tallon D (eds) (2010) *Contract law*, 2nd edn. Hart, Oxford-Portland

- Beatson J, Schrage EJH (eds) (2003) Unjustified enrichment. Hart, Oxford-Portland
- Blackie J (2003) Tort/Delict in the work of the European civil code project of the study group on a European civil code. In: Zimmermann R (ed) Grundstrukturen des Europäischen Deliktsrechts. Nomos, Baden-Baden, pp 133–146
- Blackie J (2007) The torts provisions of the study group on a European civil code. In: Bussani M (ed) European tort law: Eastern and Western perspectives. Stämpfli, Berne, pp 55–80
- Blackie J (2009) The provisions for ‘non-contractual liability arising out of damage caused to another’ in the draft common frame of reference. King’s Law J 20:215–238
- Brügemeier G (2004) Common principles of tort law: a pre-statement of the law. BIICL, London
- Brügemeier G, Colombi Ciacchi ALB, O’Callaghan P (eds) (2010) Personality rights in European tort law. CUP, Cambridge
- Bussani M (1998) Current trends in European comparative law: the common core approach. Hastings Int Comp Law Rev 21:785–801
- Bussani M (2007a) European tort law: a way forward? In: European tort law. Eastern and Western perspectives. Stämpfli, Berne, pp 365–379
- Bussani M (ed) (2007b) European tort law. Eastern and Western perspectives. Stämpfli, Berne
- Bussani M (2007c) The geopolitical role of a European contract code. In: Boele-Woelki K, Grosheide FW (eds) The future of European contract law. Liber Amicorum E.H. Hondius Kluwer, The Hague, pp 43–52
- Bussani M, Infantino M (2015) Tort law and legal cultures. Am J Comp L 63: forthcoming
- Bussani M, Mattei U (1998) The common core approach to European private law. Colum J Eur Law 3:339–356
- Bussani M, Mattei U (2000) Le fonds commun du droit privé européen. Rev Int Dr Comp 52:29–48
- Bussani M, Mattei U (eds) (2003) The common core of European private law. Essays on the project. Kluwer, The Hague
- Bussani M, Mattei U (eds) (2007) Opening up European law. Stämpfli, Berne
- Bussani M, Palmer VV (2003) Pure economic loss in europe. CUP, Cambridge
- Bussani M, Palmer VV (2005) Pure economic loss in Europe. Law Press China Beijing (in Chinese)
- Bussani M, Werro F (eds) (2009) European private law: a handbook, vol I. Stämpfli, Berne
- Bussani M, Infantino M, Werro F (2009) The common core sound: short notes on themes, Harmonies and Disharmonies in European tort law. King’s Law J 20:239–255
- Cartwright J, Hesselink M (eds) (2008) Precontractual liability in European private law. CUP, Cambridge
- Eidenmüller H, Faust F, Grigoleit HC, Jansen N, Wagner G, Zimmermann R (2008) The common frame of reference for European private law: policy choices and codification problems. Ox J Leg Stud 28:659–708
- European Commission Directorate-General for Internal Market and Services (2007) Handbook on implementation of the services directive. Available at http://ec.europa.eu/internal_market/services/docs/services-dir/guides/handbook_en.pdf
- European Group of Tort Law (2005) Principles of European tort law. Springer, Wien
- Faure M (2003) How law and economics may contribute to the harmonisation of tort law in Europa. In: Zimmermann R (ed) Grundstrukturen des Europäischen Deliktsrechts, Nomos, Baden, pp 31–82
- Fauvarque-Cosson B (2001) Comparative law and conflict of laws: allies or enemies? New perspectives on an old couple. Am J Comp Law 49:407–427
- Giliker P (2009) Can 27(+) ‘Wrongs’ make a right? The European tort law project: some skeptical reflections. King’s Law J 20:257–279
- Giliker P (2010) Vicarious liability in tort. A comparative perspective. CUP, Cambridge
- Gordley J (ed) (2001) The enforceability of promises in European contract law. CUP, Cambridge
- Graziadei M, Mattei U, Smith L (eds) (2005) Commercial trusts in European private law. CUP, Cambridge
- Hartlief T (2002) Harmonization of European tort law. Some critical remarks. In: Faure M, Smits J, Schneider H (eds) Towards a European ius commune in legal education and research, Intersentia, Antwerp, pp 225–236
- Hay P (2007) Contemporary approaches to non-contractual obligations in private international law (conflict of laws) and the European community’s “Rome II” regulation. Eur Legal F 7:137–151
- Hesselink MW (2004) The politics of a European civil code. Eur Law J 10:675–697
- Hinteregger M (ed) (2008) Environmental liability and ecological damage in European law. CUP, Cambridge
- Hondius E, Grigoleit HC (eds) (2014) Unexpected circumstances in contract law, CUP, Cambridge
- Howells G (2008) Is European product liability harmonised? In: Koziol H, Schulze R (eds) Tort law of the European community. Springer, Wien, pp 121–134
- Infantino M (2010) Making European tort law: the game and its players. Cardozo J Int Comp Law 18:45–87
- Infantino M (2012) La causalità nel diritto della responsabilità extracontrattuale, Stämpfli, Berne
- Kasirer N (2002) The common core of European private law in boxes and bundles. Eur Rev Private Law 10:417–437
- Kieninger E-M (ed) (2004) Security rights in movable property in European private law. CUP, Cambridge
- Koch BA (2007) The “principles of European tort law”. ERA F 8:107–115
- Koch BA (2009) Principles of European tort law. King’s Law J 20:203–214
- Koch BA, Koziol H (eds) (2002) Unification of tort law: strict liability. Kluwer, The Hague
- Koziol H (ed) (1998) Unification of tort law: wrongfulness. Kluwer Law International, The Hague
- Koziol H (2007) Comparative law – A must in the European union: demonstrated by tort law as an example. J Tort Law 1:1–10

- Larouche P (2000) *Ius commune casebooks for the common law of Europe: presentation, progress, rationale*. *Eur Rev Private Law* 8:101–109
- Larouche P (2001) *L'intégration, les systèmes juridiques et la formation juridique*. *McGill Law J* 46:1011–1036
- Magnus U (ed) (2001) *Unification of tort law: damages*. Kluwer, The Hague
- Magnus U (2002) Towards European civil liability. In: Faure M, Smits J, Schneider H (eds) *Towards a European ius commune in legal education and research*. Maklu, Antwerpen, pp 205–24
- Magnus U (2010) Why is US tort law so different? *J Eur Tort Law* 1:102–124
- Magnus U, Martin Casals M (eds) (2004) *Unification of tort law: contributory negligence*. Kluwer, The Hague
- Markesinis B (2003) *Comparative law in the courtroom and classroom: the story of the last thirty-five years*. Hart, Oxford-Portland
- Markesinis B, Unberath H (2002) *The German law of torts*, 4th edn. Hart, Oxford-Portland
- Menyhárd A (2009) Eastern tort law. In: Bussani M, Werro F (eds) *European private law: a handbook*, vol I. Stämpfli, Berne, pp 337–375
- Micklitz HW, Stuyck J, Terryn E, Droshout D (eds) (2010) *Consumer law*. Hart, Oxford-Portland
- Möllers TMJ, Heinemann AB (eds) (2008) *The enforcement of competition law in Europe*. CUP, Cambridge
- Ogus A (1999) Competition between national legal systems: a contribution of economic analysis to comparative law. *Int Comp Law Quart* 48:405–418
- Oliphant K (2009) Introduction: European tort law. *King's Law J* 20:189–202
- Oliphant K (ed) (2010) *Aggregation and divisibility of damage*. Springer, Vienna
- Oliphant K (2012) Cultures of tort law in Europe. *J Eur Tort Law* 3:147–157
- Palmer VV, Bussani M (2008) *Pure economic loss: new horizons in comparative law*. Routledge-Cavendish, London-New York
- Pozzo B (ed) (2007) *Property and environment*. Stämpfli, Berne
- Quézel-Ambrunaz C (2012) *Essai sur la causalité en droit de la responsabilité civile*. Dalloz, Paris
- Reimann M (2003) Product liability in a global context: the hollow victory of the European model. *Eur Rev Private Law* 2:128–154
- Reimann M (2012) General report. In: Id (ed) *Cost and fee allocation in civil procedure*. Springer, Dordrecht, pp 3–56
- Reimann M (2014) Products liability. In: Bussani M, Sebok AJ (eds) *Comparative tort law: a global approach*. Elgar, Cheltenham, forthcoming
- Rogers HWW (ed) (2004) *Unification of tort law: multiple tortfeasors*. Kluwer, The Hague
- Sacco R (1991) Legal formants: a dynamic approach to comparative law. *Am J Comp Law* 39:1–34, 343–401
- Schepel H (2007) The European brotherhood of lawyers: the reinvention of legal science in the making of European private law. *Law Soc Inquiry* 32:183–199
- Schick D, Waddington L, Bell M (eds) (2007) - *Non-discrimination law*. Oxford-Portland
- Schlesinger RB (ed) (1968) *Formation of contracts: a study of the common core of legal systems*. Oceana Publications, Dobbs Ferry
- Schlesinger RB (1995) The past and future of comparative law. *Am J Comp Law* 43:477–481
- Sefton-Green R (ed) (2005) *Mistake, fraud and duties to inform in European contract law*. CUP, Cambridge
- Spier J (ed) (1998) *The limits of expanding liability: eight fundamental cases in a comparative perspective*. Kluwer, The Hague
- Spier J (ed) (2000a) *The limits of liability: keeping the floodgates shut*. Kluwer, The Hague
- Spier J (ed) (2000b), *Unification of tort law: causation*. Kluwer, The Hague
- Spier J (ed) (2003) *Unification of tort law: liability for damage caused by others*. Kluwer, The Hague
- Stapleton J (2007) Benefits of comparative tort reasoning: lost in translation. *J Tort Law* 1:1–45
- Ussing H (1952) The Scandinavian law of torts – impact of insurance on tort law. *Am J Comp Law* 1:359–372
- van Boom WH (2009) Harmonizing tort law. a comparative tort law and economics analysis. In: Faure M - (ed) *Tort law and economics*, 2nd edn. Elgar, Cheltenham, pp 435–450
- van Boom WH, Pinna A (2008) Le droit de la responsabilité civile de demain en Europe: questions choisies. In : Winiger B (ed) *La responsabilité civile européenne de demain—projets de révision nationaux et principes européens*. Bruylant, Bruxelles, pp 261–277
- van Dam C (2006) *European tort law*, 1st edn. OUP, New York
- van Dam C (2013) *European tort law*, 2nd edn. OUP, Oxford
- van den Bergh R, Visscher L (2006) The principles of European tort law: the right path to harmonization? *Eur Rev Private Law* 14:511–543
- van der Merwe C, Verbeke A-L (eds) (2012) *Time-limited interests in land*. CUP, Cambridge
- van Erp S, Akkermans B (eds) (2012) *Property law*. Hart, Oxford-Portland
- van Gerven W (1996) *Casebooks for the common law of Europe: presentation of the project*. *Eur Rev Private Law* 4:67–70
- van Gerven W (2002) Comparative law in a regionally integrated Europe. In: Harding A, Örüçü E (eds) *Comparative law in the 21st century*. Kluwer, The Hague, pp 155–178
- van Gerven W (2007) A common framework of reference and teaching. In: Bussani M (ed) *European tort law: Eastern and Western perspectives*. Stämpfli, Berne, pp 345–364
- van Gerven W (2009) Judicial convergence of laws and minds in European tort law and related matters. In: Colombi Ciacchi A, Godt C, Rott P, Smith LJ (eds) *Haftungsrecht im dritten millennium – liability in the third millennium*. Baden-Baden, Nomos, pp 29–47
- van Gerven W, Lever L, Larouche P, von Bar C, Viney G (1998) *Tort law – scope of protection*. Hart, Oxford-Portland
- van Gerven W, Lever J, Larouche P (2000) *Tort law*. Hart, Oxford-Portland

- von Bar C (1998) The common European law of torts, vol I. OUP, Oxford
- von Bar C (1999) A European civil code, international agreements and European directives (European parliament, directorate general for research, working paper, JURI 103 EN, 1999), Available at http://www.europarl.europa.eu/workingpapers/juri/pdf/103_en.pdf, pp 133–8
- von Bar C (2000) The common european law of torts, vol II. OUP, Oxford
- von Bar C (2001) Le groupe d'études sur un code civil européen. *Rev Int Dr Droit Comp* 53:127–139
- von Bar C (ed) (2009) Non-contractual liability arising out of damage caused to another. Sellier, Munich
- von Bar C, Drobnič U (2004) The interaction of contract law and tort and property law in Europe. A comparative study. Sellier, Munich
- Wagner G (2005) The project of harmonizing European tort law. *Common Market Law Rev* 42:1269–1312
- Wagner G (2009) The law of torts in the draft common frame of reference. In: Id (ed) The common frame of reference: a view from law & economics. Sellier, Munich, pp 225–271
- Werro F, Palmer VV (eds) (2004) The boundaries of strict liability in European tort law. Carolina Academic Press, Durham
- Widmer P (ed) (2005) Unification of tort law: fault. Kluwer, The Hague
- Zimmermann R (ed) (2003a) Grundstrukturen des Europäischen Deliktsrechts. Nomos, Baden-Baden
- Zimmermann R (2003b) Principles of European contract law and principles of European tort law: comparison and points of contact. In: Koziol H, Steininger B (eds) European tort law yearbook 2002. Springer, Wien, pp 2–31
- Zimmermann R (2009) The present state of European private law. *Am J Comp Law* 57:479–512
- Zimmermann R, Whittaker S (eds) (2000) Good faith in European contract law. CUP, Cambridge

Harmonization: Consumer Protection

Franziska Weber

Faculty of Law, Institute of Law and Economics,
University of Hamburg, Hamburg, Germany

Abstract

This contribution gives an overview of law and economic works on consumer protection. It does so by taking three complementary angles: (1) assessing harmonization efforts regarding

consumer protection in Europe, (2) presenting selected topics of substantive consumer law, and (3) analyzing the European consumer law enforcement from an economic point of view.

Synonyms

Consumer law; Consumer legislation; Consumer policy

Introduction

Consumer protection laws refer to a set of rules aimed at stipulating and guarding consumers' rights. It is a challenging field and requires knowledge of different legal areas (contract law, tort law, regulatory law, competition law, etc.) because consumer problems by nature lie at the borderline of private/public problems and social/commercial ones. Substantive consumer laws encompass public and private law, which is why redress is necessarily made to both public and private enforcement instruments. Whereas harmonization efforts at the European level have historically primarily concerned substantive consumer laws, European legislation increasingly aligns procedural laws also.

This contribution is structured as follows: Section “[The Legal Approach and Its Economic Analysis: Research on Consumer Behavior](#)” gives a short introduction to research into consumer behavior. Section “[Harmonizing the European Consumer Law](#)” starts with an economic analysis of harmonizing consumer law throughout Europe. It then brings into focus selected topics of consumer protection more specifically. On top of the substantive law dimension, a condensed economic analysis of the European consumer law enforcement is presented. Section “[Conclusion](#)” concludes.

The Legal Approach and Its Economic Analysis: Research on Consumer Behavior

The consumer movement started developing in the 1960s and 1970s. Advocates of this movement

I wish to thank Roger van den Bergh for the valuable comments on an earlier version of this contribution.

took the consumers' weaker position for granted and, therefore, strongly favored strengthening consumers' rights. Up until today the paternalistic consumer protection argument continues to be used when it comes to justifications for passing legislation in the field of consumer law (Micklitz et al. 2010, p. 1).

Whereas the term protection hence implicates a paternalistic element, an economic insight into consumer markets is the fact that increased consumer protection comes at the cost of higher product prices (Faure 2008). Typically an intervention in the market will lead to costs for the producer, such as increased expectancy of damage payments. It is therefore straightforward that these costs will be reflected in the product's price. Consequently, any legal intervention needs to be justified by the existence of a market failure. The most commonly occurring market failure in consumer law is information asymmetries, such as quality uncertainty (Akerlof 1979; Van den Bergh 2007a). The debate on consumer policy in law and economics in elaboration of the neoclassical approach has taken three different theoretical angles: information economics, new institutional economics, and behavioral economics (Rischkowsky and Döring 2008). In information economics the notion of asymmetric information is prevalent. The consumer is regarded as unable to appropriately perceive quality differences (Stigler 1961; Stiglitz 2000). A lack of information on the consumers' side and the resulting information costs impact on the consumers' decision-making process. This may be cured by providing consumers with more information. New institutional economics, on the other hand, acknowledges that the pure provision of information is not sufficient and calls for regulation of certain institutions (e.g., contracts). It expands the focus to looking at institutional arrangements in place to cure information asymmetries and deals more closely with specific rules. New institutional economics is more broadly concerned with transaction costs (Williamson 1975) – not only information costs – that impact market transparency and consumer behavior. Another core notion of this stream of economics relevant also to consumer behavior is that of “bounded rationality”

(Simon 1957). According to this concept individuals, even if presented with all necessary information, are unable to correctly process it due to the limited capacity of the human brain to do so. Behavioral economics challenges the rational choice theory even more radically by showing that consumers – even when exposed to complete information – cannot successfully process it (Kahneman et al. 1982; Sunstein and Thaler 2009). Consumers are shown to systematically deviate from rational behavior. Behavioral economics provides insights on how consumers perceive and use available information and act in the presence of choice. Perception is dependent on consumers' internal constraints in the form of cognitive, emotional, and situational factors (Luth 2010). A notion put forward within behavioral economics is that of libertarian paternalism (Thaler and Sunstein 2003) – a form of paternalism that protects people from making bad choices by not employing coercion.

Consumer research is being carried out within all of these disciplines, with behavioral economics being the most recent and most popular one (Vandenberghe 2011; Sunstein and Reisch 2014). In the last years policy-makers and law-makers at the European or Member State (MS) level and also, for instance, in the United States start to take more account of such evidence. Whereas insights are regarded as very valuable to understand real consumer behavior, some challenges need to be overcome still before strong normative conclusions may be drawn. There is, for instance, evidence of cognitive biases pointing in opposite directions; furthermore behavioral economics keeps lacking a framework to systematically use its insights for policy-making (Vandenberghe 2011). It is an unresolved matter how general conclusions can be drawn from a limited set of experiments. Note should be taken of the fact that marketing research has long dealt with better understanding consumer behavior (for an early contribution, see Kotler 1965).

Given its importance in the European context, it should be mentioned that much of the consumer law research is carried out from the angle of comparative law and economics, as the name suggests the intersection between comparative law and the

economic analysis of law (e.g., Faure et al. 2009; Kovac 2011; Weber 2014).

Harmonizing the European Consumer Law

Once an economist has identified a market failure, this paves the way to some type of – cautious or intrusive – legal intervention. One decision to be taken is, consequently, the type of legal intervention – e.g., imposition of information duties. A layer additionally complicating the nature of passing consumer laws in Europe is the availability of various lawmakers: For the European Union (EU), the competences are generally either situated at the MS level or at the Union level. Not all Union legislation, furthermore, leads to the same degree of unification within Europe. A starting point would be whether a regulation, directive, or other instrument is chosen; legal instruments can aim at different levels of harmonization.

The legal basis stipulating law-making competences does not necessarily coincide with economic theory on the desirable level of law-making powers at all times. The trade-off between centralized and decentralized decision-making has been assessed from the point of view of “economics of federalism” (Oates 1972) and “regulatory competition” (Ogus 1999; Van den Bergh 1994, 2000, 2002), also specifically for the European consumer law (Van den Bergh 2007b; Faure 2008). The concepts can be used to give meaning to the principle of subsidiarity (Van den Bergh 1994) and identify the adequate level of law-making. Competition between different laws is regarded as desirable as it enables citizens to choose the jurisdiction according to their preferences (Tiebout 1956). Differences, furthermore, may facilitate learning between the states. Decentralized rule-making may have advantages in terms of being better able to tailor legislation to preferences, and local authorities may have an information advantage regarding local specificities (Van den Bergh 2007b). Reasons to harmonize on the other hand are the need to internalize interstate externalities, the potential danger of a race to the bottom, economics of scale,

and the reduction of transaction costs. Another factor to be taken into consideration is political distortions that depend on the strength of lobby groups at the various levels (Van den Bergh 2002).

The findings will differ for the area of law considered. Regarding the case of consumer protection laws, the discussion on the adequate level of law-making and the need for harmonization was recently steered up again by the Consumer Rights Directive (Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, OJ L 304, 22.11.2011, pp. 64–88). According to the original draft proposal, four directives previously aiming at minimum harmonization were to be changed into maximum harmonization (Faure 2008). (The directives concerned are Directive 85/577/EEC on contracts negotiated away from business premises, Directive 93/13/EEC on unfair terms in consumer contracts, Directive 97/7/EC on distance contracts, and Directive 1999/44/EC on consumer sales and guarantees.) This original proposal triggered heavy criticism by law and economics scholars (Van den Bergh 2007b; Faure 2008), providing strong arguments against full harmonization and instead pleading in favor of diversity: mainly because preferences of consumers in the MS diverge a lot, not least based on differences in income levels. The final text of the Consumer Rights Directive does not claim maximum harmonization as strongly as the first draft. It is said to having been watered down to “some kind of quite undefined ‘targeted’ (i.e., partial or incomplete) full harmonisation” (Gomez and Ganuza 2012). Its content concerns primarily information duties for distance, off-premises, and other contracts and the right of withdrawal for distance and off-premises contracts (see recital 9).

Substantive Law Dimension

There is a growing body of research discussing the costs and benefits of consumer protection laws (Cseres 2012). Recent collections of law and economics work on contract law were published in the context of discussing the Draft Common Frame of Reference (Wagner 2009; Larouche and Chirico 2010); some contributions deal more specifically with consumer protection clauses

(Luth and Cseres 2010 and for the Common European Sales Law, Bar-Gill and Ben-Shahar 2013).

Much of consumer law is written as mandatory laws; other consumer protection provisions are stipulated as default rules. From a legal point of view, there is no clear normative stance identifying a preference for one type over the other; apparently the protection aspect plays an important role. From a law and economics point of view, certain costs and benefits of both types of rules have been identified. It is crucial that mandatory rules prevent “sharper deals forgone” by consumers who would have abstained from enjoying certain types of costly consumer protection rules; however, they are prohibited by law from doing so (Mackaay 2013, p. 423). This factor, on top of the costs of enacting and enforcing such a rule, needs to be weighed against potential benefits. For default rules the costs are different as opting out of the default is possible and may lead to costs (Mackaay 2013). In addition, the default may end up being unsuited or deviation may practically be impossible. Law and economics literature has dealt extensively with the question of what the default should be (Ayres 2012). In the first wave of this literature, the view prevailed that the default rule should be the one that the majority would have chosen as this would reduce total costs. A second stream of literature concerned a debate on whether there is such a thing as a penalty default or information-forcing default. Such a rule basically punishes the non-deviating party by, for instance, granting his/her a high interest rate only if he/she does not reveal certain information. Lately, not least under the influence of behavioral law and economics, the debate has turned toward the design of different default rules – personal default rules or those enabling active choice (Sunstein and Reisch 2014) – and the notion of “altering rules,” i.e., the rules that set out how one can deviate from the default option (Ayres 2012). Cognitive biases are used to illustrate how certain types of formulations and conditions to deviate from the default affect behavior. Some rules can be regarded as “sticky,” i.e., choice is only an illusion. These findings play in the context of all types of mandatory and default rules that may be relevant for consumer protection.

Standard contract terms in consumer contracts have specifically attracted scholar’s attention. There are empirical studies confirming that consumers do not read these terms (e.g., Bakos et al. 2014) and publications discussing the consequences of this insight for policy-making (Luth 2010). An empirical study regarding the content of such terms in the context of software companies, for instance, confirms a bias toward terms favoring the contract writer – in this case the software company (Marotta-Wurgler 2007). The basic market failure asserted is that of an information asymmetry (Schäfer and Leyens 2010). Due to the consumer not being able to observe the terms and condition’s quality, quality may overall decrease, driving good terms out of the market. A consequence that may be drawn is that a judicial control of standard terms is desirable for certain types of contracts.

Further research concerns the costs side of a right of withdrawal (Eidenmüller 2011). One such cost may be a potential moral hazard problem (Rekaiti and Van den Bergh 2000). Unsurprisingly, given that a lot emphasis is put on the problem of information asymmetries between consumers and the opposite party, information disclosure rules are an important aspect of consumer protection law (Beales et al. 1981; Eisenberg 2003; De Geest and Kovac 2009, and from a comparative law and economics point of view: Kovac 2011). Law and economics has furthermore dealt with the topic of warranty contracts and their different functions being that of insurance, a quality signal and acting as an incentive to provide high quality (Emons 1989; Werth 2011). The classical consumer topic “unfair commercial practices” is dealt with from an economic point of view by Gómez (2006). Within the scope of this short overview, it is not possible to set out all the research that has been carried out (see for a broader overview Cseres 2012). Suffice it to say that also advertising and banking laws have attracted scholarly attention.

Law Enforcement Dimension

From an economic viewpoint, the threat of enforcement steers people’s behavior, more particularly, their incentives to obey the law. The

interplay between substantive laws and their enforcement forms the incentives and deterrents that induce law-abiding behavior. Law and economics scholars traditionally discuss enforcement matters from the angle of *Becker's* deterrence theory developed for criminal law (Becker 1968). According to the deterrence theory, a rational wrongdoer can be induced not to violate the law if the sanction for a potential wrongdoing (multiplied by the probability of detection and conviction) is at least as large as the benefit he/she could obtain from committing the wrong. Meanwhile, this approach has been expanded to law enforcement more broadly.

There are a number of different enforcement mechanisms at play in consumer law, and group litigation is a powerful procedural tool. Law and economics scholarship is long to realize that, for instance, classical public and private enforcement schemes clearly each have a number of economic strengths and weaknesses (Shavell 1993; Van den Bergh 2007a; Faure et al. 2009; Weber 2014). Crucial economic factors at play when assessing whether individuals will initiate a lawsuit or report a wrong are the following: Due to “rational apathy” an individual will not act if the costs of doing so outbalance his/her benefits, for instance, when harm is very small and the investment to enforce the law is costly (Van den Bergh 2007a). If societal harm is nevertheless large, no action may be taken because of a divergence between the individual and social incentive to sue (Shavell 1997). The other extreme is “frivolous lawsuits” that are not based on merits and socially not desirable. Another possible distortion of incentives concerns “free-riding” problems. This problem can occur if in certain situations, in which many victims suffer from a law infringement but all gain as soon as one of them sues, it is efficient for everybody to wait for someone else to sue and then profit from the result (Van den Bergh and Visscher 2008a, p. 14). As mentioned information asymmetries are the core market failure in consumer law, e.g., unobservable characteristics of a consumer good. For the enforcement side it is furthermore true that they are one of the key triggers of litigation. Law enforcement mechanisms/procedures are therefore also discussed regarding the extent to which

they mitigate these asymmetries by generating information. Investigative powers are a main means to achieving this. Continuing to think along the lines of enforcer's incentives to carry out tasks as desired, literature discusses capture as an incentive problem, meaning the exertion of influence on public administration that leads to public officials pursuing, e.g., industry interests (Ogus 1994, p. 57). This concept can be expanded to other enforcers. Principal-agent problems are another matter discussed in this literature. In these relationships the principal (e.g., a client) cannot fully control the quality of the agent's (e.g., lawyer's) performance (Shavell 1979). The basis for any principal-agent problem is an information asymmetry between two parties, which can lead to moral hazard. Lastly, two additional types of costs are decisive: Error costs refer to courts taking mistaken decisions. Error costs can generally be divided in two groups: Error I costs are those that occur when an individual who is guilty might mistakenly not be found liable (“mistaken acquittal”). Error II costs on the other hand occur if an innocent individual might mistakenly be found liable (“mistaken conviction”). Administrative costs then refer to all costs incurred by having a particular enforcement system in place, which in reality is challenging to measure.

By way of example, private law enforcement, if litigation costs are substantial, may suffer from people being rationally apathetic, and depending on the remedy that is sought, “free-riding” may be an issue. Public law enforcement scores highly when it comes to reducing information asymmetries by having investigative and monitoring powers in place. It is generally differentiated between judges that are less likely to be captured compared to public officials, self-regulatory institutions, or certain group representatives. Group litigation, if designed well, is able to reduce rational apathy problems, same as “free-riding” problems. However, capture may be an issue, same as a more complex principal-agent structure and high administrative costs. Criminal law on the other hand may deter even judgment-proof wrongdoers, those that are insolvent, by unique sanctions such as imprisonment (Bowles et al. 2008). Again, administrative costs are

presumably high, same as investigative powers would be. Error costs are discussed as a concern in particular for administrative law enforcement as opposed to criminal law enforcement. In conclusion the optimal solution might be to create a mixture of public and private enforcement that draws upon their comparative advantages (Van den Bergh 2007a; Faure et al. 2009) – the so-called optimal mix of public and private enforcement. Furthermore this mix needs to vary for different sectors of consumer law. Therefore recent research attempts to design optimal enforcement mixes concretely for certain consumer violations by arguing along the lines of the different established economic criteria (e.g., safety laws (Van den Bergh and Visscher 2008b), misleading advertising and package travel (Weber 2014)).

The allocation of enforcement between various mechanisms is one of four crucial parameters of law enforcement, according to law and economics literature (Shavell 1993). The other three are the optimal sanction (injunction, administrative fine, criminal fine), the optimal timing of the intervention (ex ante monitoring and/or ex post enforcement), and the optimal governmental level – either centralized or decentralized – to house the enforcement powers (see also section “[Harmonizing the European Consumer Law](#)”). Needless to say, the parameters are heavily interlinked and cannot be discussed in isolation.

At the EU level various legal initiatives in the context of consumer law enforcement are ongoing: alternative dispute resolution and online dispute resolution (legislative proposal for a Directive of the European Parliament and of the Council on alternative dispute resolution for consumer disputes and amending Regulation (EC) No 2006/2004 and Directive 2009/22/EC (Directive on Consumer ADR), Brussels, 29.11.2011 COM (2011) 793 final and proposal for a Regulation of the European Parliament and of the Council on online dispute resolution for consumer disputes (Regulation on Consumer ODR), Brussels, 29.11.2011 COM(2011) 794 final 2011/0374 (COD)), same as collective redress. (See Commission Recommendation of 11 June 2013 on common principles for injunctive and compensatory collective redress mechanisms in the Member

States concerning violations of rights granted under Union Law OJ L 201, 26.7.2013, p. 60–65, Communication from the Commission “Towards a European Horizontal Framework for Collective Redress”, C(2013) 3539/3.) The EU has, furthermore, strengthened the public law dimension with the Regulation on Consumer Protection Cooperation. (See Regulation (EC) No 2006/2004 of the European Parliament and of the Council of 27 October 2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws (the Regulation on Consumer Protection Cooperation).) These initiatives found their echo in scholarly writing (Keske 2010; Van den Bergh 2013; Wagner 2014) and also stressing the need to take due account of MS’s differing traditions when suggesting EU-wide legislation (Weber 2014). Implementation costs of the EU legislation vary for each MS depending on the prevalent enforcement tradition (Ogus 1999; Faure 2008). Countries are “path dependent” which is why deviating from their tradition may incur costs that need to be weighed against benefits of a new law.

Conclusion

Increasingly, law and economics scholars are looking into consumer protection matters. Research thereby takes various angles. One aspect concerns the questionable desirability of harmonizing consumer laws throughout Europe. Secondly, scholars assess typical consumer law topics, such as standard contract terms or duties of information disclosure from an economic perspective. Thirdly, research is ongoing regarding the fine-tuning, mixing, of different enforcement mechanisms in consumer law. Consumer research is increasingly carried out within behavioral law and economics. Furthermore studies take due account of the importance of comparative law which is particularly crucial in the context of EU law-making. It is important to suggest new European legislation in awareness of existing enforcement traditions in the MS.

In terms of future research, two main challenges can be identified: Firstly, the question on

how far behavioral insights should be translated into policy- and law-making needs answering, and a framework to doing so would need to be established. Secondly, when passing consumer legislation at the European level, it is crucial to identify each time to what extent harmonization is indeed desirable. In this overall cost-benefit assessment, an important factor to be considered is implementation costs which stem from countries' path dependencies. This is a key to avoiding inadequate European legislation.

References

- Akerlof G (1979) The market for lemons: qualitative uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Ayres I (2012) Regulating opt-out: an economic theory of altering rules. *Yale Law J* 121(8):2032–2117
- Bakos Y, Marotta-Wurgler F, Trossen DR (2014) Does anyone read the fine print? Consumer attention to standard-form contracts. *J Legal Stud* 43(1):1–35
- Bar-Gill O, Ben-Shahar O (2013) Regulatory techniques in consumer protection: a critique of European consumer contract law. *Common Mkt Law Rev* 50:109–125
- Beales H, Craswell R, Salop S (1981) The efficient regulation of consumer information. *J Law Econ* 24:491–539
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Bowles R, Faure MG, Garoupa N (2008) The scope of criminal law and criminal sanctions an economic view and policy implications. *J Law Soc* 35(3):389–416
- Cseres K (2012) Consumer protection in the European Union in regulation and economics. In: Van den Bergh R, Paccas A (eds) *Regulation and economics*, vol IX, 2nd edn, *Encyclopaedia of law and economics*. Edward Elgar, Cheltenham, pp 163–228
- De Geest G, Kovac M (2009) The formation of contracts in the draft common frame of reference. *Eur Rev Private Law* 17:113–132
- Eidenmüller H (2011) Why withdrawal rights? *Eur Rev Contract Law* 7:1–24
- Eisenberg MA (2003) Disclosure in contract law. *Calif Law R* 91(6):1648–1691
- Emons W (1989) On the limitation of warranty duration. *J Ind Econ* 37(3):287–301
- Faure MG (2008) Towards a maximum harmonization of consumer contract law?!? *Maastricht J Eur Comp Law* 15(4):433–445
- Faure MG, Ogus AI, Philipsen NJ (2009) Curbing consumer financial losses: the economics of regulatory enforcement. *Law Policy* 31(2):161–191
- Gómez FP (2006) The Unfair Commercial Practices Directive: A Law and Economics Perspective, *European Review of Contract Law*. Vol. 2 No. 1, 4–34.
- Gomez F, Ganuza JJ (2012) How to build European private law: an economic analysis of the lawmaking and harmonization dimensions in European private law. *Eur J Law Econ* 33:481–503
- Kahneman D, Slovic P, Tversky A (1982) *Judgment under uncertainty: heuristics and biases*. Cambridge University Press, New York
- Keske S (2010) *Group litigation in European competition law: a law and economics perspective*. Intersentia, Antwerp
- Kotler P (1965) Behavioral models for analyzing buyers. *J Market* 29(4):37–45
- Kovac M (2011) *Comparative contract law and economics*. Edward Elgar, Cheltenham
- Larouche P, Chirico F (eds) (2010) *Economic analysis of the DCFR – the work of the economic impact group within CoPECL*. Sellier European Law, Munich
- Luth HA (2010) Behavioural economics in consumer policy – the economic analysis of standard terms in consumer contracts revisited. Intersentia, Antwerp
- Luth HA, Cseres K (2010) The DCFR and consumer protection: an economic assessment. In: Larouche P, Chirico F (eds) *Economic analysis of the DCFR – the work of the economic impact group within CoPECL*. Sellier European Law, Munich, pp 35–273
- Mackaay E (2013) *Law and economics for civil law systems*. Edward Elgar, Cheltenham
- Marotta-Wurgler F (2007) What's in a standard form contract? An empirical analysis of software license agreements. *J Empirical Legal Stud* 4:677–713
- Micklitz H-W, Stuyck J, Terryn E (2010) *Cases, materials and text on consumer law*. Hart, Oxford/Portland
- Oates W (1972) *Fiscal federalism*. Harcourt Brace Jovanovich, New York
- Ogus AI (1994) *Regulation: legal form and economic theory*. Clarendon, Oxford
- Ogus AI (1999) Competition between national legal systems: a contribution of economic analysis to comparative law. *Int Comp Law Q* 48:405–418
- Rekaiti P, Van den Bergh RJ (2000) Cooling-off periods in the consumer laws of the EC member states a comparative law and economics approach. *J Consum Policy* 23(4):307–401
- Rischkowsky F, Döring T (2008) Consumer policy in a market economy: considerations from the perspective of the economics of information, the new institutional economics as well as behavioural economics. *J Consum Policy* 31:285–313
- Schäfer H-B, Leyens P (2010) Judicial control of standard terms and European private law. In: Larouche P, Chirico F (eds) *Economic analysis of the DCFR – the work of the economic impact group within CoPECL*. Sellier European Law, Munich, pp 97–119
- Shavell S (1979) Risk sharing and incentives in the principal and agent relationship. *Bell J Econ* 10(1):55–73. The RAND Corporation
- Shavell S (1993) The optimal structure of law enforcement. *J Law Econ* 36:255–288
- Shavell S (1997) The fundamental divergence between the private and the social motive to use the legal system. *J Legal Stud* 16:575–612

- Simon HA (1957) Part IV in models of man. Wiley, New York, pp 196–279
- Stigler GJ (1961) The economics of information. *J Polit Econ* 69:213–225
- Stiglitz J (2000) The contributions of the economics of information to the twentieth century economics. *Q J Econ* 114:1441–1478
- Sunstein CR, Reisch LA (2014) Automatically green: behavioral economics and environmental protection. *Harvard Environ Law* 38(1):127–158
- Sunstein CR, Thaler R (2009) *Nudge* improving decisions about health, wealth, and happiness, 2nd edn. Yale University Press, New Haven
- Thaler R, Sunstein CR (2003) Libertarian paternalism. *Am Econ Rev* 93(2):175–179
- Tiebout CM (1956) A pure theory of local expenditures. *J Polit Econ* 64:416–433
- Van den Bergh RJ (1994) The subsidiarity principle in European Community Law: some insights from law and economics. *Maastricht J Eur Comp Law* 1:337–366
- Van den Bergh RJ (2000) Towards an institutional legal framework for regulatory competition in Europe. *Kyklos* 53(4):435–466
- Van den Bergh RJ (2002) Regulatory competition or harmonization of laws? Guidelines for the European regulator? In: Marciano A, Josselin J-M (eds) *The economics of harmonizing European law*. Edward Elgar, Cheltenham, pp 27–49
- Van den Bergh RJ (2007a) Should consumer law be publicly enforced? In: Van Boom WH, Loos MBM (eds) *Collective enforcement of consumer law*. Europa Law, Groningen, pp 179–203
- Van den Bergh RJ (2007b) The uneasy case for harmonizing consumer law. In: Heine K, Kerber W (eds) *Zentralität und Dezentralität von Regulierung in Europa*. Lucius & Lucius, Stuttgart, pp 184–206
- Van den Bergh RJ (2013) Private enforcement of European competition law and the persisting collective action problem. *Maastricht J Eur Comp Law* 1:12–34
- Van den Bergh RJ, Visscher LT (2008a) The preventive function of collective actions for damages in consumer law. *Erasmus Law Rev* 1:5–30
- Van den Bergh RJ, Visscher LT (2008b) Optimal enforcement of safety law. In: de Mulder RV (ed) *Mitigating risk in the context of safety and security. How relevant is a rational approach?* Erasmus University, Rotterdam, pp 29–62
- Vandenberghe AMIB (2011) Behavioural approaches to contract law. In: De Geest G (ed) *Contract law and economics, encyclopaedia of law and economics*, 2nd edn. Edward Elgar, Cheltenham, pp 401–429
- Wagner G (2009) The common frame of reference: a view from law and economics. Sellier European Law, Munich
- Wagner G (2014) Private law enforcement through ADR: wonder drug or snake oil? *Common Mkt Law Rev* 51:165–194
- Weber F (2014) The law and economics of enforcing European consumer law – a comparative analysis of package travel and misleading advertising, *Markets and the law series*. Ashgate – Gower – Lund Humphries, Farnham, Surrey
- Werth K (2011) Warranties in contract law and economics. In: De Geest G (ed) *Contract law and economics, encyclopaedia of law and economics*, 2nd edn. Edward Elgar, Cheltenham, pp 256–277
- Williamson O (1975) *Markets and hierarchies, analysis and antitrust implications: a study in the economics of internal organization*. Free Press, New York

Harmonization: Legal Enforcement

Katalin J. Cseres

Amsterdam Centre for European Law and Governance, Amsterdam Center for Law and Economics, University of Amsterdam, Amsterdam, The Netherlands

Abstract

While harmonization and convergence of national substantive laws are well advanced in the European Union, a similar convergence and harmonization of the procedural rules and institutional frameworks has not taken place. The implementation, supervision, and enforcement of EU law are left to the Member States in accordance with the so-called national procedural and institutional autonomy. This “procedural competence” of the Member States means that the Member States have to provide remedies and procedures governing actions intended to ensure the enforcement of rights derived from EU law.

Harmonized substantive rules are implemented through diverging procedures and different kinds of enforcement bodies, but this decentralized enforcement challenges the coherent and uniform application of EU law. In an enforcement system where Member States apply divergent procedures, may impose a variety of sanctions and remedies administered by various actors the effectiveness of EU law, effective judicial protection and effective law administration may be at risk. This chapter analyzes the development of EU law

concerning law enforcement and takes a critical look at the EU's aim to harmonize the national procedural rules when EU law is enforced. It will examine the question of which legal basis can be used in order to harmonize procedural rules, whether harmonizing procedural rules would be more efficient than the existing legal diversity and the economics of harmonization will be applied to assess the top-down harmonization by the EU and comparative law, and economics is applied to evaluate the bottom-up voluntary harmonization of the Member States.

Introduction

The implementation, supervision, and enforcement of EU law are left to the Member States in accordance with the so-called national procedural and institutional autonomy. This “procedural competence” of the Member States means that the Member States have to provide remedies and procedures governing actions intended to ensure the enforcement of rights derived from EU law, provided that the principle of equivalence and the principle of effectiveness are observed. (*Case 33/76, Rewe-Zentralfinanz eG and Rewe-Zentral AG v Landwirtschaftskammer für das Saarland (Rewe I)*, [1976] ECR 1989, para 5.) Competence allocation in the European multilevel governance is politically sensitive (Bakardjieva-Engelbrekt 2009; Cafaggi and Micklitz 2009; Van Gerven 2000), and as such, it has been extensively discussed in the literature and in the case law of the European courts (Jans et al. 2007). (*Case C-410/92, Johnson* [1994] ECR I-5483, para. 21; *Case C-394/93, Unibet (London) Ltd and Unibet (International) Ltd*. [2007] ECR I-2271, para 39; *Joined Cases C-430/93 and C-431/93 Van Schijndel and van Veen* [1995] ECR-I4705; *Joined cases C-295/04 to C-298/04 Manfredi v Lloyd Adriatico Assicurazioni SpA and Others* [2006] ECR para 62.) This discussion has mainly addressed the competence allocation between the Member States and the EU and the increasing influence of EU law and policy with regard to procedural and remedial autonomy (Delicostopoulos 2003; Kakouris 1997;

Lenaerts et al. 2006; Prechal 1998; Trstenjak and Beysen 2011; Reich 2007; Van Gerven 2000). It has, however, not addressed the question to which authorities of the Member States allocate regulatory powers for the enforcement of EU law and how they organize and structure these enforcement agencies in their national administrative law system. (Institutional autonomy is the Member States' competence to design their own institutional structure and allocate regulatory powers to public administrative agencies that enforce EU law (Verhoeven 2010). In *International Fruit Company II*, the CJEU has stated that

[A]lthough under Article 5 of the Treaty the Member States are obliged to take all appropriate measures, whether general or particular, to ensure fulfilment of the obligations arising out of the Treaty, it is for them to determine which institutions within the national system shall be empowered to adopt the said measures [...] when provisions of the Treaty or of Regulations confer power or impose obligations upon the states for the purposes of the implementation of Community law, the question of how the exercise of such powers and the fulfilment of such obligations may be entrusted by Member States to specific national bodies is solely a matter for the constitutional system of each state

Joined cases 51–54/71 International Fruit Company II 15 December 1971: [1971] E.C.R. 1107. paras 3–4.)

While harmonization and convergence of substantive laws are well advanced, a similar convergence and harmonization of the procedural rules and institutional frameworks have not taken place. In the following sections, first the development of EU law concerning law enforcement will be examined, and then a critical look is taken at the EU's aim to harmonize the national procedural rules when EU law is enforced. First, it will be examined whether legally it is feasible, i.e., what legal basis can be used in order to further harmonize procedural rules. Second, it will be examined whether harmonizing procedural rules would be more efficient than the existing legal diversity. The economics of harmonization will be applied to assess the top-down harmonization by the EU and comparative law, and economics is applied to evaluate the bottom-up voluntary harmonization of the Member States.

Development of EU Law on Law Enforcement

In the EU Member States, highly convergent and harmonized substantive rules are implemented through diverging procedures and different kinds of enforcement bodies. This decentralized enforcement poses a challenge to the coherent and uniform application of EU law. In an enforcement system where Member States apply divergent procedures and may impose a variety of sanctions and remedies administered by various actors, the effectiveness of EU law, effective judicial protection (Articles 6 and 13 of the ECHR, Article 47 of the Charter of fundamental rights of the European Union which has now been reaffirmed in Article 19(1) TEU), and effective law administration may be at risk. In EU competition law, for example, it has been questioned whether consistent policy enforcement and the effective functioning of the European Competition Network require a certain degree of harmonization of procedures, resources, experiences, and independence of the NCAs (Bakardjieva-Engelbrekt 2009; Cengiz 2009; Gauer 2001; Frédéric 2001). Similarly, in consumer law, the Unfair Commercial Practices Directive 2005/29 has allowed the Member States to establish their own specific enforcement systems. The national enforcement regimes are very diverse: some Member States have predominantly private enforcement; others rely predominantly on public bodies. In accordance with the principles of procedural and institutional autonomy, the Member States can entrust public agencies or private organizations with the enforcement of consumer laws, enabling those institutions to decide on the internal organization, regulatory competences, and powers of public agencies (Balogh and Cseres 2013). The variety of national enforcement architectures is remarkable in light of the far-reaching harmonization goal of the Directive. Moreover, the broader institutional framework comprising of locally developed enforcement strategies may further differentiate the Member States' enforcement models.

The following two sections will first examine which developments have taken place toward

harmonization of law enforcement in the EU and then examine the economic rationale of such harmonization.

Europeanizing Law Enforcement

The influence of EU law on Member States' procedures and remedies of law enforcement has been gradually growing since 1992 (de Moor-van Vugt 2011) and has been intensified when enlargement to the Central and Eastern European countries in 2004 took place (Bakardjieva-Engelbrekt 2009). According to Nicolaides, enforcement became a priority area of EU policy with the process of enlargement due to the fact that, first, the CEECs emerged from many years of communism and they had to build institutions that were accountable to citizens and functioned in very different environments than in the past. Second, EU integration has progressed, and the impediments in the internal market were found in administrative weaknesses and incorrect implementation of EU law. Third, the legal body of the *acquis* expanded considerably, especially in the area of internal market, and it made proper enforcement key to make the single market work (Nicolaides 2003, pp. 47–48).

De Moor-van Vugt has identified patterns in the Europeanization of law enforcement. She shows that as from the early 1990s, the Commission began to monitor the Member States' enforcement of EU law due to insufficient implementation of EU law resulting in fraud with EU structural funds (de Moor-van Vugt 2011). As a result, the Commission implemented a stricter policy which limited the Member States' procedural autonomy and obliged them to comply with the principles of sincere cooperation, non-discrimination, and effectiveness (de Moor-van Vugt 2011). The CJEU's judgment in *Greek Maize* opened the way for the Commission to lay down obligations for the Member States in consequent directives and regulations that concerned subsidies in order to take appropriate measures in case of infringements of EU law. (Since the CJEU's judgment in C-68/88 *Greek Maize* the Member States are required by Article

4 (3) TFEU to take all measures necessary to guarantee the application and effectiveness of EU law. While the Member States remain free to choose the appropriate enforcement tools, they must ensure that infringements of EU law are penalized under conditions, both procedural and substantive ones, which are analogous to those applicable to infringements of national law of a similar nature and importance and which, in any event, make the penalty *effective, proportionate, and dissuasive*. Case 68/88 *Commission v Hellenic Republic*, Judgment of the Court of 21 September 1989, I-2965, paras 23–24.) De Moor-van Vugt demonstrates that the same pattern of Europeanizing national enforcement models has been followed by the EU in several other sectors such as agriculture, environmental law, financial services, and sector regulations such as telecom and energy (de Moor-van Vugt 2011, p. 72).

The process of Europeanizing enforcement was well visible in the modernization of EU competition law from the late 1990s and further since Regulation 1/2003 entered into force in 2004. The improvement of cross-border enforcement laid also at the heart of the Regulation 2006/2004 on consumer protection cooperation. (Regulation (EC) No 2006/2004 of the European Parliament and of the Council of 27 October 2004 on cooperation between national authorities responsible for the enforcement of consumer protection, O.J. L 364/1, 9.12.2004.) This Regulation has set up an EU-wide network of national enforcement authorities and enabled them to take coordinated action for the enforcement of the laws that protect consumers' interests and to ensure compliance with those laws.

Similarly, in the liberalized network industries, the EU gradually extended the EU principles of *effective, dissuasive, and proportionate sanctions* as formulated in the case law of the European courts to a broader set of obligations and criteria for national supervision in EU legislation. The liberalization of state-owned enterprises has been accompanied by the obligation for Member States to create regulatory agencies in order to maintain elements of public control and to provide reassurance of independence from government in

creating a level playing field for new entrants (Gorecki 2011; Scott 2000; Thatcher 2002).

Europeanizing market supervision (Ottow 2012) in the liberalized network industries also obliged Member States to establish independent national regulatory agencies with core responsibilities for monitoring markets and safeguarding consumers' interests (Micklitz 2009). Member States have strengthened the role of regulatory agencies and have empowered them with a growing number and diversity of regulatory competences. In law enforcement, a shift has taken place from the state to individuals and their collectives and also a shift from judicial enforcement to more administrative law enforcement (Cafaggi and Micklitz 2009).

Harmonization of Law Enforcement

The above-described process of Europeanizing law enforcement in the Member States through the harmonization of national procedural rules has been driven by the EU Commission. However, EU harmonization of civil or administrative procedures faces problems of legitimacy (e.g., private enforcement of competition law in fact is a question of national private law rules, contract, tort, and corresponding civil procedural rules. Case C-453/99 *Courage v. Crehan* ECR [2001] I-6297) because the Commission lacks the competence and a clear legal basis to harmonize procedural rules. In accordance with the so-called national procedural autonomy and the principle of subsidiarity, the Member States have the competence to lay down private law consequences of EU law infringements as well as the administrative procedures. The Member States provide for remedies to effectuate damage actions, and it is for the national courts to hear cases. (The CJEU has consistently held that

[I]n the absence of Community rules governing the matter, it is for the domestic legal system of each Member State to designate the courts and tribunals having jurisdiction and to lay down the detailed procedural rules governing actions for safeguarding rights which individuals derive directly from Community law, provided that such rules are not less favourable than those governing similar domestic

actions (principle of equivalence) and that they do not render practically impossible or excessively difficult the exercise of rights conferred by Community law (principle of effectiveness)

Joined cases C-295/04 to C-298/04 *Manfredi v Lloyd Adriatico Assicurazioni SpA and Others* [2006] ECR para 62; Case 33/76, *Rewe-Zentralfinanz eG and Rewe-Zentral AG v Landwirtschaftskammer für das Saarland (Rewe I)*, [1976] ECR 1989, para 5, Case C-261/95 *Palmisani* [1997] ECR I-4025, para 27, Case C-453/99 *Courage and Crehan*, par. 29.)

In accordance with Article 5 TEU, the Union is only empowered to act within the competences conferred upon it by the Treaty. With regard to the harmonization of procedural rules, one could turn to Article 114 TFEU, which forms the legal basis for harmonization measures when such measures have as their objective the establishment and the functioning of the internal market. For example, the Public Procurement Remedies Directives were issued on this legal basis. (Council Directive 89/665/EEC of 21 December 1989 on the coordination of the laws, regulations, and administrative provisions relating to the application of review procedures to the award of public supply and public works contracts [1989] OJ L 395/33; Council Directive 92/13/EEC of 25 February 1992 coordinating the laws, regulations, and administrative provisions relating to the application of Community rules on the procurement procedures of entities operating in the water, energy, transport, and telecommunications sectors [1992] OJ L 76/14.) However, this Article has been strictly interpreted by the EU courts, and it can be applied only when it can be proved that without the harmonization measures, the functioning of the internal market would be endangered and competition distorted. The CJEU, among others, said that the goal of the Commissions' intervention has to be precisely stated by explaining the actual problems consumers face in the internal market and the actual obstacles to the free movement principles as well as the distortions of competition. In *Germany v. Parliament and Council*, the ECJ has explicitly said that

a measure adopted on the basis of Article 100a of the Treaty must genuinely have as its object the

improvement of the conditions for the establishment and functioning of the internal market. If a mere finding of disparities between national rules and of the abstract risk of obstacles to the exercise of fundamental freedoms or of distortions of competition liable to result there from were sufficient to justify the choice of Article 100a as a legal basis, judicial review of compliance with the proper legal basis might be rendered nugatory. (Case C-376/98 *Germany v. Parliament and Council* [5 October 2000] ECR I-8419, para 84; Two years later in the 'Tobacco Labelling' judgment when applying the same arguments it approved the adoption of the Tobacco Labelling Directive on the basis of Article 95 EC and it thereby reaffirmed its interpretation of Article 95 as a legal basis for measures of harmonization. Case C-491/01 *British American Tobacco v The Queen* [2002] ECR I-; Directive 2001/37 on Tobacco Labelling OJ 2001 L194/26 paras 60–61)

Accordingly, the Commission has to precisely define the legal problems by providing clear evidence of their nature and magnitude, explaining why they have arisen and identifying the incentives of affected entities and their consequent behavior.

Since the Amsterdam Treaty Article 81 TFEU can be applied as the legal basis for harmonization of civil procedural law, this legal basis can be used with regard to civil matters which have cross-border implications and in so far as common rules are necessary for the functioning of the internal market. It could be argued that even though this legal basis concerns civil procedural measures, it could be applicable also to administrative procedural law (Eliantonio 2009, p. 4).

Both Article 114 and 81 TFEU require a justification for procedural harmonization measures by showing that the functioning of the internal market is at stake, namely, that the direct effect of substantive EU law might be at risk and market competition would not take place on equal terms, unless at least some minimal requirements concerning procedure were upheld in all Member States, then there would be adequate grounds to support the introduction of harmonized remedies in national courts. Accordingly, in order to decide upon the necessity of EU harmonization measures in the field of procedural law, the negative effects of diverging judicial remedies for European integration should be estimated.

Procedural differences in the Member States could be justified by their impact on business

actors. Competition would be distorted when business actors have to reduce their business in a certain Member State because of the difficulty that they might encounter in enforcing EU law. Consequently, the benefits of harmonized procedural rules may be transparency and legal certainty which might be appreciated especially from the perspective of economic policy and competition.

It is the next section that will discuss these economic arguments of harmonization and legal diversity.

The Economics of Harmonization

As mentioned above, in EU law, the harmonization process is governed by the principles of subsidiarity and proportionality. These principles have been argued to entail a cost-benefit analysis of legislation and require to minimize transaction costs (Van den Bergh 1994). The economics of harmonization discusses the costs and benefits of legal diversity and harmonization. It addresses the optimal level of intervention by applying the economic theory of federalism as extended to the theory of regulatory competition.

The idea that decentralized decision making may contribute to efficient policy choices in markets for legislation was first formulated by Tiebout in his seminal article on the optimal provision of local public goods (Tiebout 1956). (This economic theory argues that local authorities have an information advantage over central authorities, and, therefore, they are better placed to adjust the provision of public goods to the preferences of citizens. Under certain strict conditions, the diffusion of powers between local and central levels of government favor a bottom-up subsidiarity. The economics of federalism deals with the allocation of functions between different levels of government. Tiebout argued that buyers “vote with their feet” by choosing the jurisdiction which offers the best set of laws that satisfy their preferences. The economics of federalism rests upon a number of assumptions. When the “Tiebout conditions” are fulfilled, competition between legal orders will lead to efficient outcomes. There has to be a sufficiently large number of jurisdictions among

which consumers and firms can choose. Consumers and firms enjoy full mobility among jurisdictions at no costs. Last, there are no information asymmetries, which on the one hand means that states have full information as to the preferences of firms and citizens, and on the other, suppliers of production factors must have complete information on the costs and benefits of alternative legal arrangements. Only in the presence of these information requirements will consumers and firms be able to choose the set of laws, which maximizes their utility or profit. Further, no external effects should exist between states and regions. There must be no significant scale economies or transaction savings that require larger jurisdictions.) Tiebout’s model has been extended to legal rules and institutions. The theory of regulatory competition applies the dynamic view of competition to sellers of laws and choice between legal orders offering a number of criteria to judge whether centralization or decentralization is more successful in achieving the objectives of the proposed legislation (Van den Bergh 1994, 1996, 2002). In this section, these criteria will be applied to the Commission’s harmonization proposals in order to consider the likely costs and benefits of top-down rule-making.

One reason to harmonize procedural rules is that these rules differ across countries that they may lead to adverse externalities for other Member States. Such negative spillover effects might be present with regard to different procedures as well. While such negative externalities can be internalized by harmonization, bargaining between the Member States can also solve this problem. According to the Coase theorem, when property rights are well specified, transaction costs are low and information is complete and bargaining can be an efficient solution (Van den Bergh and Camesasca 2001, p. 132).

Another reason in favor of harmonization is that different legal rules carry the risk of destructive competition. Such a “race to the bottom” development has often been linked and criticized as a result of competition among jurisdictions. It has been argued that competition among legal rules drives social, environmental, cultural, and other standards down. This argument has been

mainly embraced in international corporate law by making reference to the “Delaware effect.” However, the risk of such declining levels of standards has not yet been proved, and the little empirical evidence is not conclusive enough (Wagner 2005; Josselin and Marciano 2004, pp. 477–520; McCahery and Vermeulen 2005). Furthermore, international trade may even stimulate race to the top (Van den Bergh and Camesasca 2001, pp. 153–154). Gomez also argues that the outcome of such competitive process cannot be examined without taking into account the relative power of the affected groups (Gomez 2008). Such competition might not harm powerful and well-organized groups but could have different effects for small- and medium-sized enterprises and consumers.

A third argument often raised to support harmonization is to achieve scale economies and to reduce transaction costs. Transaction costs can be high when firms and consumers have to search and comply with different sets of national rules. In the case of uniform rules, the search costs of information could be saved, and complying with one set of rules can achieve scale economies. Uniform rules can guarantee more stable and predictable jurisprudence and considerably contribute to transparency and legal certainty.

In particular, it has been argued that business would profit from clear and transparent system in which they would be able to enforce their claims against public authorities all over Europe pursuant to the same procedural rules (Elia Antonio 2009, p. 6). This raised the question whether procedural harmonization may be pursued in a “compartmentalized” way for specific policy areas. With regard to the private enforcement of EU competition law, two remarkable suggestions were made. Both concern a separate harmonization of economic torts or in the present case economic administrative procedures. Heinemann proposed that general tort rules of the DCFR could be examined against the backdrop of the special needs of competition law (Möllers and Heinemann 2007, p. 377). Boom proposed a compartmentalized approach to work with the existing modest body of European tort law. By addressing the policy issues involved in each of these torts one by one, the

European Union can make harmonized tort law more attainable. He pointed out that a likely candidate for harmonization is the category of economic torts, such as the protection of intellectual property through tort law, liability for infringement of competition rules, and the liability for misleading advertising (Van Boom 2008, pp. 133–149).

However, transaction costs can be especially relevant for large firms operating in interstate commerce, the same might not hold for small- and medium-sized undertakings operating mainly in national markets or for consumers. Therefore, as mentioned above, the impact of such harmonization also has to be analyzed with having regard to the relative power of the affected groups.

Furthermore, while uniform rules help to maintain economies of scale, which is an important argument for centralization, but they can only be advantageous from an *ex ante* perspective, when neither the Member States nor the Community have as yet adopted certain legislation (Van den Bergh 1998). This is neither the case with regard to administrative or civil procedural rules, which are rooted in old legal traditions and characteristics of the different legal systems.

When all parties in one region have identical preferences, cost efficiency considerations might point to harmonizing through one single instrument that suits all. This is clearly in-line with the preferences of the business community as they are in favor of uniform rules. However, the preferences of consumers and public administration can significantly diverge. In fact, it has been argued that the legal systems of the Member States are built through habits, customs, and practices which dictate how law is going to be interpreted (Legrand 2002, p. 230), and that public law “has particularly deep roots inside a cultural and political framework” (Harlow 2002, p. 208). This is clearly the case with regard to enforcement bodies as these are a wide variety of institutions that enforce EU law in the Member States (Cseres 2013; Balogh and Cseres 2013).

Accordingly, the possibility of achieving a common procedural (administrative or civil) law in Europe is doubtful because the political conditions are missing and because the national legal

systems are based on very different conceptions (Eliantonio 2009, p. 8). The same is true for bridging the gaps between the various economic policies and institutional settings. Harmonization of administrative procedures might conflict with legitimate national interests, such as the need to protect fairness and efficiency in the administration of justice (Eliantonio 2009, p. 7). Due to these fundamental differences in national administrative procedures, it would be very difficult to agree on common rules for all 27 jurisdictions, and in fact, it has been argued that “a general codification could be achieved only by reducing the requirements to the level of a common denominator, in which case it would prove as a barrier rather than an asset for an effective and uniform enforcement of Community law” (Schwarze 1996, p. 832). Moreover, some Member States may prefer to implement criminal law procedures for the enforcement of most severe law violations as this is the case already in a significant number of Member States concerning consumer, environmental, or even competition law.

In sum, there are not sufficient economic arguments in favor of harmonization, but there are good economic arguments in supporting legal diversity. One such argument is that a larger set of legislations can satisfy a wider range of preferences which leads to allocative efficiency. The broad range of preferences can easily be seen behind the various different regulations on unilateral conduct, but it also holds for administrative procedures. Another argument is information asymmetries that support decentralization by maintaining the principle of subsidiarity and procedural autonomy. When information at local level is more valuable for rule-making and law enforcement, decentralization is more efficient. Competition between these legal rules has the advantages of a learning process. National laboratories produce different rules that allow for different experiences and that can improve the understanding of alternative legal solutions. (Justice Brandeis’ famous metaphor for states as laboratories of law reform and his plea for decentralization has been laid down in his dissenting opinion in *New State Ice Corp. v. Liebmann*, 285 US 262, 311 (1932).) These advantages are

relevant for both the formulation of substantive rules as well as law enforcement. Moreover, legal diversity and competition does not necessarily exclude harmonization. In fact, dynamic competition between legal rules can lead to voluntary harmonization which in turn can be more effective and successful than forced coordination of legislations. Instead of forced harmonization, the Commission could guarantee the conditions for regulatory competition and let this process work up to voluntary harmonization. These conditions could in fact be ensured within the various European networks of national regulatory agencies that were set up in the last decade.

Voluntary Harmonization

This section will analyze the underlying rationales of voluntary convergence by making use of insights from comparative law and economics.

Comparative law and economics compares and evaluates the law of alternative legal systems with the “efficient” model offered by economic theory (Mattei et al. 2000, pp. 506–507) (Mattei 1994). Comparative law and economics deals with “legal transplants” by measuring them with the tool of efficiency, and it offers an economic analysis of institutional alternatives tested in legal history (Komesar 1994). It “deals with the transplants that have been made, why and how they were made, and the lessons to be learned from this.” While it offers comparative lawyers the measuring tools of economics, it, at the same time, places the notion of efficiency in a dynamic perspective by offering a comparative dimension with concrete alternative rules and institutions (Watson 1978).

The insights of comparative law and economics offer a dynamic approach to study legal divergence and convergence and compare that against the benchmark of efficiency offered by economics. In order to explain convergence between different legal systems that depart from different points, it uses economic efficiency to evaluate changes that are so-called legal transplants in a legal system. Convergence between different legal rules toward an efficient model may take

place as a result of a legal transplant or as an outcome of a competitive process between different legal formants (Mattei et al. 2000, pp. 508–511). In the first case, legal transplants are implemented because they proved to be efficient in other legal systems. In the second case, convergence toward efficiency is the result of the interaction between different legal formants. So while legal transplants are governed by hierarchy, the second scenario is governed by competition among legal formants (Mattei et al. 2000, pp. 510–511).

In EU competition law, for example, the Member States voluntarily harmonized various elements of their national administrative procedures (European Competition Network 2013). Yet, this convergence exhibits some shortcomings in terms of the benchmark it uses and in terms of the methods to achieve convergence. First, convergence between the different national rules uses Regulation 1/2003 and some accompanying soft-law instruments as its benchmark. Thus, convergence so far took place through legal transplants imposed by the Member States and by in fact implementing similar procedural rules as those of the Commission's. The underlying reasons might be that once these rules and enforcement methods work effectively and efficiently in the hands of the Commission, they will prove successful in the hands of the NCAs as well. However, the efficiency of these rules and their comparative advantage vis-à-vis other national rules has neither been analyzed nor confirmed. Furthermore, the success of such legal transplants is not guaranteed in the different institutional frameworks of the Member States, where agencies often have to divide resources between several legislative competences. The actual outcome of enforcement depends heavily on the existing institutional framework (Cseres 2014a, b).

The influence of the institutional framework also plays a role in measuring actual law enforcement and in understanding why a certain legal rule proves to be successful or fails in different institutional contexts (Stiglitz 2002; North 1995, p. 13).

Despite the blueprint convergence of procedural rules, the NCAs could not or did not actually

enforce these rules due to certain constraints present in their institutional framework. The strengthened enforcement tools have not always delivered the expected results in actual enforcement. This is, for example, the case with regard to the power to investigate private premises (Commission Staff Working Paper, par. 202.). Similar experience has been found with regard to leniency programs which are often praised as the model for procedural convergence in EU competition law and a clear result of the cooperation mechanism within the European Competition Network (Cseres 2014a, b).

Convergence of national laws in EU competition law is also steered from the center by the Commission establishing the EU rules as the benchmark for harmonization (Cseres 2014a, b). While the Commission was seemingly decentralizing enforcement powers, in fact it has retained a central policy-making role but without a control mechanism.

References

- Bakardjieva-Engelbrekt A (2009) Public and private enforcement of consumer law in central and Eastern Europe: institutional choice in the shadow of EU enlargement. In: Cafaggi F, Micklitz H-W (eds) *New frontiers of consumer protection. The interplay between private and public enforcement*. Intersentia, Antwerp, pp 47–92
- Balogh V, Cseres KJ (2013) Institutional design in Hungary: a case study of the unfair commercial practices directive. *J Consum Policy* 36:343–365
- Cafaggi F, Micklitz H-W (2009) Introduction. In: Cafaggi F, Micklitz H-W (eds) *New frontiers of consumer protection. The interplay between private and public enforcement*. Intersentia, Antwerp, pp 1–46
- Cengiz F (2009) Regulation 1/2003 revisited. TILEC discussion paper no. 2009-042, 17
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cseres KJ (2013) Integrate or separate. Institutional design for the enforcement of competition law and consumer law. ACELG working paper 2013-01
- Cseres KJ, Schrauwen A (2013) Empowering consumer-citizens: changing rights or merely discourse? In: Schiek D (ed) *The EU social and economic model after the global crisis: interdisciplinary perspectives*. Ashgate, Farnham, pp 117–139
- Cseres KJ (2014a) Accession to the EU's competition law regime: a law and governance approach. *Yearb Antitrust Regul Stud* 7(9):31–66

- Cseres KJ (2014b) The European competition network as experimentalist governance: the case of the CEECs, ECPR standing group on regulatory governance conference, Barcelona, 25–27 June 2014
- De Moor-van Vugt AJC (2011) Handhaving en toezicht in een Europese context. In: Pront-van Bommel S (ed) *De consument en de andere kant van de elektriciteitsmarkt: inleidingen op het openingscongres van het Centrum voor Energievraagstukken Universiteit van Amsterdam op 27 januari 2010*. Universiteit van Amsterdam, Centrum voor Energievraagstukken, Amsterdam, pp 62–95
- Delicostopoulos J (2003) Towards European procedural primacy in national legal systems. *Eur Law J* 9(5):599–613
- Elia Antonio M (2009) The future of national procedural law in Europe: harmonisation vs. judge-made standards in the field of administrative justice. *Electron J Comp Law* 13(3):1–11
- Frédéric J (2001) Does the effectiveness of the EU network of competition authorities depend on a certain degree of homogeneity within its membership? In: Ehlermann CD, Atanasiu I (eds) *European competition law annual 2000: the modernisation of EC antitrust policy*. Hart Publishing, Oxford, pp 208–210
- Gauer C (2001) Does the effectiveness of the EU network of competition authorities require a certain degree of harmonisation of national procedures and sanctions? In: Ehlermann CD, Atanasiu I (eds) *European competition law annual 2000: the modernisation of EC antitrust policy*. Hart Publishing, Oxford, pp 187–201
- Gomez F (2008) The harmonization of contract law through European rules: a law and economics perspective. *Eur Rev Contract Law* 4(2):89–118
- Gorecki P (2011) Economic regulation: recentralisation of power or improved quality of regulation? *Econ Soc Rev* 42:177–211
- Harlow C (2002) Voices of difference in a plural community. In: Beaumont P, Lyons C, Walker N (eds) *Convergence and divergence in European public law*. Hart Publishing, London, pp 199–204
- Jans JH, de Lange R, Prechal S, Widdershoven R (2007) *Europeanisation of public law*. Europa Law Publishing, Groningen
- Josselin J-M, Marciano A (2004) Federalism and subsidiarity in national and international contexts. In: Backhaus JG, Wagner RE (eds) *Handbook of public finance*. Springer Science & Business Media, United Kingdom, pp 477–520
- Kakouris CN (1997) Do the member states possess judicial procedural “autonomy”? *Common Mark Law Rev* 34:1389–1412
- Komesar NK (1994) *Imperfect alternatives: choosing institutions in law, economics and public policy*. University of Chicago Press, Chicago
- Legrand P (2002) Public law, Europeanisation and convergence: can comparatists contribute? In: Beaumont P, Lyons C, Walker N (eds) *Convergence and divergence in European public law*. Hart Publishing, London, pp 225–256
- Lenaerts K, Arts D, Maselis I (2006) *Procedural law of the European union*. Sweet & Maxwell, UK, pp 1–002
- Mattei U (1994) Efficiency in legal transplants: an essay in comparative law and economics. *Int Rev Law Econ* 14:3–19
- Mattei U, Antonioli L, Rossato A (2000) *Comparative law and economics*. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics*. Edward Elgar, Cheltenham
- McCahery JA, Vermeulen EPM (2005) Does the European company prevent the ‘delaware-effect’? TILEC discussion paper no. 2005-010. Available at SSRN: <http://ssrn.com/abstract=693421>
- Micklitz H (2009) *Universal services: nucleus for a social European private law*. EUI working paper law no. 2009/12
- Möller TMJ, Heinemann A (eds) (2007) *The enforcement of competition law in Europe*. Cambridge University Press, Cambridge
- Nicolaides PH (2003) Preparing or accession to the EU: how to establish capacity for effective and credible application of EU rules. In: Cremona M (ed) *The enlargement of the European union*. OUP, Oxford University Press, Oxford, pp 43–78
- North DC (1995) The new institutional economics and third world development. In: Harriss J, Hunter J, Lewis CM (eds) *The new institutional economics and third world development*. Routledge, London and New York, pp 17–26
- Ottow A (2012) Europeanization of the supervision of competitive markets. *Eur Public Law* 18(1):191–221
- Prechal S (1998) Community law in national courts: the lessons from Van Schijndel. *Common Mark Law Rev* 35:686
- Reich N (2007) Horizontal liability in EC law: hybridization of remedies for compensation in case of breaches of EC rights. *Common Mark Law Rev* 44:708
- Schwarze J (1996) The Europeanization of national administrative law. In: Schwarze J (ed) *Administrative law under European influence: on the convergence of the administrative laws of the EU member states*. Sweet & Maxwell, London, p 832
- Scott C (2000) Accountability in the regulatory state. *J Law Soc* 27(1):38–60
- Stiglitz J (2002) Participation and development: perspectives from the comprehensive development paradigm. *Rev Dev Econ* 6(2):163–182, 164
- Thatcher M (2002) Delegation to independent regulatory agencies: pressures, functions and contextual mediation. *West Eur Polit* 25:125–147
- Tiebout CM (1956) A pure theory of local expenditures. *J Polit Econ* 64:416–424
- Trstenjak V, Beysen E (2011) European consumer protection law: *curia semper dabit remedium*? *Common Mark Law Rev* 48:95–124
- Van Boom WH (2008) European tort law an integrated or compartmentalized approach? In: Vaquer A - (ed) *European private law beyond the common frame of reference – essays in honour of Reinhard Zimmermann*. Europa Law Publishing, Groningen, pp 133–149

- Van den Bergh R (1994) The subsidiarity principle in European Community law: some insights from law and economics. *Maastricht J Eur Comp Law* 1:337–366
- Van den Bergh R (1996) Modern Industrial Organisation versus old-fashioned European competition law. *Eur Compet Law Rev* 7–19
- Van den Bergh RJ (1998) Subsidiarity as an economic demarcation principle and the emergence of European private law. *Maastricht J Eur Comp Law* 5(2): 129–152
- Van den Bergh R (2002) Regulatory competition or harmonization of laws? Guidelines for the European regulator in the economics of harmonizing European law. In: Marciano A, Josselin J-M (eds) *The economics of harmonizing European law*. Edward Elgar, Cheltenham, pp 27–49
- Van den Bergh RJ, Camesasca PD (2001) European competition law and economics: a comparative perspective. Intersentia, Antwerpen
- Van Gerven W (2000) Of rights, remedies and procedures. *Common Mark Law Rev* 37:502
- Verhoeven M (2010) The ‘*Constanzo* Obligation’ and the principle of National Institutional Autonomy: supervision as a bridge to close the gap? *Rev Eur Admin Law* 3/1:23–64
- Wagner G (2005) The virtues of diversity in European private law. In: Smits J (ed) *The need for a European contract law; empirical and legal perspectives*. Europa Law Publishing, Groningen
- Watson A (1978) Comparative law and legal change. *Camb Law J* 37:313–336

Hate Groups and Hate Crime

Matt E. Ryan¹ and Peter T. Leeson²

¹Department of Economics, Duquesne University, Pittsburgh, PA, USA

²Department of Economics, George Mason University, Fairfax, VA, USA

Definition

Hate crimes are criminal offenses motivated by offenders’ animus toward their victims based on victims’ race, religion, sexual orientation, or other such characteristic. Hate groups are organizations of individuals whose organizing purpose is thought to reflect animus toward certain people based on their race, religion, sexual orientation, or other such characteristic.

The Economics of (Hate) Crime

Gary Becker (1968) pioneered the economics of crime by applying rational choice theory to analyze decisions to break the law. In his rendering, and in most of the enormous literature in the economics of crime following Becker, criminal activity is motivated by the prospect of material gain.

Although material gain undoubtedly motivates the vast majority of criminal behavior, it does not motivate all of it. One example of criminal behavior not so motivated is that which reflects “hate crime.” The Federal Bureau of Investigation defines such crime as “criminal offense against a person or property motivated in whole or in part by an offender’s bias against race, religion, disability, sexual orientation, ethnicity, gender, or gender identity.” A Neo-Nazi who vandalizes a Jewish synagogue because of his animus toward Jews, for example, commits a hate crime. As Gale et al. (2002) point out, hate crimes thus differ from most other crimes in that the perpetrator’s goal is not (at least in whole) to make himself better off, but rather to make his victim worse off.

Correlates of Hate Crime

The correlates of hate crime include at least two factors commonly found to be important determinants of crime more generally: unemployment and income. Medoff (1999), Gale et al. (2002), Ryan and Leeson (2011), and Curthoys (2013), for example, find a positive relationship between unemployment and the incidence of hate crime in the United States (though Green et al. 1998 and Mulholland 2013 find no consistently significant relationship). Similarly, Sharma (2015) finds that underemployment is a significant determinant of cross-caste crime in India, and Iparraguirre (2014) finds that “employment deprivation” is associated with more hate crime against older people in England and Wales. Medoff (1999) finds a negative relationship between wages and hate-crime incidence. And Gale et al. (2002) and Sharma (2015) find a positive relationship between the ratio of income of a targeted group to the

(presumed) targeting group and race-based and cross-caste crime, respectively.

A correlate of hate crime that may be more specific to this particular type of offense is the social perceptions of persons who are typically victims of hate crime. Hanes and Machin (2014), for example, find that the terrorist attacks of September 11, 2001 in the United States and on July 7, 2005 in London were associated with increased hate crimes against the Asian and Arab populations in England.

Do Hate Groups Matter?

The Southern Poverty Law Center (SPLC) defines hate groups as groups which “have beliefs or practices that attack or malign an entire class of people, typically for their immutable characteristics.” Among such groups, the SPLC counts the Ku Klux Klan, Neo-Nazi, White Nationalist, Racist Skinhead, Christian Identity, Neo-Confederate, Black Separatist, and “General Hate” groups. To a certain extent, a particular group’s status as a “hate group” may be a matter of debate. The SPLC, for example, considers “patriot groups” hate groups; however, it is questionable whether all such groups can be reasonably likened to the Ku Klux Klan, Neo-Nazis, or other such groups whose organizing purpose is clearly grounded in racial, religious, or some other such prejudice.

Since hate groups are those that malign (or are thought to malign) people based on their race, religion, sexual orientation, etc., an obvious question is whether the prevalence of such groups is an important determinant of criminal offenses committed against such people – hate crimes. Theoretically, the answer to this question is ambiguous. If an important activity of hate groups is the perpetration of crimes against hated classes, one would expect more hate groups to lead to more hate crime. Alternatively, hate groups may act as substitutes for, rather than complements to, hate crime. For example, if hate groups serve predominantly as outlets for individuals to congregate and “safely” vent their hatred of certain persons, which might otherwise manifest in the form of criminal acts against these persons, such

groups may displace hate crime with the hateful bluster. In this case, it is possible that more hate groups could lead to less hate crime.

Empirical examination of the relationship between hate groups and hate crime in the United States has delivered somewhat mixed results. Ryan and Leeson (2011), who consider the relationship between the number of hate groups and hate crime at the state level between 2002 and 2008, find no evidence that more hate groups lead to more hate crime. Mulholland (2013), who uses county level data between 1997 and 2007, finds that an active white supremacist group within a county is associated with an increased rate of hate crime in that county. Adamczyk et al. (2014), who also use county level data from 1990 to 2012, find that ideologically motivated homicides are more likely in counties with larger numbers of hate groups. Most recently, Larson and Brandt (2016) consider the number of Ku Klux Klan chapters in 72 combined statistical areas (CSAs) between 2006 and 2014 and find that while CSAs with a larger number of active Ku Klux Klan chapters experienced higher rates of racially motivated crime, they also experienced lower rates of religiously motivated crime.

Cross-References

- ▶ [Becker, Gary S.](#)
- ▶ [Crime: Expressive Crime and the Law](#)
- ▶ [Crime and Punishment \(Becker 1968\)](#)
- ▶ [Criminal Sanctions and Deterrence](#)
- ▶ [Economic Analysis of Law](#)

References

- Adamczyk A, Gruenewald J, Chermak S, Freilich J (2014) The relationship between hate groups and far-right ideological violence. *J Contemp Crim Justice* 30(3):310–332
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Curthoys A (2013) Identifying the effect of unemployment on hate crime. Mimeo, Syracuse
- Gale L, Heath W, Ressler R (2002) An economic analysis of hate crime. *East Econ J* 28(2):203–216

- Green D, Glaser J, Rich A (1998) From lynching to gay bashing: the elusive connection between economic conditions and hate crime. *J Pers Soc Psychol* 75:82–92
- Hanes E, Machin S (2014) Hate crime in the wake of terror attacks: evidence from 7/7 and 9/11. *J Contemp Crim Justice* 30(3):247–267
- Iparraguirre J (2014) Hate crime against older people in England and Wales – an econometric enquiry. *J Adult Protect* 16(3):152–165
- Larson A, Brandt P (2016) Hotbeds of hate: analyzing spatial and temporal disparities in hate crime across U.S. cities. Mimeo, Dallas
- Medoff H (1999) Allocation of time and hateful behavior: a theoretical and positive analysis of hate and hate crimes. *Am J Econ Sociol* 58(4):959–973
- Mulholland S (2013) White supremacist groups and hate crime. *Public Choice* 157(1-2):91–113
- Ryan M, Leeson PT (2011) Hate groups and hate crime. *Int Rev Law Econ* 31(4):256–262
- Sharma S (2015) Caste-based crimes and economic status: evidence from India. *J Comp Econ* 43(1):204–226

Hayek, Friedrich August von

Edward Peter Stringham^{1,2} and Todd J. Zywicki³

¹Economic Organizations and Innovation, Trinity College, Hartford, CT, USA

²American Institute for Economic Research, Great Barrington, MA, USA

³Antonin Scalia Law School, George Mason University, Arlington, VA, USA

Abstract

F.A. Hayek focused on many traditional economic questions, and also made important contributions to law and economics. His framework differed from Kaldor Hicks efficiency and the wealth maximization norm common among neoclassical law and economics scholars. But he talked about how the common law evolves and helps shape economic outcomes. Underlying this approach was Hayek's conviction that the essence of law is not created by the state, but rather pre-exists in the conventions and understandings within a community. Hayek argued that the role of the judge in a common-law system is to discover the law in the imminent consensus of norms and expectations. Hayek's work has

many implications for positive analysis and normative discussions of what judges should or should not do. To Hayek, the primary purpose of the law is not a wealth maximization problem, but to provide a stable institutional framework in a dynamic world that enables individuals to plan and coordinate.

Biography

Friedrich A. Hayek was born in Vienna in 1899 and won the economics Nobel Prize in 1974 for “pioneering work in the theory of money and economic fluctuations” and “penetrating analysis of the interdependence of economic, social and institutional phenomena.” After studying under Ludwig von Mises in the 1920s, Hayek had a series of debates with John Maynard Keynes about government's ability to manage the business cycle. In the 1930s and 1940s, Hayek joined Mises in a debate against those who argued that government should centrally plan the economy. He maintained that economic problems cannot simply be solved with a system of simultaneous equations. In “The Use of Knowledge in Society,” Hayek (1945) interpreted the role of prices in a market society as the most effective means of allowing people to use their knowledge of time and place to develop economic plans consistent with those of the rest of society. And Hayek described the market itself as a “discovery process” for generating that very knowledge. His best-selling work was *The Road To Serfdom*, where Hayek (1944) argued that even setting aside the fatal economic flaws, socialism and milder forms of government planning are incompatible with freedom in the long run. Much of this work amounts to a vindication of the notion of spontaneous order – patterns that emerge through human action but not human design. Hayek came to see that the notion applied just as well to the formation of law, thereby making an early contribution to the field of land economics. (This entry focuses on Hayek's contribution to law and economics. For a more comprehensive discussion of Hayek's thought, see Caldwell (2004).)

Innovative and Original Aspects

In the earlier part of his career, Hayek followed the Continental *Rechtsstaat* tradition – going back to Napoleon – where law is understood as statute-based, formally expressed rules. Hayek believed these rules should be forward-looking, well-articulated, and applied equally to all individuals. Such thought can be found in his *The Road to Serfdom* (1944) and *The Constitution of Liberty* (1960).

But Hayek came to shift his views radically, as can be seen in his three-volume series *Law, Legislation, and Liberty* (1973; 1976; 1979). Hayek underplayed the significance of the shift, viewing it as a move from a focus on “public” law to “private” law. Empirically, Hayek had now turned his attention to the common law of England from the Continental law. Analytically, and viewed from the outside, Hayek began breaking from the prevailing scholarly understanding of the ideal of law – and of law itself.

According to the later works of Hayek, not only was the common law a spontaneous order, the common law was analogous to the market process. Properly instituted, common law evolved according to the population’s changing expectations concerning disputes. In contrast, top-down civil law evolved based on legislators’ understanding of what rules best resolve disputes. Legislators attempt to institute detailed rules to anticipate all future legal circumstances. In common law, judges instead act on the facts of particular cases, but the unintended result of all such decisions is a body of law that not only coordinates most people’s expectations, but makes use of detailed, tacit, and local knowledge and promotes liberty. Another effect of the decentralized legal process is that abstract, coherent principles emerge. “In an ever changing society,” Hayek wrote, “judges must seek to find rules that will aim at securing certain abstract characteristics of the overall order of our society that we would like it to possess to a higher degree” (Hayek 1973, p.105).

The common law is not only abstract, but non-arbitrary, despite emerging without central design, as it aggregates knowledge implicit in the pre-

existing, commonly understood conventions of people in particular communities (Hume 1740, Book III, p.541; Hayek 1973, p. 162). To be sure, certain conventions needed for a society to thrive can be objectively identified: stability of ownership (property law), transference by consent (contract law), and protection of property (tort law). Such a perspective can be consistent with a natural law- or rights-based perspective where certain basic rights or side constraints, to use Robert Nozick’s language, exist. But within that basic framework, many of the specifics of legal procedures and law itself can be seen as evolving. Under the right conditions, a Darwinian process of competition among communities with differing bodies of law tends to lead law to converge away from systems that fail to protect investment and resource conflict toward systems that lead to peaceful cooperation (Zywicki 2000).

Hayek found further merit in common-law systems for making law predictable. Counter-intuitively, civil law, despite its greater verbal precision, provides less predictability. Common law works in this respect because people in any given community have an intuitive understanding of the rules reflecting their widely shared sense of what is just. Having this understanding, potential litigants can predict on their own how cases would be decided and can adjust their planning accordingly. They may not be able to articulate the rules (their knowledge of the law is tacit; Zywicki 1998), but regardless, their intuitive understanding means they often need not consult case law. In contrast, potential litigants in a civil-law system must often seek legal advice to learn the specifics of complicated rules.

Given that the rules constituting common law are difficult for litigants to articulate, it should be no surprise that the role of the judge is to discover conventions, not construct them. Such a judge considers the implicit consensus of norms and expectations in a given community and applies it to the facts in particular cases (Zywicki and Sanders 2008). A legislator, in contrast, considers what external standard should govern disputes. The body of judge-discovered law emerges spontaneously (Hayek 1973, p.118); the civil law arises by design. Hayek (1973, p.123) wrote,

“The difference between the rules of just conduct which emerge from the judicial process, the *nomos* or law of liberty—and the rules laid down by authority ... lies in the fact that the former are derived from the conditions of a spontaneous order which man has not made, while the latter serve the deliberate building of an organization serving specific purposes.” Put differently, common law “has never been ‘invented’ nor can it be ‘promulgated’ or ‘announced’ before hand” (1973, pp. 72, 118).

Hayek considers the common law’s discovery process as analogous to that of price discovery in markets (Zywicki and Sanders, 2008). The owner of a gas station, for example, must discover and articulate what price the consumer will pay for gasoline, based on the supply and demand resulting from the outcome of billions of individuals’ interactions. Similarly, a judge discovers and articulates the preexisting rules. Both the gas station owner and the judge engage in a process of trial and error.

Discovery of the law involves a search for what conventions govern the particular circumstances of time and place. The specific conventions evolve over time and differ across place because individuals’ needs for exploiting their local knowledge differ. But though the conventions differ, they reflect the underlying, universal, and nonarbitrary principles for resolving disputes and promoting coordination. As society becomes more complex, larger, and more heterogeneous, the need arises for someone to articulate the conventions. But the law itself is the consensus of principles and expectations, not the judges’ articulation of the consensus.

Beyond the advantage of allowing judges to discover legal rules, the decentralized process of law making ensures there is no central decision-maker for interested parties to capture who can then impose a rule that will generate long-term rents, thereby reducing the opportunity and incentive for rent-seeking litigation.

Despite all of the perceived advantages of the common law, Hayek ultimately saw room for legislatures to step in when common law reaches a “dead end.” The judge’s proper role is at most to articulate new rules within an existing framework of rules and expectations – not to alter expectations and amend the framework. To Hayek, the legislature’s proper role is to devise new rules

when existing rules have become obsolete or if rules inconsistent with a market economy emerge. In his ideal, the legislature should act consistently with the purpose-independent, abstract principles of the common law (*cosmos*) rather than transform it through command-oriented rules (*nomos*). So even though Hayek supported judges setting most precedents, he supported legislatures overseeing the process to remove bad precedents.

One could argue that a tension exists in Hayek’s thinking here. On one hand, Hayek saw the evolution of law as a discovery process akin to the market’s discovery process. But at the same time, he analogized the monopolistic court system to the market system. But given that markets are disciplined by profits and losses, can a monopolistic court system have similar discipline? Can such a system learn whether the rules are consistent with underlying principles and expectations without the aid of competition? Does his other analysis about markets as a discovery process logically imply that an effective legal system must require competition as well (Stringham and Zywicki 2011a, b)?

Impact and Legacy

Hayek’s work has many implications for positive analysis and normative discussions of what judges, or other government officials such as regulators, should or should not do. In a Posnerian normative framework, judges should evaluate the costs and benefits of different decisions or laws in general and implement those that maximize wealth. But a Hayekian framework stresses the knowledge problems that confront judges seeking to determine and implement the wealth maximizing decision or set of rules. A Hayekian framework also emphasizes the subjective nature of individual cost and choice and the challenges this provides for any judge seeking to ascertain the efficient rule in any scenario. In this sense, judges seeking to maximize societal wealth or Kaldor Hicks efficiency are in a similar position to any central planner seeking to allocate resources to optimize wealth. They lack the necessary information to weigh costs and benefits and choose what is best for society. Such logic also applies

to judges who wish to tweak the details of the law. For instance, a movement from contributory to comparative negligence may have implications not only for other elements of the tort system (such as liability rules or damages), but also contract law, procedure, and remedies.

The more neoclassical law and economics models also implicitly assume that the world is in equilibrium where the judge simply compares the end states of his possible decisions. A Hayekian influenced law and economics, in contrast, recognizes that the world is in a state of change as billions of consumers around the world seek to constantly adjust to millions of constant and simultaneous shocks. Such change affects basic economic conditions but can also disrupt contractual relationships and lead to conflict among individuals (Zywicki and Sanders 2008). In Hayek's perspective, equilibrium cannot describe the world in the abstract. Instead we should focus on the ability of individuals to mesh their plans at any given time and to form expectations about how parties will perform in the future.

In this view, the primary purpose of the law is not a wealth maximization problem, but instead to provide a stable institutional framework that enable individuals to plan and coordinate their affairs in a world of constant dynamism. A set of relatively stable and clear rules enable people to predict one another's behavior (Hayek, 1978). Relatively clear and stable rules create boundaries for property rights, and other legal obligations enable individuals to adapt their behavior to the ever-changing world that surrounds them. Adding a constantly changing legal system – even if in the name of “modernization” or “updating” – to this chaotic world could create uncertainty and undermine the ability of individuals to coordinate their plans in the face of constant need for adaptation.

Zywicki (1998) argues that getting rid of letting government constantly change or adjust the law actually allows ordinary people to have a greater degree of confidence in the content of the law without needing to consult lawyers or run across an unexpected legal obligation. It also enables radical decentralization of decision-making authority to individuals who have the tacit and local knowledge most relevant to the particular decision. Some law and economics

scholars influenced by Hayek focus on legal norms or even entire legal systems that clearly emerged without central design or even any government control. Bernstein (1992) and Ellickson (1991) show that New York diamond merchants and California farmers and cattle ranchers devise various rules and regulations to govern their behavior independent of what legislatures or government judges say. Benson (1990) argues that the entire law merchant was created by business people to meet their needs of exchange. Stringham (2015) highlights that in seventeenth century Amsterdam, eighteenth century London, and nineteenth century New York, complex financial contracts emerged even though government was not enforcing, or actively trying to prohibit, them. The rules that govern the most advanced markets in the world market emerged from the market. In a Hayekian law and economics perspective, the world is not an optimization problem that government planners, regulators, or judges evaluate and solve. Instead the world, including law itself, can be viewed as a spontaneous order.

Cross-References

- ▶ [Austrian Perspectives in Law and Economics](#)
- ▶ [Constructivism, Cultural Evolution, and Spontaneous Order](#)
- ▶ [Harmonization of Tort Law in Europe](#)
- ▶ [Law and Economics](#)
- ▶ [Law and Economics, History of](#)
- ▶ [Rule of Law](#)

References

- Bernstein L (1992) Opting out of the legal system: extra-legal contractual relations in the diamond industry. *J Leg Stud* 21:115–57
- Benson BL (1990) *The Enterprise of law: justice without the state*. Pacific Research Institute for Public Policy, San Francisco
- Caldwell B (2004) *Hayek's challenge: an intellectual biography of F. A. Hayek*. University of Chicago Press, Chicago
- Ellickson RC (1991) *Order without law: how neighbors settle disputes*. Harvard University Press, Cambridge
- Hayek FA (1944) *The Road to Serfdom*. The University of Chicago Press, Chicago
- Hayek FA (1945) The use of knowledge in society. *Am Econ Rev* 35(4):519–530

- Hayek FA (1960) *The constitution of liberty*. University of Chicago Press, Chicago
- Hayek FA (1968/2002) *Competition as a discovery procedure*. (Marcellus S. Snow Trans.) *Q J Austrian Econ* 5:9–23
- Hayek FA (1973) *Law, legislation and liberty*, volume 1: Rules and order. University of Chicago Press, Chicago
- Hayek FA (1976a/1990) *The denationalisation of money*, 3rd edn. Institute of Economic Affairs, London
- Hayek FA (1976b) *Law, legislation and liberty*, volume 2: The mirage of social justice. University of Chicago Press, Chicago
- Hayek FA (1978) Tom Hazlett interviews Friedrich A. Hayek, November 12, 1978. UCLA Oral History Program and the Pacific Academy of Advanced Studies. www.hayek.ufm.edu. Accessed 12 Dec 2010
- Hayek FA (1979) *Law, legislation and liberty*, volume 3: the political order of a free people. University of Chicago Press, Chicago
- Hume D (1740) *A treatise of human nature*. Book III. John Noon, London
- Stringham EP (2015) *Private governance: creating order in economic and social life*. Oxford University, New York/Oxford
- Stringham EP, Zywicki TJ (2011a) Hayekian Anarchism. *J Econ Behav Organ* 78(3):290–301
- Stringham EP, Zywicki TJ (2011b) Rivalry and superior dispatch: an analysis of competing courts in medieval and early modern England. *Public Choice* 147:497–524
- Zywicki TJ (1998) Epstein and Polanyi on simple rules, complex systems, and decentralization. *Constit Polit Econ* 9:143–150
- Zywicki TJ (2000) Was Hayek right about group selection after all? Review essay of unto others: the evolution and psychology of unselfish behavior by Elliott sober and David Sloan Wilson. *Rev Austrian Econ* 13(1):81–95
- Zywicki TJ, Sanders AB (2008) Posner, Hayek, and the economic analysis of law. *Iowa Law Rev* 93:559–604

Hedge Fund Activism in Corporate Governance

Alessio M. Paces

Erasmus School of Law, Rotterdam Institute of Law and Economics, Erasmus University
Rotterdam, Rotterdam, The Netherlands

Definition

Hedge fund activism is an important disciplinary mechanism in corporate governance. It consists in actions aimed to change the way a company is managed, without trying to gain control. Because

hedge funds aim to profit from these actions, this is called entrepreneurial shareholder activism.

A hedge fund campaign can only succeed if a majority of institutional shareholders supports it. The question whether hedge fund activism leads to inefficient short termism in corporate governance can only be answered by looking at which investors are decisive on such campaigns. Although the empirical evidence on this is still unclear, index funds seem to be decisive in most instances.

Because of their incentives, index fund managers cannot be trusted to make an informed decision about whether the company should be managed for the short term or the long term. It turns out that the optimal choice in this respect varies with the individual companies and with time. As a result, companies should be enabled to opt out of hedge fund activism by way of dual-class or at least loyalty shares, if a majority of institutional investors agrees to this solution ex-ante.

Introduction

Some ten years ago, a prominent law and economics commentator defined hedge funds (and private equity funds) “the newest big thing in corporate governance” (Macey 2008: 241). Activist hedge funds are definitely the big thing in corporate governance today. They intervene in corporate governance with the goal to make changes happen. So do private equity funds, but there are two differences. For one, hedge funds are supposed to unlock existing value, whereas private equity purports to create new value. Secondly, hedge funds confront the management of public companies, whereas private equity makes friendly deal with them to take the company private. This entry focuses on hedge fund activism as a disciplinary mechanism in the governance of public companies.

Although hedge fund activism has emerged as an US phenomenon, it has now become international (Becht et al. 2016). Hedge fund activism is increasingly prominent in Europe, for instance, in the UK. Moreover, hedge funds are active in

concentrated ownership structures, too. Somewhat surprisingly, activist hedge funds have succeeded in extracting concessions also in the presence of dominant shareholders. Importantly, hedge funds do not always achieve outcomes from engaging the management and do not need to engage overtly in order to achieve outcomes.

Activist hedge funds screen the market for underperforming companies and buy a significant stake in them in order to engage their management. The goal of activists is to make a profit from improving the company's performance, which will be reflected by an increase in the value of their stake when it is sold back to the market. In doing so, hedge funds are coping with the fundamental agency problem stemming from separation of ownership and control, namely, the failure by management to maximize shareholder value. However, because hedge funds profit from a short-term change in the stock price of the target company, they may also be responsible for short termism in corporate governance. Feeling the pressure of hedge funds, managers may pursue short-term stock returns at the expenses of long-term value creation. For this reason, hedge fund activism is highly controversial.

From a law and economics standpoint, the question is whether hedge fund activism remedies or exacerbates market failure. On the one hand, the reduction of agency costs stemming from activism improves the efficiency of corporate governance. On the other hand, if the profitability of hedge funds activism depends on short-term pricing, the ability of corporate governance to sustain certain long-term projects may be undermined. To answer these questions, it is crucial to understand what drives hedge fund activism as well as the factors affecting its success.

Hedge Fund Business Model

Shareholder activism is not new. Activists, such as individuals or public pension funds, have always been prompting managers to act on some issue, sometimes for ideological reasons. Before the advent of hedge funds, however, such activism has not been very effective in achieving outcomes

of sort, let alone in affecting stock prices (Bratton and Mc Cahery 2015).

Hedge fund activism is different. Like other activists, hedge funds take actions aimed to change the way a public company is managed, without trying to gain control (Gillan and Starks 2007). However, differently from traditional activism, hedge funds strive to profit from the changes they provoke. This is called "entrepreneurial activism" (Klein and Zur 2009). The changes can be quite radical, such as the departure of the CEO or some other executives, if not the restructuring of the company. Likewise, activist hedge funds may seek to stop a change wanted by the management, for instance an acquisition.

The mark of entrepreneurial activism's success is whether the desired change happens or not. Hedge funds have a different business model than other institutional investors. Hedge fund managers charge a performance fee in addition to a percentage of the asset under management. The remuneration usually follows the so-called 2-20 rule: 2% of the asset under management plus 20% of any increase in the value of the portfolio. This remuneration structure aligns the hedge fund incentives with investors having a relatively high appetite for risk. Hedge funds profit from investing in stock that they can buy, hold, and resell at a higher price.

Two factors are key for the success of entrepreneurial activism. First, the hedge fund needs to be able to buy the bulk of its stake in the company, while the stock market does not anticipate the engagement. The moment the engagement is revealed, investors will anticipate gains, and discounting those for the probability that the engagement fails, the stock price will rise disabling further profit from activism. Second, the activist needs to be able to persuade the management to implement the desired changes. To increase its leverage with the management, the activist can use several techniques, ranging from news campaigns to threatening a lawsuit. The last resort, however, is a shareholder vote. Reached that point, the success of the engagement will depend on whether the activist has managed to attract sufficient support from other shareholders to get a favorable vote. This explains the

importance of proxy contests for activism in the USA and of the rules for initiating and executing a shareholder vote in the European jurisdictions.

The support by institutional investors is crucial for successful engagement. The typical hedge fund stake in the target company is slightly above 6%, which is not nearly a controlling one. As a result, activists must persuade other investors to vote for them. Similarly, engagement may succeed based on the sheer threat of winning a contested vote. From the moment the hedge funds formulate their demands to the management, both parties start to speak with the largest institutional investors, while the investing public is in the dark about the engagement. Management will give in to the activists' demands when it is clear they are going to lose the vote, whereas hedge funds will withdraw from engagement when they realize that not enough institutional investors will vote for them. The fight becomes public only when the outcome is ambiguous. Consequently, a substantial portion of hedge fund engagement takes place behind closed door. This is consistent with the theoretical literature on litigation vs. settlement, as noted by Bebchuk et al. (2017a).

As explained by Gilson and Gordon (2013), the tremendous influence activists have gained in corporate governance depends on the reconcentration of ownership occurred in the past few decades. The bulk of equity investment is no longer in the hands of dispersed individual stockholders, but is managed by institutional investors. In 2016, institutional investors collectively held 63% of US public equities. Institutional ownership is also concentrated. The largest five and 20 institutional investors control, respectively, above 20% and 30% (on average) of the voting rights in the 20 largest US public companies (Bebchuk et al. 2017b). This is very important for activists, who need to be able to speak rapidly with the people casting a majority of the votes. The situation is similar in Europe. A recent OECD study reveals that institutional investors own nearly 90% of UK listed equities (Isaksson and Celik 2013). Although the style of engagement differs considerably across countries, hedge funds activism consistently gets traction wherever institutional ownership is concentrated.

Although institutional investors are crucial for the success of hedge fund activism, they do not act as activist themselves. The reason is agency costs. Institutional investors are prohibited from charging performance fees. Instead, they charge fees as a proportion of the assets under management, which is for them the key variable to maximize. Differently from hedge funds, institutional investors care about relative, not absolute performance. Asset managers do not have incentives to monitor individual companies because competitors will free ride on their efforts. Therefore, the activists' teaming up with institutional investors seems to be beneficial for corporate governance. On the one hand, activists lower the agency costs of institutional ownership. On the other hand, institutional investors screen the activists' proposals and should sanction only the value-increasing ones.

Reducing agency costs undoubtedly improves the efficiency of corporate governance. However, this does not imply that hedge fund activism is always value increasing. Several objections have been raised concerning the judgment of institutional investors.

Critique of Hedge Fund Activism

A first point of criticism is that institutional investors may fail to exercise judgment in the presence of hedge fund engagement. Rather, they would blindly follow the recommendations of proxy advisors, notably including global market leaders such as Institutional Shareholders Services (ISS) and Glass-Lewis, to decide whether to vote for or against hedge funds. The incentives of proxy advisors to get corporate governance right are even more questionable than those of institutional investors, if only because advisors do not have stakes in the companies for which they recommend a vote. Nevertheless, US legislation has effectively encouraged institutional investors to purchase proxy advisory services to meet the obligation to disclose their voting and avoid embarrassment (Rock 2015). The European Union is now following suit (Pacces 2017).

Empirical research suggests that the impact of proxy advisors on the voting by institutional

investors may be not as decisive as it looks. To begin with, only the smaller institutional investors systematically follow the proxy advisors. Large asset managers, such as Blackrock, Vanguard, or State Street, seem to vote independently. Moreover, every study of proxy advisors' impact faces a fundamental reverse causality problem. In the end, although there are clear synergies in judgment (McCahery et al. 2016), it is impossible to determine how much proxy advisors influence large institutional investors and how much they are influenced by them. Although there are no specific studies on the impact of proxy advisors on hedge funds activism, a US study of uncontested elections reveals that ISS advice against the management shifts at most 10% of votes.

Another fundamental critique levered at hedge funds is that they may succeed without any screening by institutional investors, if they act as a coalition, namely as a so-called "wolf pack." Empirically, wolf packs account for about 22% of the engagements observed internationally and are associated with a higher success rate than individual engagements – 78% as opposed to 46% (Becht et al. 2016). On this basis, Coffee and Palia (2016) have argued that wolf packs are a nearly riskless strategy for hedge funds, suggesting that this may lead to over-engagement. However, the impact of wolf packs seems to be overestimated. Firstly, in more than one-fifth of observed wolf pack engagements, the engagement has been unsuccessful. Because unsuccessful engagements result in losses, even wolf packs face risks. Second, wolf packs are never large enough to control a majority of the votes, which makes institutional investors still decisive. Finally and most importantly, 78% of the overt engagements mapped internationally are not wolf packs. This cannot be random because hedge funds choose their battles. If they decide to join and form a wolf pack only when success is more likely, the success rate of wolf packs is overestimated.

The recurrent objection to hedge funds activism is short-termism. This critique is more difficult to handle because short-termism means different things to different people. In one respect,

this critique is not borne out by the empirical evidence. Both in the USA and internationally, the announcement of hedge fund engagement leads, on average, to a significant increase of the stock price. This increase is not reversed for up until 5 years down the road (Bebchuk et al. 2015), provided that the engagement is effective in determining change (Becht et al. 2016). Therefore, hedge funds are not short-termist in the conventional sense of "cutting and running." While useful to defend hedge fund activism from an easy rhetoric against them, this result says nothing about whether the stock markets is myopic relative to some horizon longer than the activists' holding period (1.7 years on average), let alone whether it makes sense to consider such a longer horizon to assess the performance of any particular company.

Theory and Evidence on Hedge Fund Activism

Underlying the short-termism discussion, there is a fundamental question about the desirability of hedge fund activism. Hedge fund activism is an important feedback mechanism in corporate governance, but its efficiency is not obvious. Assuming that other stakeholders either protect themselves by contract or are protected from externalities by efficient regulation (Pacces 2012), efficient corporate governance requires maximization of shareholder value. If financial markets were informationally efficient, there would be no difference between short-term and long-term maximization of shareholder value because any asymmetry would be arbitrated away. If, however, stock markets overweight the short-term profits of a company relative to its long-term profits, albeit temporarily, there is market failure. This is a problem for corporate governance to the extent that short-termism leads managers to make value-destroying choices. Whether this is actually the case is uncertain, though, as the management may more or less intentionally err towards the long term to justify underperformance (long-termism). Defining short- and long-termism is conceptually difficult

in the absence of a consensus on what constitutes the “right term” to maximize profit.

What is “right” term depends on the “right” strategy to maximize profit; both are difficult to identify. Stock markets are an impressive source of information in this respect, but alas, they are imperfect. Because they overreact to news, mis-price risks, and are prone to asset bubbles, stock market prices do temporarily fail to incorporate the value of future profit opportunities. When this is the case, the Efficient Capital Market Hypothesis (ECMH) does not hold true. Therefore, there may be a conflict between the pursuit of short-term results, which are immediately impounded in market prices, and long-term projects, whose expected results are underweighted or even overlooked by stock prices. Because the hedge fund business model is based on stock market prices, pressure from activist hedge funds may well turn the short-termism of stock market into short-termism of managerial choice.

The question whether public companies should be managed for a shorter or a longer term cannot be answered empirically. Data reveal that successful activism, on average, is associated with a stock price increase. However, this result is uninformative because hedge fund activism produces unobservable effects, too, and because the companies for which we observe engagements cannot be meaningfully compared to those that are not engaged.

The first part of the problem is that we only observe a portion of the true activism, the overt part, whereas a great deal of activism takes places behind closed doors. Moreover, the distribution between overt and covert activism is not random. Better-managed companies react in anticipation of hedge fund engagement, whenever this is a credible threat. Activists, on the other hand, may have to make their campaign public precisely when the targeted is more mismanaged, which overestimates the observable returns from engagement.

The second part of the problem is that companies that are or can be targeted by activists fundamentally differ from those that are not and cannot be targeted. The fact that companies successfully engaged outperform a market index, on average,

does not really show that activism improves performance. It only shows that target companies were undervalued relative to a market benchmark and that activism brings performance back in line with that benchmark. These studies cannot rule out the possibility that a target company would outperform the benchmark by a larger extent, if not engaged, because this counterfactual company does not exist and, if it existed, it would be a different firm. This fallacy affects as well the studies arguing that hedge fund activism is value decreasing on the grounds that comparable companies, which have not been engaged, outperform engaged companies in the long run (Cremers et al. 2016).

In theory, whether hedge fund activism is efficient depends on context. Some companies benefit from the correction of underperformance fostered by activist hedge funds, particularly in the presence of investor expropriation or misuse of free cash. For other companies, though, underperformance is temporary and the change of strategy promoted by hedge funds can indeed destroy value. The disagreement on the proper length of time in which to assess performance reflect a more fundamental conflict between two views of the target firm, one by the activist hedge fund and the other by the incumbent management. These views normally differ on strategic issues, such as whether the company should be leaner, more focused on certain businesses and cost-effective in carrying them out, which hedge funds typically like to see perhaps because they are impatient to cash in the profit from engagement. The opposite view that the company should pursue longer-term goals, typically fostered by the management, is equally legitimate although it may procrastinate the acknowledgment of mistakes or conceal the extraction of private benefits of control. For this reason, hedge fund activism should be framed as a conflict of entrepreneurship (Pacces 2016).

Who Decides on Hedge Fund Activism

Framing hedge fund activism as a conflict of entrepreneurship brings up the question who

decides on the conflict and whether this is efficient. As explained before, hedge funds need to garner institutional investors' support in order to succeed. Institutional investors, however, differ considerably from each other in terms of investment strategy and incentives. Their propensity to exercise voice to support or to stop hedge funds likewise differs.

Based on Bushee (1988), institutional investors can be broadly characterized as transient (small stakes, high turnover), dedicated (large stakes, low turnover), and quasi-indexers (small stakes, low turnover). According to Hirschman (1970), a fundamental activator of voice is loyalty, that is, a commitment not to exit. Transient and dedicated investors are not loyal. The former enter or exit whenever there is a tiny profit to make, as is the presence of hedge fund engagements. The latter compete with hedge funds on entering and exiting portfolio companies timely. Only quasi-indexers, exemplified by index funds, are loyal in Hirschman's sense.

Index funds are likely to be the decisive institutional investors in a hedge fund campaign. They cannot exit an investment they are dissatisfied with, so long as this investment is part of the index they track. Thus, they are committed to exercising voice. Empirical evidence seems to confirm this. Ownership by index funds in the USA increases the frequency and the success rate of hedge fund engagements requiring high voting support to succeed (Appel et al. 2016). Apart from this, however, very little is known about how index funds vote or are expected to vote on a hedge fund campaign. Assuming that index funds are indeed pivotal, whether they are the right arbiters between the opposing views of hedge funds and incumbent managers depends on context.

To illustrate the conflict of entrepreneurship with an example, a strategic issue on which the views of activists and incumbent management typically collide is quality and quantity of R&D expenditures. Activist hedge funds want companies to focus on developing specific products, which usually results in cuts of R&D expenditures and larger short-term profits. The managers usually ask for the return on R&D expenditures to be

assessed over a longer horizon. Reducing R&D expenditures does not necessarily imply that a company is less innovative. There is evidence that activism increases R&D productivity and output. Hedge fund activism, however, systematically reduces R&D input. This may be not the right choice for a number of companies. In particular, nonlinear innovation with long lifecycles benefits from conglomerate structures in which R&D input is large and can be redirected internally from a project to another. Those are precisely the structures that activists seek to break up.

Drawing on their long-term commitment to index tracking, managers of large index funds, such as Blackrock and State Street, have recently made public statements to distance themselves from the short-termism of activist hedge funds. However, such statements must be taken with a grain of salt. Whereas regulation effectively compels index fund managers to vote, they do not have the right incentives to decide on firm strategy. Index fund managers cannot benefit from firm-specific monitoring because their competitors can free ride (Bebchuk et al. 2017b). In pursuing relative performance, index fund managers will rather choose low-cost voting policies that are generally appreciated by investors. That is to say, index funds may support a hedge fund's request to cut on R&D expenditures not because it is efficient, but because the target has poor corporate governance. Whether it is efficient for the particular company to engage in linear or nonlinear innovation is a more idiosyncratic question than an index fund manager can answer. And yet the choice whether to invest for the long term or focus on short-term should depend on firm-specific variables, such as the competitive environment and the length of the innovation cycle.

Relying on the judgment of index funds is efficient in other situations. Often, what prompts hedge fund activism is waste of resources. Because index funds are expected to vote in a standardized, predictable fashion on a hedge fund's memo showing waste, hedge fund activism protects investors by way of committing management to being efficient. Facing the threat of hedge funds teaming up with institutional investors, managers have to be more careful about misusing

fee cash or being unresponsive to the competitive environment. In this perspective, hedge funds activism is a tremendous tool to stop management from being lazy or building empires.

In conclusion, index funds cannot always be trusted to screen hedge fund activism because they do not have incentives to make an informed decision on individual company's strategy. Nevertheless, the incentives of index funds are aligned with the interest of the investing public regarding the control of agency costs. Therefore, the problem whether a company should be exposed to hedge funds activism does not warrant a one-size-fits-all solution. Different companies may need different degrees of exposure to activism at different points in time. As a result, individual companies should be able to choose the exposure of management to the scrutiny by hedge funds and to alter this choice over time.

Policies Towards Hedge Fund Activism

Similarly to hostile takeovers, hedge fund activism is a disciplinary mechanism in corporate governance. If anything, activism is more effective than hostile takeovers to keep managers on their toes because it is cheaper to operate. Unsurprisingly, hedge fund activism has attracted as much criticism as hostile takeovers used to do, if not more. As a result, a number of one-size-fits-all policies have been proposed to fend off hedge fund activism.

The business model of activist hedge funds is based on the purchase of a "toehold" of undervalued stock to secure a reward for screening the market for potential targets. If target identification were revealed before the engagement, other investors would free ride on hedge funds' investment and reduce, if not eliminate, their reward. Free-riding is a mechanism that fundamentally undermines hedge fund activism, as it undermines hostile takeovers for comparable reasons. By nurturing free riding, the regulation of ownership disclosure and shareholder identification curbs hedge fund activism across the board.

The purpose of ownership disclosure is to reveal the build-up of significant stakes in a

company. This is important information for the investing public and the company's management. Regulation mandates transparency of large ownership on both sides of the Atlantic. The specific rules, however, differ between the USA and the EU. In the USA, ownership disclosure is triggered by the crossing of a 5% beneficial ownership threshold, after which the shareholder has 10 days to disclose its stake. These rules are sufficiently lenient to make the USA one of the legal environments most favorable to hedge fund activism. Although EU regulation also mandates a 5% beneficial ownership threshold, the time window to disclose it is shorter (4 days), and more important, these are only minimum requirements. The EU member states can and do set lower thresholds and shorter time windows, for instance in the UK, Italy, and the Netherlands. These stricter rules make hedge fund activism less profitable and thus less likely to happen.

In the USA, attempts have been made to curb hedge fund activism through stricter rules on ownership disclosure. Antiactivist legislation is well illustrated by the so-called Brokaw Act, proposed in 2016 by Senators Bernie Sanders and Elizabeth Warren (Brav et al. 2016). This proposed legislation took on board several previous anti-activist proposals (Coffee and Palia 2016). In the Brokaw Act, the most important curbs to hedge fund activism are: (a) shortening the disclosure period from ten to two days and (b) conferring upon the Securities Exchange Commission (SEC) the authority to determine when a group of hedge fund acted as wolf pack, and to mandate disclosure accordingly. The Brokaw Act has not become law at the time of writing.

In the EU, curbs on hedge fund activism have been adopted not by altering the regulation of ownership disclosure, but by introducing EU-wide shareholder identification obligations (Pacces 2017). Different jurisdictions have different rules on shareholder identification. In the USA, institutional investors can opt out of identification altogether. The EU rules, instead, seek to harmonize shareholder identification rules by requiring that all shareholders above 0.5% be identified by the company's management and the investing public. Differently from ownership

disclosure, the purpose of shareholder identification is to facilitate coordination between non-controlling shareholders. Whereas this policy seems consistent with encouraging the engagement of institutional investors, which is a goal of the EU, the cost of shareholder identification for corporate governance may be higher than the benefits (Enriques et al. 2010). Particularly the cost for hedge fund activism can be substantial, unless hedge funds manage to circumvent identification and purchase several toeholds without crossing the 0.5% threshold.

Ownership disclosure and shareholder identification are two examples of regulation curbing hedge fund activism across the board. As argued before, such one-size-fits-all curbs are inefficient. Law should enable individual companies to tailor exposure to activism to their circumstances, for instance, depending on whether is optimal for them to profile on short-term or longer-term strategies.

In several jurisdictions, companies can effectively opt out of hedge fund activism through dual-class shares, which are interesting relative to other antiactivist tools (such as low-trigger poison pills) because they commit some of the controller's own wealth to the claim that a project must be evaluated in the long term. The problem is that dual class-shares can only be introduced before the company has gone public, unless they are presented as loyalty shares. Formally, loyalty shares provide super-voting rights to any owner that retains the shares for long enough – say, 2 years. Practically, however, loyalty shares are only interesting for controlling owners, because institutional investors are reluctant to give up the higher liquidity of common stock regardless of the average length of their investment horizon.

Both dual-class and loyalty shares confer upon the controlling management super-voting rights. A founder concerned with the adverse impact of activism on certain styles of innovation can go public with dual-class shares, as for instance, Google did. In already listed companies, however, the management cannot exchange existing shares for two classes of shares, one with higher voting rights for the controlling group and one with lower voting rights for the investors. Such dual-class recapitalizations are prohibited. Loyalty shares

are a way out of this prohibition. Companies can issue them after having gone public because, formally, loyalty shares do not discriminate between shareholders. European legislation, for example in France, has sometimes gone as far as to allow the introduction of loyalty shares even despite the opposition of a majority of institutional investors.

Loyalty shares are a good instrument for individual companies to tailor exposure to hedge fund activism. However, managers could abuse their power to introduce them. It would be more efficient to allow dual-class recapitalizations explicitly and subject them to a veto by institutional investors. The ex-ante approval of such transactions by a Majority of Minority shareholders (MOM) is a good regime to support contracting between management and shareholders in the midstream. If a MOM vote is required to introduce dual-class shares, the management will have to persuade institutional investors that insulating the company from hedge fund activism, perhaps for a limited time, is value increasing.

Conclusion and Future Research

This entry discusses the role of hedge fund activism in corporate governance. Activist hedge funds are an important source of feedback in corporate governance. However, their influence is not always efficient. Although other institutional investors are decisive on a hedge fund campaign, their judgment cannot always be trusted. In particular, index funds appear to be most often decisive on a hedge fund campaign. However, they do not have the right incentives to decide on company's strategy. On the contrary, the optimal decision whether a company should be managed for the short or the long term depends on the circumstances faced by the individual company. Therefore, individual companies should be able to tailor exposure to hedge fund activism to their needs and to alter this choice over time.

We still do not know enough about hedge fund activism, because most part of it happens behind closed doors. In order for companies and legislatures to fine-tune the rules on hedge fund activism, it would be useful to know more about how

institutional shareholders actually vote and whether different categories of institutional investors are decisive on different campaigns. Moreover, it would be good to know more about the role of coalitions in determining success of hedge fund activism. These are interesting topics for future research.

Cross-References

- [Concentrated Ownership](#)
- [Governance](#)
- [Hirschman, Albert. O.](#)
- [Law and Finance](#)

References

- Appel IR et al (2016) Standing on the shoulders of giants: the effect of passive investors on activism. NBER working paper No. 22707
- Bebchuk LA et al (2015) The long-term effects of hedge fund activism. *Columbia Law Rev* 115:1085
- Bebchuk LA et al (2017a) Dancing with activists. *Columbia Business School research paper* No. 17–44. SSRN: <https://ssrn.com/abstract=2948869>
- Bebchuk LA et al (2017b) The agency problems of institutional investors. SSRN: <https://ssrn.com/abstract=2982617>
- Becht M et al (2016) Returns to hedge fund activism: an international study. *Review of financial studies*, forthcoming. European Corporate Governance Institute (ECGI) – finance working paper No. 402/2014. <https://doi.org/10.2139/ssrn.2376271>
- Bratton W, Mc Cahery J (eds) (2015) *Institutional investor activism – hedge funds and private equity, economics and regulation*. Oxford University Press, Oxford
- Brav A et al (2016) Anti-activist legislation: the curious case of the Brokaw Act. Working Paper. <https://doi.org/10.2139/ssrn.2860167>
- Bushee BJ (1988) The influence of institutional investors on myopic R&D investment behaviour. *Account Rev* 73(3):305–333
- Coffee JC, Palia D (2016) The wolf at the door: the impact of hedge fund activism on corporate governance. *Ann Corp Gov* 1(1):1–94
- Cremers M et al (2016) Hedge fund activism and long-term firm value. Working Paper. <https://doi.org/10.2139/ssrn.2693231>
- Enriques L et al (2010) Mandatory and contract-based shareholding disclosure. *Uniform Law Rev* 15(3–4):713–742
- Gillan S, Starks LT (2007) The evolution of shareholder activism in the United States. *J Appl Corp Financ* 19(1):55–73
- Gilson RJ, Gordon JN (2013) The agency costs of agency capitalism: activist investors and revaluation of governance rights. *Columbia Law Rev* 113:863
- Hirschman AO (1970) *Exit, voice, and loyalty: responses to decline in firms, organizations, and states*. Harvard University Press, Cambridge, MA
- Isaksson M, Celik S (2013) Who cares: Corporate Governance in today's equity markets. OECD Corporate Governance working papers 8:1
- Klein A, Zur E (2009) Entrepreneurial shareholder activism: hedge funds and other private investors. *J Financ* 64(1):187–229
- Macey J (2008) *Corporate governance: promises kept, promises broken*. Princeton University Press, Princeton
- McCahery J et al (2016) Behind the scenes: the corporate governance preferences of institutional investors. *J Financ* 71(6):2905–2932
- Pacces AM (2012) *Rethinking corporate governance: the law and economics of control powers*. Routledge, Abington
- Pacces AM (2016) Exit, voice and loyalty from the perspective of hedge funds activism in corporate governance. *Erasmus Law Rev* 9(4):199–216
- Pacces AM (2017) Hedge fund activism and the revision of the shareholder rights directive. European Corporate Governance Institute (ECGI) – law working paper No. 353/2017. <https://doi.org/10.2139/ssrn.2953992>
- Rock EB (2015) Institutional Investors in Corporate Governance. In: Gordon J, Ringe WG (eds) *The Oxford handbook of corporate law and governance*. Oxford University Press, Oxford

Heilbroner, Robert

Cyrille Ferraton¹ and Ludovic Frobert²

¹Université Montpellier 3, Montpellier, France

²Centre National de la Recherche Scientifique, Paris, France

Abstract

In this entry, we would like to show that Robert Heilbroner was not only the author of the famous best seller, *The Worldly Philosophers*, first published in 1953. Actually, he was an active and innovative scholar investigating many subjects, from development to finance, learning constantly on Marx, Keynes, or Schumpeter, and crossing boundaries between many social sciences. However the very core of his project remains a general inquiry on the nature and evolution of capitalism. He could

be considered as a late representative of institutionalism and a brother in arm of economists like John Kenneth Galbraith, Gunnar Myrdal, or Allan Gruchy. It is in the context of this general inquiry on capitalism that Heilbroner presented his main contributions to law and economics.

Introduction

Robert Heilbroner is mainly recognized as the author of an economics best seller, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economics Thinkers* (1953). It is very likely that this success partly masks his numerous other works, bearing especially on political economy, capitalism, socialism, or economic methodology.

Critical of the orientation taken by economic analysis post-*General Theory* for its historical and depoliticized nature, Heilbroner continued to argue for a situated and historical approach which takes account of the political, sociological, and moral dimensions of economic change. In this way, economic philosophers such as Adam Smith, David Ricardo, Karl Marx, Thorstein Veblen, Joseph Schumpeter, and John Maynard Keynes, all embraced economic, social, political, and psychological problematics.

Their analyses aimed simultaneously to understand and act, with the view of adapting and controlling the capitalism of their times. Therefore they were not reducible to technical dimensions or expert assessments as many contemporary economic works are. As the objectivity which those aimed for seems to him to be a vain quest, Heilbroner considers that all research supposes a “vision,” a notion borrowed from Joseph Schumpeter, who defines this as a “pre-analytic cognitive act” comprising political motivations, social stereotypes, value judgements, etc. (cited in Heilbroner 1993b, p. 89). This vision permeates the analysis for it is constitutive of all social thought, and also because it is a psychological, even an existential necessity. And as such it does not constitute “correctable weakness” (Heilbroner 1996, p. 47). Ideology leads to not recognizing the rôle played by this “vision.”

The object of economic study is not given; it is constructed and isolated from social reality. On this point, Heilbroner moves away from conventional economics which reduces its object to a method, a science of choices, so as to privilege a “substantive” definition, in the sense of Karl Polanyi. A definition takes account of institutional configurations and social interactions in the supply of resources to societies. Thus he considers that any economic system, capitalism today, or in the past traditional or planned economies, remains transitory, historically situated, and susceptible to undergoing future changes.

Biography

Heilbroner was born in 1919 in New York. While studying at the University of Harvard, he was influenced notably by Alvin Hansen, Paul Sweezy, and Joseph Schumpeter. But his predilection for economic philosophers and history was truly expressed some years later at the New School for Social Research at the side of his master Adolph Lowe (1883–1995) during a “fascinating seminar on Adam Smith” (Heilbroner 1953, p. 7). Heilbroner shared with Lowe little interest in contemporary economic analyses which treated the environment – social, cultural, political, etc. – as given, while the originality of most economists up to Keynes was precisely the proposal of a general and heuristic vision of the present and the future of the capitalist system, integrating into their analyses elements as important as politics, culture, and social questions.

Heilbroner attempted in turn to adopt this approach in his work and particularly in those bearing on capitalism, which constituted one of his preferred objects of study (see especially *The Nature and Logic of Capitalism* (1986), *Behind the Veil of Economics: Essays in the Worldly Philosophy* (1988) and *twenty-first Century Capitalism* (1993a)).

He was astonished at the almost complete disappearance in economic discourse of the notion of capitalism to the advantage of supposedly more neutral terminology such as “free market society.” Reduced to a simple technical study of

market functioning, contemporary economic work neglected the fundamental rôle of the state by evacuating the cultural values and social foundations upon which the capitalist economy rested. The analysis of institutions, involving collective representations and judicial rules as much as social organizations, constitutes an indispensable condition for understanding the *nature* and *logic* of capitalism. Conventional economics thus conveys a political and social “vision” directly inherited from capitalism which it refuses to admit. Economic reality is not identified with an unconscious emergent social order; the state and other institutions have participated and participate in its structuring. The great majority of economists do not recognize “the fundamentally social and political nature of all economies – that is, the core of political and social norms that provide the drive and the acquiescence without which no economic system, especially if marked by great inequalities, would long exist” (Heilbroner 1998, p. 4).

Innovative and Original Aspects

He borrowed from Marx the idea according to which accumulation of capital constitutes the driving force of capitalism but taking into consideration the psychological and psychoanalytic deciding factors that underlie this logic of acquisition. So, it is less conscious than unconscious motivations which explain acquisitive logic and “specifically the universal infantile need for affect and experience of frustrated aggression” (Heilbroner 1998, p. 37).

The principal motivation for the accumulation of capital should be sought beside the power and the prestige that it confers. It is about satisfying a need for domination anchored in childhood which in capitalism has no inhibition, contrary to noncapitalist societies which have succeeded in limiting it culturally. The search for profit plays an equivalent rôle to territorial conquest for military regimes or to increasing the number of the faithful for religions. Hierarchy, power, and domination are thus inseparable from the accumulation of capital. The analysis of individual choices which conventional economics is focused on is

incapable of giving an account of this ensemble of logic which can only be understood from a socio-political approach.

The markets which permit the allocation of resources and the coordination of economic activity constitute the institutional economic basis of capitalism. But taken in isolation following the model of contemporary economists, they are insufficient to qualify capitalism and notably cannot permit an understanding of the logic of acquisition.

Finally, the distinction between a public sector, place of power, and a private sector, in which market economic activity is deployed, represents a last central characteristic of the capitalist system. This distinction constitutes a historical particularity as economic activity relying on its own rules, autonomous from the rest of social activities, does not exist in noncapitalist societies. Neither do societies isolate the exercise of authority in a delimited sector. This separation is thus a source of liberties, as political dissidence can find in the economic sphere sources of expression that are inconceivable in noncapitalist societies.

However, far from being opposed, the two sectors are mutually dependent, as the accumulation of capital is not able to be achieved without the support of the state on the one hand and a prosperous economy is necessary for the financing of public activities on the other. The state even has a priority: supporting the accumulation of capital upon which depends its financial resources. Keynesianism and its inverse, the economics of the fight against inflation of the 1980s, have displayed how the capitalist system is now connected to and dependent upon the political order. The state intervenes in the economic field not to transform it or for ideological reasons but so as to secure its own future. Economic thought recognized, with Keynes, the state as an economic actor but always as a necessary interference and not as an integral part of the social capitalist order.

Identically, there is, in Heilbroner, a delegation of public functions to capitalist functions, in particular in the organization of work within the productive sphere. This is a case of an “unrecognized transfer of political power from the state into private hands” (Heilbroner 1985, p. 100). The opposition between public and

private, often presented as constitutive of capitalism, is thus brought back into cause since there is mutual interpenetration between the two sectors. Heilbroner undeniably here sees things from the perspective opened up by authors such as Karl Marx or Karl Polanyi for whom the state is a central actor of the institutionalization of the modern economy. The crises which have multiplied since the end of the nineteenth century have shown that the state can use the budgetary instrument to temporarily shore up the defects of the system. But it is firstly by the intermediary of law that the state favors the accumulation of capital. Citing Ellen Meiksins Wood, Heilbroner even considers that capitalism represents the ultimate “privatization” of politics “to the extent that functions formerly associated with coercive political power... are now firmly lodged in the private sphere” (cited in Heilbroner 1985, p. 100).

Assimilating capitalism to a uniquely private regime constitutes a gross error. The state, through socialization, by the protection and stimulation of certain economic activities, actively supports capital. Recurrent economic crises have given rise to necessary public interventions showing the adaptable and evolving character of the social capitalist order.

Impact and Legacy

The economic philosophers were perfectly aware of this interpenetration of the private and public sectors which constituted in their eyes an indispensable condition to the accumulation of capital of a stable social order. Heilbroner emphasizes that the economic philosophers always envisaged a “bounded future” for capitalism (the stationary state in David Ricardo, associationist socialism in John Stuart Mill, etc.) (Heilbroner 1985, p. 143). It was to prevent or delay these potentially dramatic episodes that they conceived such vast enquiries with economic dimensions but also with social and political and positive and normative aspects.

Cross-References

► Capitalism

References

- Carroll MC (1998) A future of capitalism. The economic vision of Robert Heilbroner. St. Martin's Press, New York
- Heilbroner R (1953) The worldly philosophers. The lives, times, and ideas of the great economic thinkers, 7th edn. Touchstone, New York. 1999
- Heilbroner R (1985) The nature and logic of capitalism. W. W. Norton, New York/London
- Heilbroner R (1988) Behind the veil of economics. Essays in the worldly philosophy. W. W. Norton, New York/London
- Heilbroner R (1993a) 21st century capitalism. W. W. Norton, New York/London
- Heilbroner R (1993b) Was Schumpeter right after all? J Econ Perspect 7(3):87–96
- Heilbroner R (1996) The embarrassment of economics. Challenge 39(6):46–49
- Heilbroner (1998) The “disappearance” of capitalism. World Policy J 15(2):1–7

Heterarchy

Erwin Dekker¹ and Pavel Kuchar^{2,3}

¹Erasmus School of History, Culture and Communication, Erasmus University Rotterdam, Rotterdam, The Netherlands

²Department of Economics and Finance, University of Guanajuato, DCEA-Sede Marfil, Guanajuato, GTO, Mexico

³Facultad de Economía-División de Estudios de Posgrado, Universidad Nacional Autónoma de México, Ciudad de México, México

Abstract

Whenever agents choose A instead of B, B instead of C, and C instead of A, a logical contradiction arises. This contradiction – also known as a value anomaly – characterizes genuine choices. Some organizations and firms, but also legal systems, markets, or even the human brain can be regarded as complex systems that manage the value anomaly by operating with multiple mutually incompatible ordering principles. Such a management – as opposed to a mere elimination – of these mutually incompatible values allows these systems to better cope with uncertainty, and to benefit

from the recognition of complexity. One of the central implications of these heterarchical systems is that there is no single scale on which unequals can be compared and, consequently, that commensuration is an active process which involves friction and opportunities for entrepreneurship. We argue that in a world that naturally seems to be characterized by these value anomalies, heterarchical organizations and, in particular, heterarchy as a complex system of valuation might well be a good response.

Definition

Heterarchy is a complex adaptive system of governance, an order with more than one governing principle. Heterarchies include elements of hierarchies and networks, but in a number of important ways, heterarchies are different from both of these systems of governance. The model of heterarchical governance is like plate tectonics: mutually self-contained orders with unclear hierarchies among them.

Origins of the Concept: A Value Anomaly

The concept of heterarchy dates back to 1945 when Warren S. McCulloch, a neurophysiologist, presented a problem defined by a logical contradiction that is characteristic for any system (be it a group of neurons, an individual, or an organization) that chooses A instead of B, B instead of C, and C instead of A. This value anomaly is present in any system that has to make a choice between two or more potential acts that are incompatible (McCulloch 1945, 90; cf. Shackle 1979). Without the logical contradiction – without two or more potential acts that are incompatible – there is no space for a genuine choice based on disparate evaluation criteria. A decision ultimately relying on an algorithm, on an overarching metric, suffices.

Law and economics as a discipline has been defined by such a logical contradiction. The field has forever faced a tension between the internal

order of the market and the internal order of the law. One tendency, when faced with such a heterarchy of values, is to attempt to regulate or otherwise clear up the conflict by defining and applying a hierarchy of values. This approach has been adopted by scholars developing the economic analysis of law, which is aimed at demonstrating that law really is about efficiency (Posner 1973).

Hierarchies of Values, Commensuration, and Genuine Choice

Whether we like it or not, conflicting demands and contradicting values form integral parts of complex systems which cannot be reduced to a single ordering principle. What is right may be, sometimes at least, fundamentally at odds with what is efficient, and getting rid of the heterarchy of values might be the wrong approach to begin with. A growing number of social scientists have recognized that heterarchies actually have strengths of their own, which we often fail to appreciate. These scholars suggest that our brain (Bruni and Giorgi 2015), firms and organizations (Hedlund 1986, 1993; Hedlund et al. 1990; Girard and Stark 2003), constitutional legal systems (Joerges et al. 2004; Halberstam 2008), and systems of global governance (Baumann and Dingwerth 2015) are all examples of heterarchies and that the value anomaly which McCulloch identified is a feature that defines these systems, not a bug to be done away with. Heterarchy seems to be the natural state of the world, and heterarchical organizations (including the brain) might well be a response to that world.

In a heterarchy of values, different and often incompatible logics of what is valuable are at play. Yet, the difficult requirement to reconcile incompatible situational logics may be eliminated by transforming qualities into quantities, in other words, through commensuration. When an overarching value like efficiency, equality, or fairness is found, the choice is reduced to figuring out which option scores better on the value metric. Commensuration transforms the system with a heterarchy of values into one with a

hierarchy of values. In such a system, a genuine choice between incommensurables is not necessary, it is in fact ruled out. While commensuration seems to be a *sine qua non* for a rational choice, the acceptance of a hierarchy of values might have performative collateral effects – commensuration may transform the character of one’s decision-making effectively turning one’s passions from something boundless and elusive to something graspable and orderly (Nussbaum 1986). The danger of such a transformation is that when it happens, we “must see the beauty or value of bodies, souls, laws, institutions, and sciences, as all qualitatively homogeneous and intersubstitutable, differing only in quantity” (Nussbaum 1984, 68; cf. Mill 1863); when such a transformation happens in economics, we become convinced that the relation between any two desired things is always conceived of as a problem of making the right trade-off.

Heterarchy, Incommensurability, and the Exchange of Unequals

Claims about incommensurability are typically made where different modes of valuing overlap and conflict with one another (Espeland and Stevens 1998). This happens at the intersection of diverse systems of worth that sustain disparate mental models and justify conflicting institutions and situational logics. But while some kind of commensuration is necessary for markets to exist, making artifacts commensurable requires effort; no artifacts – be it persons, goods, and organizations – are essentially commensurable on their own (Dekker and Kuchař 2016).

When people trade they exchange artifacts which they value less for other ones they value more. Trade is thus an exchange of unequals but possibly also of competing principles or notions of worth (think, e.g., about the legal differences between goods and persons). If one artifact is to be traded for another in a market, conflicting principles of valuation have to be reconciled through commensuration. Legal scholars often treat this question as a formal matter, what is necessary is a system that turns possession into property. It

seems to be the case, however, that equally necessary, perhaps even a prerequisite for trade to happen in markets, is a social process that by way of commensuration legitimates trade. And indeed, market interactions are embedded in frameworks that make comparisons of unequals possible and that legitimate these comparisons. These institutional frameworks enable agents to negotiate between different modes of valuing that make up the heterarchy of values.

Entrepreneurship and Legitimation of Markets

The social process of commensuration requires entrepreneurial action because trade is not naturally legitimate. We highlight three points why this is so: First, as we have demonstrated elsewhere, some broader agreement on what an artifact is good for greatly facilitates trade and is required for its legitimacy (Dekker and Kuchař 2017). Such an agreement may emerge when the meaning of exchange is contested. Second, by introducing a good, a price tag is put on the object which makes it comparable with many other artifacts (within a circuit of commerce, cf. Kopytoff 1988). In some cases, this might be considered illegitimate, allowing us to explain the existence of separate circuits of commerce, rather than unified markets (or one big market). This happens when commensuration with (some subset of) other goods is contested. Third, the entire idea of monetary valuation might be resisted, because it devalues the “sacred.” How much does a seat in parliament cost? On the margin, how much freedom should we give up to make our markets “more competitive”? Artifacts like voting rights or justice may under some circumstances be considered as unique. When this happens, commensuration with money is contested. In such cases, a unified metric is lacking, and we face a genuine choice.

None of the abovementioned examples rely on strict commensuration in that the artifacts in question are 3 ft long or 5 ft high. But they rely on a commensuration in terms of moral categories. Within these different moral categories and

hence within different circuits of commerce, we will find different norms of ownership and different norms of exchange and redistribution that will lead to different forms of property and contract law on which well-functioning markets rest.

Applications of the Concept Across Disciplines

The idea of heterarchy has been applied in fields such as anthropology, political science and institutional economics, international political economy, organization theory, or sociology. For example, Carole Crumley, an anthropologist, defined heterarchy as “the relation of elements to one another when they are unranked or when they possess the potential for being ranked in a number of different ways” (Crumley 1995, 3). She found hierarchical theories to be inadequate in explaining the emergence of modern states. In her seminal paper, Crumley criticized the idea that order in human society must rest on hierarchical systems of governance mechanisms pointing toward a “conflation of hierarchy with order [that] makes it difficult to imagine, much less recognize and study, patterns of relations that are complex but not hierarchical” (Crumley 1995, 3).

That order and hierarchy do not necessarily come together is a prominent message throughout the work of Elinor Ostrom who suggested that a heterarchical organization of water supply in California (Ostrom 1965) or polycentric organizational structure of public safety provision (Ostrom and Parks 1973) may, in fact, be a good idea for public policy: “not a single case was found where a large centralized police department outperformed smaller departments serving similar neighborhoods” (Ostrom 2010, 644).

The concept of heterarchy has also influenced the study of international relations and sociology of law through the idea of legal pluralism. Legal pluralism may lead to conflicts, tensions, and boundary disputes which often result from collisions between plural discourses of contradictory legal regimes. These tensions present international actors with paradoxical demands. Paradoxical double-bind situations typically arise when

the social environment makes ambivalent or contradictory demands to which organizations must respond. The organizational response to these paradoxical situations tends to be neither contract nor hierarchy but hybrid networks (Hutter and Teubner 1993) that may allow for the transformation of external incompatibilities into internally manageable contradictions.

Heterarchies as Complex Systems of Valuation

Inconsistency and contradiction between institutional frameworks can bring about opportunities for innovation and change. On the other hand, the overlap between institutional orders that sustain different modes of valuation are also sites of fierce struggle between different situational logics that give rise to different standards of proper and permissible behavior. When institutional frameworks that sustain conflicting modes of valuation overlap, uncertainty arises. This uncertainty creates opportunities for proponents of alternative modes of valuation that legitimate certain applications of novel artifacts. In such a perspective, entrepreneurship is not seen as a property of individual proponents of particular modes of valuation. Rather, entrepreneurship is a property of a group that may or may not allow for a heterarchy of values (Stark 2009).

Heterarchies are complex adaptive systems precisely because they interweave a multiplicity of criteria according to which performance may be evaluated, esteemed, or appraised. These heterarchies of worth are organizational structures in which a given element may simultaneously be expressed in multiple overlapping networks that maintain separate situational logics and that constitute separate orders of worth (Stark 2017; cf. Boltanski and Thévenot 2006). Contending frameworks can themselves be a valuable organizational resource that fosters entrepreneurship by bringing about uncertainty. Consequently, entrepreneurship as a feature of an organizational structure consists in the ability to keep multiple evaluative principles in play and to exploit the resulting friction that arises as a result of their interplay. Entrepreneurship exploits indeterminacy

by keeping open diverse performance criteria rather than by fostering consensus about one set of rules that would apply to every element in the whole system.

Cross-References

- [Consensus](#)
- [Economic Analysis of Law](#)
- [Entrepreneurship](#)
- [Law and Economics](#)

References

- Baumann R, Dingwerth K (2015) Global governance vs empire: why world order moves towards heterarchy and hierarchy. *J Int Relat Dev* 18(1):104–128
- Boltanski L, Thévenot L (2006) On justification: economies of worth. Princeton University Press, Princeton
- Bruni LE, Giorgi F (2015) Towards a heterarchical approach to biology and cognition. *Prog Biophys Mol Biol* 119(3):481–492
- Crumley CL (1995) Heterarchy and the analysis of complex societies. *Archeol Pap Am Anthropol Assoc* 6(1):1–5
- Dekker E, Kuchař P (2016) Exemplary goods: the product as economic variable. *Schmollers Jahr* 136(3):237–256
- Dekker E, Kuchař P (2017) Emergent orders of worth: must we agree on more than a price? *Cosmos + Taxis* 4(1):23–35
- Espeland WN, Stevens ML (1998) Commensuration as a social process. *Annu Rev Sociol* 24:313–343
- Girard M, Stark D (2003) Heterarchies of value in Manhattan-based new media firms. *Theory Cult Soc* 20(3):77–105
- Halberstam D (2008) Constitutional heterarchy: the centrality of conflict in the European Union and the United States. University of Michigan Public Law working paper, no. 111. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1147769
- Hedlund G (1986) The hypermodern MNC – a heterarchy? *Hum Resour Manag* 25(1):9–35
- Hedlund G (1993) Assumptions of hierarchy and heterarchy, with applications to the Management of the Multinational Corporation. In: Ghoshal S, Eleanor Westney D (eds) *Organization theory and the multinational corporation*. Palgrave Macmillan, London, pp 211–236
- Hedlund G, Dag R, Bartlett CA, Doz Y, Hedlund G (1990) Action in heterarchies: new approaches to managing the MNC. *Manag Glob Firm*:15–46
- Hutter M, Teubner G (1993) The parasitic role of hybrids. *J Inst Theor Econ (JITE)/Zeitschrift Für Die Gesamte Staatswissenschaft* 149(4):706–715
- Joerges C, Sand I-J, Teubner G (eds) (2004) *Transnational governance and constitutionalism*. International studies in the theory of private law. Hart, Oxford
- Kopytoff I (1988) The cultural biography of things: commoditization as process. In: *The social life of things: commodities in cultural perspective*, 1st Paperback Edition. Cambridge University Press, Cambridge, pp 64–95
- McCulloch WS (1945) A Heterarchy of values determined by the topology of nervous nets. *Bull Math Biophys* 7(2):89–93
- Mill JS (1863) *Utilitarianism*. Parker, Son & Bourn, West Strand, London
- Nussbaum MC (1984) Plato on commensurability and desire. *Proc Aristot Soc Suppl* vol 58:55–80
- Nussbaum MC (1986) *The fragility of goodness: luck and ethics in Greek tragedy and philosophy*. Cambridge University Press, Cambridge, UK
- Ostrom E (1965) *Public entrepreneurship: a case study in ground Water Basin management*. University of California, Los Angeles
- Ostrom E (2010) Beyond markets and states: polycentric governance of complex economic systems. *Am Econ Rev* 100:641–672
- Ostrom E, Parks RB (1973) Suburban police departments: too many and too small? In: Masotti LH, Hadden JK (eds) *The urbanization of the suburbs*. Sage, Beverly Hills, pp 367–402
- Posner RA (1973) *Economic analysis of law*. Aspen Publishers, New York
- Shackle GLS (1979) *Imagination and the nature of choice*. Columbia University Press, New York
- Stark D (2009) *The sense of dissonance: accounts of worth in economic life*. Princeton University Press, Princeton
- Stark D (2017) For what it's worth. *Research in the sociology of organizations* 52 (justification, evaluation and critique in the study of organizations: contributions from French pragmatist sociology)

Hidden Economy

- [Informal Sector](#)

Higher Education

Karsten Mause
Department of Political Science (IfPol),
University of Münster, Münster, Germany

Abstract

This entry discusses the role of the market, the state, and the law in the higher education (HE) sector from a law and economics perspective.

Definition

The term “higher education” (HE) is mainly used in three related contexts. First, it is used to denote the level to which individuals have pursued their educational careers. A person who finished high school and, thereafter, received a degree from a college, university, or another HE institution is someone who possesses an HE – compared to individuals with lower levels of formal education as measured by degrees or years of schooling. Second, as will be discussed below, some claim that HE institutions offer (potential) students HE as some kind of commodity. And third, social scientists consider HE as a subsector or subsystem of modern societies alongside other societal areas such as the economic and legal system.

The Market

It is often argued that HE is a commodity which is supplied by universities and other HE institutions and demanded by students in the HE marketplace. On closer inspection, however, this simple HE-as-commodity analogy is misleading and needs some clarification from an economic perspective. What HE institutions in fact provide are academic programs which contain goods and services (lectures, tutorials, course material, etc.) that students may use during their course of studies. Moreover, HE institutions award degrees which students receive after the successful completion of a degree program. Students normally have to work hard to pass the required examinations – however, there are also so-called diploma or degree mills that sell credentials that require little or no study (Grolleau et al. 2008). While selling “fake” degrees or doctorates is not necessarily illegal, whether a “degree” from an obscure “university” really pays for the respective “degree” holder in terms of labor market success, social prestige or in another respect is a different matter.

Another issue, which is frequently discussed under the keyword “grade inflation,” is that it is by no means clear that graduates holding excellent degrees (as measured by excellent grades)

awarded by more or less prestigious HE institutions are likewise excellent in terms of knowledge and skills (Popov and Bernhardt 2013). Moreover, given the vast amount of institutions sailing under the flag “college” or “university,” one has to take a careful and critical look at the curriculum a particular degree-granting institution is providing (Altbach 2001). In the labor market, potential employers therefore have to cope with the problem of asymmetric information in the sense that they have to find out whether graduates’ labor market signals are credible – and which knowledge and skills a particular graduate of a certain college or university really possesses (Rospigliosi et al. 2014). In the HE systems of many countries, however, institutions have evolved that may help (potential) students and employers to overcome their informational problems. For example, university rankings and guidebooks try to separate good from not-so-good program providers. Accreditation agencies check whether programs and their providers fulfill certain quality standards – and, if so, award a seal of approval. Governmental authorities evaluate the quality of HE institutions and programs and award the label “state recognized.” And the good or bad reputation of a program and its provider can be considered as a market-based institution providing quality information (McPherson and Winston 1993; Mause 2010).

While academic programs and degrees are supplied by public and private HE institutions in markets, these institutions do not directly provide the market good “education.” The latter is in the possession (more precisely, in the head) of an individual. And what students can do is use the educational services (lectures, seminars, mentoring, textbooks, etc.) contained in academic programs to expand their already existing personal stock of education. The fundamental distinction between HE and educational services offered by HE institutions is not just a wordplay but refers to the real-world issue that “not all ‘schooling’ is ‘education,’ and not all ‘education’ is ‘schooling.’ Many highly schooled people are uneducated, and many highly ‘educated’ people are unschooled” (Friedman and Friedman 1980, p. 187).

Following this conceptual differentiation, HE is not a commodity like cheese or wine which can be bought in a market, as it is sometimes claimed in public debate. By contrast, the “customers” of HE institutions are producers of their own education: i.e., students have to invest time and effort to educate themselves with the help of the HE institutions’ staff members and services. These institutions may create an excellent learning environment and offer a state-of-the-art curriculum; yet, whether and how students use the provided support is another issue. And students or external observers (e.g., parents, education experts, or politicians) that are dissatisfied with “bad” lectures, seminars, professors, and so on should reflect that successful “service provision” within an academic program to some extent requires the active participation of the particular “client” (Rothschild and White 1995). In other words, there are limits to what HE institutions can do to achieve that their students are really learning and enhancing their knowledge and skills.

The State

In many if not all countries, governmental authorities play a role in the HE marketplace as characterized above. Governments at different jurisdictional levels, for instance, may be providers of academic programs. In Germany, the USA, and many other countries, there are not only private HE institutions but also public universities which are state owned and publicly subsidized to some extent. A political rationale for such governmental intervention is that “the state” intends to safeguard that the HE sector produces a sufficient number of graduates in various fields of study; a highly qualified workforce is regarded as essential to ensure the wealth of a nation and its international competitiveness. From an economic perspective, the expectation that there will be a private “underdemand” (i.e., “too few” students and graduates) and/or private “undersupply” (“too few” program providers) in certain fields of study may be an economic argument to justify government intervention in the form of public ownership

and/or public funding of HE institutions. However, due to the difficulty to forecast the future, there is a danger that governmental efforts to avoid “underproduction” and to steer the *quantity* of graduates (e.g., public funding, information campaigns) may lead to “overproduction.” That is, a well-meaning government action to stimulate demand and supply in fields where a “shortage” of graduates is diagnosed may promote the production of large numbers of graduates who subsequently have problems finding a job to match their academic qualification (Freeman 1976; Büchel et al. 2003).

In addition to governmental efforts to steer the demand and supply of academic programs, the state may intervene in the HE marketplace by regulating the *price* students have to pay for attending these programs. For example, to politically enable citizens from all social strata to go to university, currently in Germany students at public HE institutions generally do not have to pay tuition fees. The state may also regulate the *quality* of academic programs. The government may, for instance, influence what is taught and how it is taught at “its” public HE institutions. Moreover, governmental authorities could, via mandatory evaluations, monitor what is going on at private HE institutions (quality of curricula, teaching staff, graduates, etc.) and intervene if deemed necessary. From an economic perspective, it should be mentioned that there are private governance mechanisms which may be a complement or – under certain conditions – even a substitute for a direct governmental quality regulation by public bureaucrats: private accreditation agencies, university rankings by reputable organizations and media, or an HE institution’s good or bad reputation may likewise help (potential) students to separate good from bad quality and to make informed choices (McPherson and Winston 1993; Mause 2010). Whether a public, private, or public-private quality regulation in a jurisdiction’s HE system is perceived as appropriate, in democratic societies, is decided by the political system.

Furthermore, the government may control the *market entry* of providers of academic programs. In Germany, for example, governmental authorities check whether an HE institution and its

programs fulfill certain quality standards. This form of mandatory licensing is practiced mainly for the purpose of quality assurance and “consumer protection.” From an economic perspective, however, it is by no means clear that spending public money for public bureaucrats who control market entry is necessary: because it can be expected that no potential student (or only a few) will enroll in a program whose provider is not able to credibly signal in some way (e.g., via a seal of quality awarded by a reputable accreditation agency, well-appointed buildings and rooms, hiring some renowned professors, etc.) that it will offer students a high-quality program, value for money, and a degree that is recognized in the labor market. In other words, private governance mechanisms under certain conditions may be a substitute for governmental market entry control (Mause 2008).

And the Law

In many societies “the law,” which again may be influenced by government action, plays an important role in facilitating exchange in the HE marketplace (Russo 2010, 2013). For example, in the USA, the UK, and other countries, students enter contractual relationships with HE institutions. Students dissatisfied with the contractually agreed services can appeal to a court in order to legally enforce their “student contract.” Such contracts can be interpreted as another form of consumer protection in the HE sector: by this means, “student-customers” are to some extent protected by contract law (Farrington and Palfreyman 2012; Kaplin and Lee 2014).

There are other settings where students do not sign an explicit contract. This is currently the case, for example, at public universities in Germany. In this particular case, however, students study under the umbrella of public and administrative law. If, for instance, a department deviates from what is stipulated in the official study and examination regulations, dissatisfied students can appeal to university’s internal bodies (e.g., examination office), to the Ministry of Education of the state in which the public university is located, or to an

administrative court. To conclude, there are not only governmental but also private governance and legal mechanisms that govern the HE marketplace. All these governance mechanisms are imperfect and have their respective limitations. The relative performance of these mechanisms can only be assessed by means of an in-depth comparative institutional analysis of real-world markets.

References

- Altbach PG (2001) The rise of the pseudouniversities. *Int High Educ* 25:2–3
- Büchel F, de Grip A, Mertens A (eds) (2003) *Overeducation in Europe: current issues in theory and policy*. Edward Elgar, Cheltenham
- Farrington D, Palfreyman D (2012) *The law of higher education*, 2nd edn. Oxford University Press, Oxford
- Freeman RB (1976) A cobweb model of the supply and starting salary of new engineers. *Ind Labor Relat Rev* 29:236–248
- Friedman M, Friedman R (1980) *Free to choose: a personal statement*. Harcourt, New York
- Grolleau G, Lakhal T, Mzoughi N (2008) An introduction to the economics of fake degrees. *J Econ Issues* 42:673–693
- Kaplin WA, Lee BA (2014) *The law of higher education*, 5th edn. Jossey-Bass, San Francisco
- Mause K (2008) Rethinking governmental licensing of higher education institutions. *Eur J Law Econ* 25: 57–78
- Mause K (2010) Considering market-based instruments for consumer protection in higher education. *J Consum Policy* 33:29–53
- McPherson MS, Winston GC (1993) The economics of cost, price, and quality in U.S. higher education. In: McPherson MS, Schapiro MO, Winston GC (eds) *Paying the piper. Productivity, incentives, and financing in U.S. higher education*. University of Michigan Press, Ann Arbor, pp 69–107
- Popov SV, Bernhardt D (2013) University competition, grading standards, and grade inflation. *Econ Inq* 51:1764–1778
- Rospigliosi AP, Greener S, Bournier T, Sheehan M (2014) Human capital or signalling, unpacking the graduate premium. *Int J Soc Econ* 41:420–432
- Rothschild M, White LJ (1995) The analytics of the pricing of higher education and other services in which the customers are inputs. *J Polit Econ* 103:573–586
- Russo CJ (ed) (2010) *Encyclopedia of law and higher education*. Sage, Los Angeles
- Russo CJ (ed) (2013) *Handbook of comparative higher education law*. Rowman & Littlefield, Lanham

Hijacking

► Piracy, Modern Maritime

Hirschman, Albert. O.

Cyrille Ferraton¹ and Ludovic Frobert²

¹Université Montpellier 3,

Montpellier, France

²Centre National de la Recherche Scientifique,
Paris, France

Abstract

This entry will provide insights into Albert Hirschman intellectual trajectory and major concepts. We will over the main steps of his academic career and his major contributions to several and multidisciplinary fields. On the issue of law and economics, the entry underlines his high preference for pragmatic institutions (voice and the political arts of participation and deliberation at different social scales) and his everlasting caution to organic institutions (exit and spontaneous market mechanisms). However, we also state that in some of his contributions, Hirschman tried to rehabilitate, in an original and balanced way, the benefit consequences of organic institutions on the issues of growth and especially development.

Biography

We can single out two main times in the work of Albert Hirschman (Adelman 2013; Ferraton and Frobert 2003; Meldolesi 1995). The first one is devoted to development economics, a domain he is considered to have pioneered. This period laid the foundations of his approach thanks to the publication of *The Strategy of Economic Development* (1958). In retrospect, Hirschman explained that in these works his ambition was to “celebrate” and “sing the epic adventure of

development, its challenge, drama and grandeur” (Hirschman 2015, XVI). Then, around 1970, he wrote more general and cross-disciplinary works around Economics, Political Science, Sociology, History, and other social sciences. It was in 1970 that *Exit, Voice and Loyalty* appeared, which was to be followed by other contributions now considered as classics, such as *The Passions and the Interests* (1977) and *The Rhetoric of Reaction* (1991).

Of course, a sense of continuity and even unity can easily be found between the two series of contributions. Many commentators, from Michael Piore to, in the present day, Jeremy Adelman or Dany Rodrik, have emphasized the special, personal dimension of Hirschman’s work. A dimension revealed by a capacity to invent notions which are at once disconcerting and enlightening – such as the *Principle of the Hiding Hand* or the *Blessing in Disguise* – by a constant effort allowing problems to be mapped out, and so to be solved, by the habit of working upon language and by combining the paths of literary and scientific enquiry, etc. Even if he always avoided offering universal methodological solutions, Hirschman nonetheless sketched out the tacit rules of his approach: possibilism. Resisting any positivism, he emphasized that his research was “pervaded by certain common feelings, beliefs, hopes, and convictions and by the desire to persuade and to proselytize which such emotions.” To clarify his credo, he adds “for the fundamental disposition of my writing has been to push back the limits of that which is or is perceived as possible, be that at the price of a weakening of our capability, real or supposed, to discern that which is probable”; and for to this purpose, to constantly “underline the multiplicity and creative disordered of the human adventure, to bring out the uniqueness of a certain occurrence, and to perceive an entirely new way of turning a historical corner” (Hirschman 1971, 26–27)

Fashioned in this way, the iconoclastic work of Hirschman is partly the result of a tumultuous life, the academic side of which began relatively late. Very early on, and notably in 1930s Germany, he was able to observe at first hand the apparent omnipresence of violence in human transactions,

the tendency of societies, including the most advanced, to be rotted away by an apparently inexorable process of brutalization. At the origin of this process, multiple processes of decline can be observed (for nations and civilizations), and of failure or falling activity (for organizations and businesses). How can warning be given and then action taken when decline or falling activity are displayed? Against all fatalism, Hirschman chooses to interest himself in the intrinsic capacity of adaptation of human communities, as well as in the conditions of its expression; whatever the scale of the measure, the action is possible and the discourse on the laws, the regularities, the instructions of history, nature or society should be nuanced.

His work constitutes a vast enquiry about principles but also about the craft that can favor the emergence, maintenance, or defense of a reasonable social, political, or economic environment. The question of action which is collective and turned towards reform is thus at the heart of his intention. This explains that he could consider since his first works on Latin America that “the fundamental problem of development consists in generating and energizing human action in a certain direction” (Hirschman 1958, 25), or that he concentrated his later thinking upon “the micro-foundations of a democratic society” and to the “constitution of a democratic personality” (Hirschman 1995).

Innovative and Original Aspects

This general orientation thus allows us to approach the question of law/economics in Hirschman. These relations depend on the new and singular fashion by which he treats the question of institutions. We are familiar with the opposition established by Carl Menger between “organic” institutions and “pragmatic” institutions (Menger 1883). “Pragmatic” institutions present themselves as rules and organizations resulting from collective, concerted, and planned action. They make concrete a design consciously developed by one or many individuals. However, organic institutions are the unintentional

consequence of intentional individual actions. Put in another way, an organic institution is a social consequence, unanticipated, even unwanted, which is the result of actions undertaken by many individuals independently of each other. Menger explains that numerous social phenomena spring from this process in terms of organic institutions: law, language, the market, and money are all institutions that are at least partly organic.

Within the framework of this alternative, the work of Hirschman swings very clearly to the side of pragmatic institutions. A key contribution such as *Exit, Voice and Loyalty* can thus be interpreted as a defense of the paths of economic, social, and political reform by the privileged means of pragmatic action. Opposing the teachings of *The Logic of Collective Action* by Mancur Olson, Hirschman emphasizes that the efficiency of the process, described by economists as simultaneously spontaneous, optimal and exclusive, that represents the market (*exit*) faced with the different risks of “slackening off” for institutions and organizations, should be questioned. Another process, more determined (*voice*) a characteristic of political action, should, at least be considered as complementary, and doubtless, most often, as a clearly preferable alternative. *Exit* appears thus as a secondary substitute.

Hirschman’s preferences concerning the two processes are in fact obvious: while *exit* is taxed as being a form of “desertion,” “*voice* is essentially an *art* constantly evolving in new directions” (Hirschman 1970, 43) and again a faculty of invention. He specifies elsewhere that the role of speaking out within an organization can be assimilated to the exercise of a democratic control founded upon the interaction of opinions and interests. In the context of the neoliberal movement of the 1980s and attacks made upon the welfare state, the tendency to favor the pragmatic over the organic will be evident in Hirschman, engagement in favor of public action becoming the supervisor in relation to an activity which is entirely turned towards private well-being (*Shifting Involvements* 1982). A little later *The Rhetoric of Reaction* studied, so as to reveal, the three rhetorical devices systematically mobilized

over the history of the two last centuries by conservative/reactionary opinion (perversity, futility, jeopardy) to stigmatize any determined attempt to improve the civil, political, and economic situation of the greatest number.

Nevertheless, as in all parts of his work, Hirschman tries to question his own opinions (his “propensity to self-subversion”) to establish dialogue with the adverse party to improve the conditions of the explanation. Thus, he explains that at the very heart of conservative tradition a concept of social change can be seen, as well as institutions which, correctly reformulated, can be included in the arsenal of the social reformer. In fact, notes Hirschman, the Hayekian paradigm of the unexpected consequences of action and the emergence of organic institutions has too clearly been monopolized, in his opinion, by conservative thought. However, this explanatory model of social change has virtues which would merit their being solicited by progressive/reformist thought.

Progressive/reformist thought, in fact most often mobilizes a model of social change that is too crude: a model stained by Prometheism and in which change intervenes as a mass, from above and according to sequences presented as entirely determinable and planifiable in advance; a model in which actions cannot, nor must not have unexpected consequences; a model in which also dominates as a consequence the expert, the money doctor and others, *deus ex machina* of a world that is entirely measurable.

However, from his first works on development, as he was confronted with the problem of generalized underemployment of existing but dormant resources which characterizes the least advanced countries, Hirschman asked for the use of a *strategic* approach. In that case, it is necessary, as in the case of the Keynesian multiplier, but to a greater degree of generality, to create an impulse, then let the fertile multiplying phenomena develop. And in this context, it is thus necessary to reflect in strategic terms, that is to say, in terms of actions of impulse at different levels, including and especially at intermediate levels, and at different moments to support much more than

fully control the chains of phenomena which will develop.

Hirschman thus explains that, in terms of models of change, a strange twinship can be suspected to exist between the Keynesian model centered on the multiplier and the Hayekian paradigm of action. Starting in *The Strategy*, Hirschman explains that “development planning” thus consists of putting “*inducement mechanisms*” systematically in place. Here there is a clear hope placed in the generalization of certain multiplying phenomena, an initial action leading to an unforeseen chain effect of positive consequences. But here the issue at stake becomes not to declare them, for cognitive reasons or other impotencies vis-a-vis these chains of phenomena, but to learn as much about them as possible so as to recognize them and watch over them. He took up this idea later: “it is possible to plan with success; to put it another way, we are not always surprised by unexpected consequences” (Hirschman 1997, 96).

Impact and Legacy

Originally marked by the experience of totalitarianism, then by field and conceptual work, in the domain of development, and finally by thinking upon the ensemble of history and the social sciences, Hirschman thus presents a complex conception of institutions and the articulation of economics/law. A concept which proceeds from the emergence and the evolution of economic phenomena – here also should be mentioned his micro-Marxism – and which brought his attention to the multiple possibilities which, collectively, offer themselves for the exploitation of their positive dimensions, thus to form obstacles to the incessant risks of decline and failure which menace human organizations and communities.

Cross-References

- [Economic Development](#)
- [Institutional Economics](#)

References

- Adelman J (2013) *Worldly philosopher: the Odyssey of Albert Hirschman*. Princeton University Press, Princeton
- Ferraton C, Frobert L (2003) *L'Enquête inachevée: introduction à l'économie politique d'Albert Hirschman*. Presses universitaires de France, Paris
- Hirschman A (1958) *The strategy of economic development*. Yale University Press, New Haven
- Hirschman A (1970) *Exit, voice and loyalty*. Harvard University Press, Cambridge, MA
- Hirschman A (1971) *A Bias for Hope : essays on development in Latin America*. Yale University Press, New Haven
- Hirschman A (1995) *A propensity to self-subversion*. Harvard University Press, Cambridge, MA
- Hirschman A (1997) *La morale secrète de l'économiste*. Les Belles Lettres, Paris
- Hirschman A (2015) *Development projects observed*. The Brookings Institution, Washington, DC
- Meldolesi L (1995) *Discovering the possible : the surprising world of Albert Hirschman*. Notre Dame University Press, Notre Dame and London
- Menger C (1883) *Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Ökonomie insbesondere*. Dunker und Humblot, Leipzig

Horizontal Effects

Philipp Schumacher
Wildau Institute of Technology, Technical
University of Applied Science, Wildau,
Brandenburg, Germany

Definition

Horizontal effects are the anti-competitive consequences of mergers where the competitive constraints existing before the transaction between the merging parties are eliminated (unilateral or non-coordinated effects), and/or the post-merger market structure reduces the intensity of competition between the remaining competitors (pro-collusive or coordinated effects).

Horizontal Effects

Non-coordinated effects refer to a situation where, post-merger, the merging parties are able to

profitably increase prices, reduce output, or otherwise act less competitively than before, while their rivals do *not* alter their strategies. Non-coordinated effects arise from the independent profit maximization of the individual firms and are due to the internalization of competition between the merging firms. Before the merger, a unilateral price increase by one party would have led to sales lost to competing firms, including the other merging party. As a result of the merger, the parties can absorb some of the competitive pressures. The incentive to increase prices strongly depends on the closeness (or degree of substitutability) of, firstly, the merging firms' products and, secondly, the existence of close substitutes offered by non-merging firms. Consequently, one of the fundamental elements of the European Commission's Horizontal Guidelines is the assessment of the closeness of competition to forecast non-coordinated effects. In order to determine the cross-price elasticity of demand between several products, economic methodology is required for robust modeling and correct interpretation of the results.

In the Horizontal Guidelines, the EC recognizes that market share is not necessarily the best indicator of market power or non-coordinated effects in the sense of ability to raise prices, and it attempts to include additional economic aspects into the assessment, e.g., the probability of post-merger anticompetitive effects in the absence of any countervailing factors and the probability of buyer power, entry of new competitors or efficiencies as factors mitigating harmful effects on competition, as well as the conditions of a failing firm defense.

A merger is said to give rise to concerns regarding *coordinated effects* when the consequent change of the nature of competition better enables the merged firm and at least one of its competitors to reach and sustain a tacit agreement not to compete effectively with one another and thereby raise prices. A merger may also make coordination easier, more stable, or more effective for firms which were coordinating prior to the merger. Coordinated effects arise if, due to the merger, non-merging competitors change their

strategies, resulting in (or reinforcing previously existing) coordinated behavior, i.e., explicit or implicit collusion. Post-merger, it may be easier for the firms in the relevant market to imitate monopoly-type behavior by acting collectively. The coordination may take various forms, for example, keeping prices above the competitive level, limiting output or new capacity, dividing the market, or allocating contracts in bidding markets (European Commission 2004, para. 40; U.S. Department of Justice 2010, p. 24–29). The likelihood of such behavior is determined by how easy the establishment of coordination is in a given market and how sustainable it is over time, dependent, in particular, on the existence of a credible monitoring and punishment mechanism for deviations. The EC Guidelines require the commission to prove that such conditions exist in the respective market before claiming that sustained coordinated behavior is likely to occur.

According to the Guidelines, coordination is more likely to emerge in markets where it is relatively simple to reach a common understanding on the terms of coordination. Additionally, the following conditions are necessary for coordination to be sustainable:

- The firms in the coordinating group must be able to sustain any tacit understanding. This requires them to be able to monitor that the tacit understanding is being adhered to by the other firms.
- If a firm does deviate from the tacit understanding, the other firms must be able to respond effectively (i.e., retaliate) to penalize that firm.
- For the tacit understanding to be sustainable, it must be immune from the potentially destabilizing reactions of firms outside the coordinating group.

Cross-References

- [Efficiency](#)
- [GE/Honeywell](#)
- [Merger Control](#)
- [Merger Remedies](#)

References

- European Commission (2004) Guidelines on the assessment of horizontal mergers under the council regulation on the control of concentrations between undertakings. OJ C31/5. Available from: [http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52004XC0205\(02\)&from=EN](http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52004XC0205(02)&from=EN). Accessed 10 June 2017
- U.S. Department of Justice (2010) Horizontal merger guidelines (Issued August 19, 2010). Available from: <http://www.ftc.gov/sites/default/files/attachments/merger-review/100819hmg.pdf>. Accessed 10 June 2017
- Further Reading**
- Bishop S, Walker M (2009) The economics of EC competition law: concepts, application and measurement, 3rd edn. Sweet and Maxwell, London
- Hildebrand D (2009) The role of economic analysis in the EC competition rules, 3rd edn. Kluwer Law International, New York
- Kirschner US (2007) European competition policy – assessment problems in merger control. VDM Verlag Dr. Müller, Saarbrücken
- Kühn KU (2008) The coordinated effects of mergers. In: Buccirossi P (ed) Handbook of antitrust economics. The MIT Press, Cambridge, MA, pp 105–144
- Motta M (2004) Competition policy – theory and practice. Cambridge University Press, Cambridge
- Schwalbe U, Zimmer D (2009) Law and economics in European merger control. Oxford University Press, Oxford/New York
- Werden G, Froeb LM (2008) Unilateral competitive effects of horizontal mergers. In: Buccirossi P (ed) The handbook of antitrust. The MIT Press, Cambridge, MA, pp 43–104

Horizontal Product Differentiation

Javier Elizalde

Department of Economics, University of Navarra, Pamplona, Navarra, Spain

Abstract

This essay provides a definition of horizontal product differentiation and a short description of its implications for Law and Economics, as the degree of product differentiation affects the degree of competition and therefore the level of prices. A short revision of the literature on horizontal product differentiation is performed.

The aim of the literature is to predict the number of varieties or the degree of differentiation (when the number of varieties is fixed) in an industry. Different works find different solutions to this problem.

Definition

Horizontal product differentiation takes place when, at a given price, some consumers prefer different varieties of the goods.

Horizontal Product Differentiation

A number of products are horizontally differentiated when, if they were all sold at the same price, all of them would be demanded by one or more consumers. The existence of horizontal product differentiation implies that the market cannot be described by perfect competition but firms enjoy some degree of market power as, if one of the firms increases its price, some consumers would switch to a competitor but some other consumers would keep loyal to their original supplier.

From the point of view of law and economics, the existence of horizontal product differentiation is relevant in order to identify the competitors for goods and therefore to define the relevant market. The degree of product differentiation determines the degree of market power, as the more similar the products, the more intense the price competition.

There are two alternative approaches to the analysis of competition with differentiated products. The first approach considered that all consumers had the same preferences for the different varieties of the goods, which were fixed (Bowley 1924). This analysis was modified by considering endogenous varieties and free entry, in what has been called in the literature the model of monopolistic competition (Chamberlin 1933). The second approach to the analysis of horizontal product differentiation is called the *address (or distance) approach* as it considers the degree of product differentiation as geographical distance between the different firms (Hotelling 1929).

Hotelling analyzed products which were differentiated in only one dimension (in his own example, he considered that cider could be simplified to be differentiated by the degree of sourness alone), and different consumers had different preferences for the different varieties of the goods. In the original Hotelling model, two firms entered simultaneously in a linear market and decided the product characteristic in the first stage and the price in the second stage. Any decision to change the product specification has a price effect and a market share effect: A firm has an incentive to offer a closer substitute to its rival's product in order to gain some of its customers, but this closeness implies a higher degree of price competition. The main goal of the address approach to horizontal product differentiation has been to define the equilibrium in the characteristics (or location) game, which represents the degree of product differentiation in the market. The prediction of Hotelling was that the two firms would be located in the center of the linear market implying minimum product differentiation. It was shown that the analysis of Hotelling failed when the firms were closely located as none of the price or market share approaches dominated the other (d'Aspremont et al. 1979).

The analysis of Hotelling was modified in order to get a perfect equilibrium in the two stages of the game. One of the assumptions that were modified was that the transportation costs of consumers are linear. By considering quadratic transportation costs, d'Aspremont, Gabszewicz, and Thisse solved the problem of nonexistence of equilibrium predicting maximum product differentiation. Other assumptions of the original work (primarily the existence of two firms, differentiation in a single dimension, simultaneous entry, and the fact that all consumers purchase one unit of the goods regardless of price) have been relaxed in recent works finding support for either maximum or intermediate degrees of product differentiation (Prescott and Visscher 1977; Salop 1979; Economides 1984, 1986; Irmen and Thisse 1998).

There are two noteworthy versions of the horizontal product differentiation analysis: The one that considered fixed prices and a fixed number of firms and looked for the location equilibrium for

each number of firms (Eaton and Lipsey 1975) and the circular market with an endogenous number of firms (Salop 1979).

References

- Bowley A (1924) The mathematical groundwork of economics. Oxford University Press, Oxford
- Chamberlin E (1933) The theory of monopolistic competition. Harvard University Press, Cambridge, MA
- d'Aspremont C, Gabszewicz JJ, Thisse JF (1979) On Hotelling's 'stability in competition'. *Econometrica* 47:1145–1150
- Eaton B, Lipsey R (1975) The principle of minimum differentiation reconsidered: some new developments in the theory of spatial competition. *Rev Econ Stud* 42:27–49
- Economides N (1984) The principle of minimum differentiation revisited. *Eur Econ Rev* 24:345–368
- Economides N (1986) Nash equilibrium in duopoly with products defined by two characteristics. *Rand J Econ* 17:431–439
- Hotelling H (1929) Stability in competition. *Econ J* 39:41–57
- Irmen A, Thisse JF (1998) Competition in multi-characteristics spaces: Hotelling was almost right. *J Econ Theory* 78:76–102
- Prescott EC, Visscher M (1977) Sequential location among firms with foresight. *Bell J Econ* 8:378–393
- Salop S (1979) Monopolistic competition with outside goods. *Bell J Econ* 10:141–156

Human Experimentation

Roberto Ippoliti

Department of Management, University of Turin, Turin, Italy

Scientific Promotion, Hospital of Alessandria – “SS. Antonio e Biagio e Cesare Arrigo”, Alessandria, Italy

Abstract

Human experimentations refer to medical investigations with humans involved as experimental subjects. These trials are prospective researches aimed at answering specific questions about biomedical interventions, generating safety and efficacy data on drugs and treatments, innovative devices or novel

vaccines, as well as new ways of using known interventions.

After a brief introduction of human experimentations, this section presents the main related issues in Law and Economics, i.e., the risk sharing between society and pharmaceutical industry.

Synonyms

Clinical trials; Medical experimentation

Definition

Human experimentation refers to a scientific investigation with humans involved as experimental subjects.

Clinical Research and its Organization

Human experimentations refer to medical investigations with humans involved as experimental subjects. These trials are prospective researches aimed at answering specific questions about biomedical interventions, generating safety and efficacy data on drugs and treatments, innovative devices or novel vaccines, as well as new ways of using known interventions. The investigation requires an ex-ante authorization by the competent Institutional Review Board (IRB), also called Ethic Committee in Europe, which is an independent authority designed to approve the experimental activity if an appropriate level of safety for the involved subjects is guaranteed.

According to a stylization elaborated by the US National Health Institute, the human experimentation can be classified as follows:

- Studies of phase I, researchers test an experimental treatment on a small group of healthy and voluntary subjects (20–80) in order to evaluate its safety, determine the dosage, and identify side effects.
- Studies of phase II, the experimental investigation is extended to a larger group of patients

(100–300) to see whether it is effective and further evaluate the safety.

- Studies of phase III, the experimentation involves 1,000–3,000 individuals and it is directed to gather a more extended dataset of information concerning its effectiveness of the treatment, side effects, the comparison with already existing treatments, and many other details.

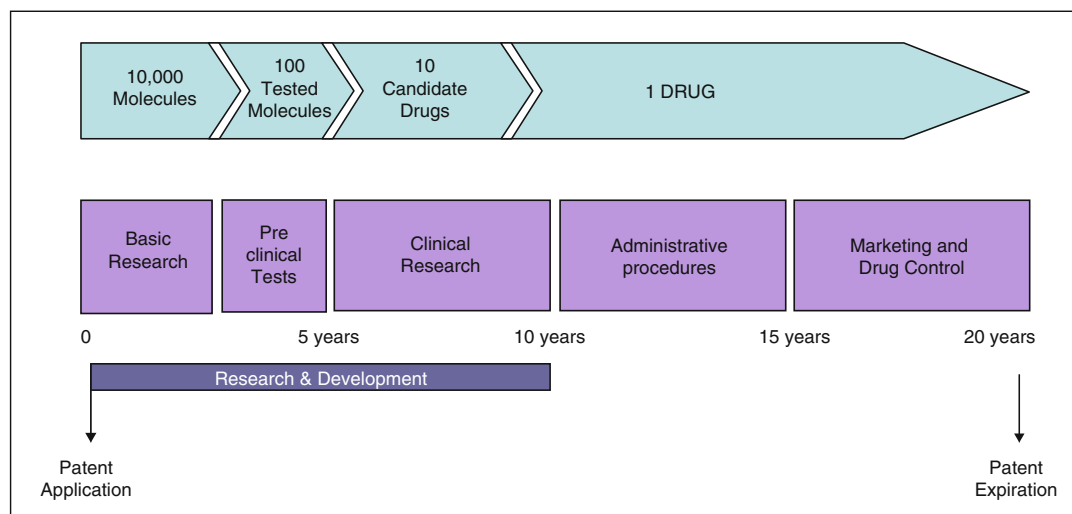
The classification is based on the development stage of the innovative product and/or treatment. At earlier stages, investigators enroll volunteers and/or patients into small (pilot) studies and then, if positive safety and efficacy data are collected, they conduct progressively larger-scale studies. This is exactly the main risk of human experimentation, i.e., the possibility of collecting high level of toxicity and negative efficacy data, raising the main Law and Economics issue which is related to the human experimentation: the risk sharing between society and pharmaceutical industry (Ippoliti 2013).

Pharmaceutical Research and Development (R&D) is a complex of activities aimed at discovering and developing new products to care patients. Generally, the pharmaceutical R&D can

be stylized into two subsequent phases (Criscuolo 2005):

- The first stage, which is the basic research devoted to advance the current knowledge by designing new productive opportunities, i.e., new molecules, new compounds, or new techniques
- The second stage, which tries to make the previous discoveries workable by testing the innovative products on animals (i.e., preclinical test) and then on humans (i.e., clinical research or human experimentation)

The following figure summarizes the entire production process of the pharmaceutical company and, as it can be easily seen, the R&D process is the main productive activity with a deep economic investment. More precisely, clinical research is directed at collecting clinical evidence to obtain from the national regulatory agencies the authorization to market the innovative products. This step is mandatory in order to commercialize a drug and thus to make the investment profitable. However, as aforementioned, another opportunity to make the R&D activity easily profitable is risk sharing (Fig. 1).



Human Experimentation, Fig. 1 Production process of the pharmaceutical industry (Source: Les Entreprises du Médicament (LEEM Report 2008))

The sharing idea is linked to unexpected adverse events which companies and patients must face. According to Ippoliti (2013), in order to expand medical knowledge, not only does society need to accept companies' market power (monopoly) but also a part of the adverse risks linked to human experimentation. Only if patients accept to bear all expected adverse risks, signing the informed consent (The Association of the British Pharmaceutical Industry 1994), the opportunities to make the investment worthwhile will raise since the expected cost will decrease. The key role of this risk sharing is the informed consent, which might be viewed as a contract between research subjects and sponsor.

Cross-References

► [Institutional Review Board](#)

References

- Criscuolo P (2005) On the road again: researcher mobility inside the R&D network. *Res Policy* 34:1350–1365
- Ippoliti R (2013) The market of human experimentation. *Eur J Law Econ* 35(1):61–85
- LEEM report (2008) *L'industrie du médicament en France, réalités économiques* Les Entreprises du Médicament (LEEM), 2008 edn, Paris.
- The Association of the British Pharmaceutical Industry (1994) *Clinical trial compensation guidelines*. ABPI, London

Imitation

Jen-Te Yao
Fu-Jen Catholic University, New Taipei City,
Taiwan

Abstract

Imitation is an idea adopted or derived from an original one and is a necessary tool for disseminating knowledge. The existence of imitation may lower the monetary returns of innovators and remove their creation incentives. This gives the rationale for the intellectual property right (IPR) protection system, whereby the returns of innovators can be secured and the incentives to innovate can be improved. The pertinence of such a viewpoint is, however, in doubt in some of innovative industries, such as software, computers, and semiconductors. Institutions that govern the creation and diffusion of inventions (knowledge) should balance the trade-off between IPR protection and imitating diffusion.

Introduction

Invention creates new knowledge, and a successful invention often leads to widespread imitation of other innovations. Imitation is an idea adopted or derived from an original one and is a necessary

tool for disseminating knowledge. Knowledge accumulation and innovation diffusion must be achieved via imitation. While invention is a discovery and proof of the workability from a new idea or method, innovation is the successful application of new inventions into marketable products.

Imitation is related to copying. Copying is an extreme case of imitation, because it is an exact reproduction of the original, such as a fake painting. Sometimes a fake painting can be so real looking that it is hard to tell the fake from the genuine article. One might recall a common proverb that “imitation is the sincerest form of flattery,” although artists and museum directors would probably disagree. An excellent invention may signal to imitators that opportunities are “good” and hence implicitly encouraging imitation. Without this signal, potential competitors might have no incentives to imitate.

When discussing imitation, we should understand the difference between imitation, counterfeiting, and piracy. Counterfeiting refers to the illegal copying of trademarks, patent or copyright infringement, where the protected rights to intellectual property are being violated. A counterfeit good is an unauthorized imitation of a branded good. Piracy is making an unauthorized exact copy – not a simple imitation – of an item covered by intellectual property rights (hereafter IPRs).

Imitation and the Incentives to Invent

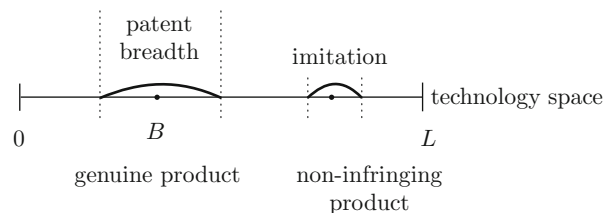
Imitation as a factor in economic development has gradually received greater attention. A number of economists have examined the subject. They argued that firms imitating major inventions and diffusing them throughout the economy – not the original inventions themselves – play a critical role in economic growth. This is because complete technological progress consists of three stages: invention, imitation, and diffusion. In the last stage, diffusion is the process of the spreading of inventions through licensing, imitation of patented innovations, or adoption of unpatented innovations. Nobel Prize winner Jean Tirole emphasizes the importance of imitation and diffusion in economic progress. He notes that “Inventing new products is not sufficient for economic progress. The innovations must then be properly exploited and diffused through licensing, imitation, or simple adoption. The difference between imitation and adoption is that imitators must pay for reverse engineering – e.g., figuring out the original firm’s technology” (Tirole 1988). Therefore, an imitating work is not free, but rather has a cost.

Inventions and innovations require monetary returns to innovators in a market-based system. However, the existence of imitation may lower the returns and remove the creation incentives of innovators. This gives the rationale for the IPR protection system, whereby the returns of innovators can be secured. In general, stronger IPR protection raises imitation costs and reduces the number of imitators, which can be realized from the mechanism of patent protection. Patent protection has two dimensions: patent length protection and patent breadth protection. Nordhaus (1969) is the first to offer a model dealing with patent life or patent length, i.e., the number of years that the patent is in force and protected

from being copied. Nordhaus (1972) extends his model to study the effectiveness of patent breadth protection. Patent breadth indicates the extent to which a given patent covers the field to which it pertains. The breadth of a patent grant measures the degree of protection in the scope of a product’s characteristics. For example, if a company invents a new drug to alleviate a heart condition, to what degree can a competitor be allowed to sell a similar drug? If a computer software firm markets a new program, how different should rival products be from the original? One can conclude that if patents are narrow, then they are easy to “invent around”; that is, it is easy to produce a non-infringing substitute for the patented inventions. This constitutes a non-infringing imitation. An extremely narrow patent does not protect a product even against trivial changes, like color or size (Scotchmer 1991).

One can apply a spatial line to express this idea, as illustrated in Fig. 1. Suppose a firm invents a new product and obtains a patent grant. The breadth B of the patent is an exclusion zone in a technology space, whereby the patent holder has exclusive control over the application of the patented technology. The technology space is a closed interval due to the limitation of technology development. Now, a competitor enters the market and needs to consider where to locate its product characteristic legally in the technology space, in order to avoid infringement of the patent. For example, the competitor needs to employ engineers to conduct reverse engineering and to figure out the original firm’s technology. Such operations are costly. Imitation entry by a second firm causes the patented firm to compete in the given technology space. However, a broader patent protection leaves a smaller remaining technology space to potential competitors (imitators), making it more difficult for them to enter the

Imitation, Fig. 1 Non-infringing imitation in a technology space



market. One thus concludes that a broader patent leads to less entry and less competition. Consequently, the market price is driven to increase due to less competition, which hurts consumer surplus, but increases the original firm's profits. Imitation entry stops when the cost of imitation can no longer be covered by competition in the market.

Two implications are derived from the above discussion. First, a broader patent increases the original firm's profits due to less competition in the market, thereby improving the incentives of the original firm to engage in innovations. Second, an increase in patent breadth means a rise in imitators' entry cost. Such a view is supported by Gallini (1992). One can further infer that an increase in patent protection will reduce the phenomenon of imitation and thus reduce competition between firms in the market. On the other hand, if a patent is very narrow (and even in some cases where there is no patent protection) and imitation is sufficiently cheap, then firms will prefer to imitate rather than innovate. Under this situation a "waiting game" takes place (Katz and Shapiro 1987) – all firms wait for other firms' inventions to occur, and the result is in fact no innovations. See also Dasgupta (1988) for an early discussion and Choi (1998) who as part of a wider paper on patent litigation, patent strength, and imitation investigates the waiting game style behavior in imitation.

The pertinence of the above viewpoints is, however, much in doubt (Takalo 2001). Ever since the pioneering study by Mansfield (1961), researchers have reported evidence of the inability of patent protection to prevent imitation, with a few exceptions such as in the pharmaceutical industry. Bessen and Maskin (2009) observe an interesting phenomenon that some of the most innovative industries of the last 40 years – software, computers, and semiconductors – have historically had weak patent protection and have experienced rapid imitation of their products. Surveys of managers in semiconductor and computer firms typically report that patents only weakly protect innovation. Patents were rated weak at protecting the returns to innovation, far behind the protection gained from lead time and learning-curve

advantages (Levin et al. 1987). Patents in the electronics industries have been estimated to increase imitation costs by only 7% (Mansfield et al. 1981) or 7–15% (Levin et al. 1987).

Imitation and Social Welfare

The strength of IPR protection substantially affects the speed of imitation. The following question arises: Does a strengthening of IPR protection have economic benefits for the society? The empirical literature on innovation shows that "imitation" is a nontrivial exercise that (even in the absence of a patent) may require substantial time and effort (Dosi 1988). Increased protection of IPRs makes imitation more difficult, which generates a trade-off on the improvement of social welfare. On the consumer demand side, stricter IPR protection implies that there will be less price competition between firms, resulting in higher prices and thus a reduction of consumer surplus. However, on the supply side, stricter IPR protection reduces competition between firms, which creates greater returns for the existing innovators and provides incentives for them to innovate. Therefore, stricter IPR protection is bad for consumers' well-being, but is good for innovators' private returns: no one wants to invest in the creation of inventions if imitation cost is very low and dissemination of inventions occurs rapidly.

To summarize, from society's point of view, the welfare effects of imitation depend on the balance of two elements: one is to provide a means for the knowledge producer to capture the benefits of efforts put forth and the other is to maximize the social dissemination of the related inventions. Institutions that govern the creation and diffusion of inventions (knowledge) should balance this trade-off, and this is why it is important to devise social mechanisms to allow inventors to capture a fraction of the benefits generated by the inventions.

Cross-References

- [Counterfeiting Models: Mathematical/Economic](#)

References

- Bessen J, Maskin E (2009) Sequential innovation, patents, and imitation. *RAND J Econ* 40(4):611–635
- Choi JP (1998) Patent litigation as an information-transmission mechanism. *Am Econ Rev* 88(5):1249–1263
- Dasgupta P (1988) Patents, priority and imitation or, the economics of races and waiting games. *Econ J* 98:66–80
- Dosi G (1988) Sources, procedures, and microeconomic effects of innovation. *J Econ Lit* 26(3):1120–1171
- Gallini N (1992) Patent policy and costly imitation. *RAND J Econ* 23:52–63
- Katz ML, Shapiro C (1987) R&D rivalry with licensing or imitation. *Am Econ Rev* 77(3):402–420
- Levin R, Klevorick A, Nelson R, Winter S (1987) Appropriating the returns from industrial research and development. *Brook Pap Econ Act* 3:783–820
- Mansfield E (1961) Technical change and the rate of imitation. *Econometrica* 29:741–766
- Mansfield E, Schwartz M, Wagner S (1981) Imitation costs and patents: an empirical study. *Econ J* 91:907–918
- Nordhaus W (1969) *Invention, growth and welfare*. MIT Press, Cambridge, MA
- Nordhaus W (1972) The optimal life of the patent: reply. *Am Econ Rev* 62:428–431
- Scotchmer S (1991) Standing on the shoulders of giants: cumulative research and the patent law. *J Econ Perspect* 5(1):29–41
- Tirole J (1988) *The theory of industrial organization*. MIT Press, Cambridge, MA
- Takalo T (2001) On the optimal patent policy. *Finn Econ Pap* 14(1):33–40

Immigration Law

Thomas Eger and Franziska Weber
Faculty of Law, Institute of Law and Economics,
University of Hamburg, Hamburg, Germany

Abstract

In this article we provide some insights into relevant law and economics research on immigration laws and policies. After discussing some typical motives of migrants, we deal with the most important welfare effects of immigration

and their distribution, and try to understand why nation states regulate immigration more restrictively than the mobility of goods and capital. We refer to the example of the free movement of EU citizens. Lastly, we present some economic insights into asylum law.

Definition

This contribution considers scholarly works on immigration law with a focus on law and economics insights into labor migration and refugee law.

Introduction

Immigration, i.e., the movement of people – usually for permanent residence – into another country or region to which they are not native, is in many respects regulated by the countries concerned. In the following, we discuss some typical motives of migrants ([why migrate?](#)), deal with the most important welfare effects of immigration and their distribution ([Some basic welfare effects of migration](#)), and try to understand why nation states regulate immigration more restrictively than the mobility of goods and capital and why international agreements on immigration are less frequent than those on trade and investment ([Why and how do modern states regulate immigration?](#)). In this chapter, we also discuss the free movement of people in the European Union as an example for a far-reaching cooperation in this field. Finally, we conclude this entry with some ideas on asylum law from an economic perspective ([Asylum laws](#)).

Why Migrate?

A fundamental line of research regarding immigration matters concerns individual's motives to migrate. Migration can be triggered by a variety of reasons. Labor migration is a common form; other migration is family based. Some migrants may be attracted by another state's social welfare system. Others expect benefits from committing crimes or terrorist attacks in the country of destination. Many people are forced to leave their

We wish to thank Jerg Gutmann, Peter Weise and Katharina Eisele for valuable comments and André Plaster for useful research assistance.

country of origin because of political or religious persecution or are even victims of human trafficking. Needless to say, migrants are often not purely motivated by one reason but by a number of inter-related reasons.

Early efforts to model the migration decision of workers did not take into account the details of the legal environment. They are based on the human-capital model by Sjaastad (1962), which makes the migration decision dependent on the discounted net revenues and the monetary and non-monetary cost of migration. More recent work puts emphasis on the complexity of migration costs (for details see, e.g., Trachtman 2009): They consist, for example, of the direct cost of changing residence, the burden of bureaucratic procedures, the utility loss from abandoning social contacts in the country of origin as well as the time and effort required to establish new contacts in the country of destination, the time and effort required to learn a new language if necessary and to adapt to a different culture, the additional cost of finding an appropriate school for the children, lower pensions for people who have changed their countries of residence during their economically active period more frequently, and so on. At the same time, the benefits may go beyond the remuneration for labor and extend to all kinds of social benefits such as children's allowances, housing subsidies, unemployment relief, social welfare, and free access to schools and universities.

Immigration law affects migration costs by determining who is allowed to enter the country for how long and what are the legal consequences and procedures when people are infringing the law. In a broader sense, immigration law also includes all legal rules that govern access of immigrants to employment, social benefits, and so on.

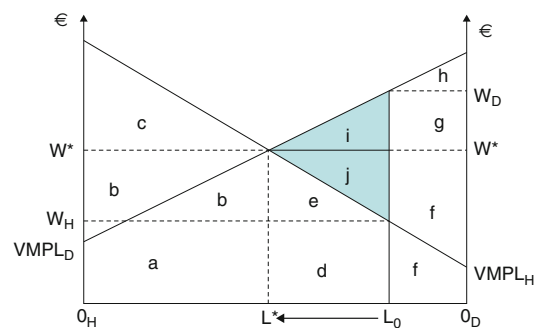
Some Basic Welfare Effects of Migration

Welfare Effects of Labor Migration

Let us start with a very simple model of the migration of workers. (An excellent presentation of the economic analysis of migrant workers

can be found in Borjas (2014). For a recent overview of the economic debate on international migration, see Hatton (2014). Be aware that labor markets are different from goods and capital markets since – different from goods and capital – labor power cannot be separated from the worker. See, e.g., Eger and Weise (1989) with further references.) We assume there are two countries, home country H and country of destination D, there is a given capital stock in each country, capital and goods are completely immobile between the countries, and labor is homogeneous and perfectly mobile between the countries.

Initially, the quantity of homogeneous labor is $O_H L_0$ in country H and $L_0 O_D$ in country D. The curves $VMPL_H$ and $VMPL_D$ represent the value of the marginal product of labor in countries H and D, respectively. With perfect competition on the labor market, the wage rate is equal to the value of the marginal product of labor in each country. Since in the initial situation the wage rate and the marginal product of labor are higher in country D, the reallocation of labor from H to D will produce a gain in allocative efficiency. If there were no migration costs and if migrant workers were induced exclusively by wage differentials, migration would stop only when wages (and the value of the marginal productivity of labor) are the same in both countries. Thus, the total factor income in the receiving country increases ($d + e + i + j$), while that in the sending country declines by a lower amount ($d + e$). The welfare gain for both countries together is $(i + j)$. With positive migration costs, the new equilibrium will be somewhere between L^* and L_0 .



Equilibrium on migration market

However, apart from the overall welfare gain, migration produces even in this simple model winners as well as losers. The winners are migrant labor, gaining $(e + j)$, remaining labor in H, gaining b , and capital owners in D, gaining $(g + i)$, whereas the losers are capital owners in H, losing $(b + e)$, and the existing labor in D, losing g .

Of course, the welfare analysis has to be modified when we relax our strong assumptions. When we allow for heterogeneous labor, it becomes less clear who the winners and losers are. For example, immigrants, who have skills that are complementary to the skill mix of the country of destination, are typically less likely to create losers in this country. On the other hand, the country of origin may suffer losses from emigrating labor force, which used to create positive externalities in that country (so-called brain drain, see, e.g., Sykes 1995; Trachtman 2009; Boeri et al. 2012; Sykes 2013a).

When we allow for inflexible labor markets and unemployment, with homogeneous labor, unemployment will increase in the country of destination and decrease in the country of origin. Without additional assumptions, it is not clear whether unemployment in both countries together will be increasing or decreasing. If immigrants have complementary skills to those of the labor force in the host country, unemployment may decrease in the country of destination (Brücker and Jahn 2011).

When we allow for the mobility of goods, we have to take into account that immigration does not only exert pressure on the wages in the host country but also changes its production and trade structure, for example, toward more labor-intensive production (Hanson and Slaughter 2002). Moreover, it should be mentioned that the mobility of labor is usually accompanied by mobility of capital in the same direction, which is due to the fact that immigration to a country with a fixed stock of capital increases the rate of return on capital (Ottaviano and Peri 2006).

Extending the Basic Model

The provision of labor cannot be looked at in isolation. Based on what we have set out in the previous section regarding individual's reasons to

migrate, we have to modify our simple migration model by getting rid of the assumption that migrants are exclusively motivated by differences in gross wages (corresponding to the values of the marginal products of labor). In a more realistic scenario, migrants will also take into account the costs and benefits of the welfare systems in the countries concerned. Actually, migrants will be induced by *differences in net compensation* in the broadest sense, i.e., gross wages minus all kinds of taxes and contributions to the welfare system plus all kinds of social benefits. In this case, part of the migration could be induced by redistributive motives instead of differences in productivity, with at least two possible negative consequences:

- First of all, *people may be induced to emigrate even though their marginal productivity (and their labor income) is lower in the country of destination than in the country of origin*, if the difference in wages is overcompensated by generous social benefits in the country of destination.
- Secondly, if workers with low skills and low income have strong incentives to migrate to countries with generous *social insurance systems*, these systems may *get under pressure* (Sinn 2003, p. 64).

It depends on the respective immigration law and its enforcement to what extent these consequences will arise. Interestingly, for long periods in history, human migration was not subject to any legal constraints (Trachtman 2009, p. 3). However, we can identify market failures that make legal intervention desirable.

Why and How do Modern States Regulate Immigration?

Preliminary Considerations

There is a strong argument from an allocative efficiency point of view to get rid of all restrictions on economic migration (Chang 1997; Trachtman 2009, p. 33; Sykes 2013a). Some scholars claim that there is, hence, no need for an immigration

policy whatsoever (Hayter 2000). It may be desirable to eliminate all immigration controls.

In order to justify a legal intervention, the economist looks for market failures. Looking at immigration generally, the prevalent view among scholars is that international migration can be accompanied by important nonpecuniary externalities (Sykes 2013a, p. 319). These include, by way of example, the additional burden on the welfare state (welfare migration) or may result from the congestion of certain public facilities, from migration which is motivated by higher returns to crime, the import of infectious diseases, and the possibility that migrants may through voting patterns redistribute resources to themselves. These are, hence, a number of varieties of externalities that require at least some border control but in many cases also some regulation of immigration. From a public choice perspective, there is a simple explanation for the existence of regulating the entry of immigrants. As we have shown in the previous chapter, there are winners and losers of immigration. If the losers are better organized than the winners, the former will lobby for a restrictive immigration policy.

The next question when assessing a legal intervention is the welfare standard one wants to apply. There are two main measures to assess immigration policies: One suggests the use of a global welfare function that weighs the welfare of all persons equally. The second trend focuses on national welfare functions in which only the effects on the host countries are considered, excluding the effects on the immigrants' welfare or the welfare of their home countries (Trebilcock 2003). When it comes to arranging for immigration policy at a global scale, hence, the insight that gains are not symmetrical for the countries makes the success of such efforts more unlikely (Sykes 2013b). Even if welfare may be increased at the international level, migration will lead to winners and losers in different countries but also within the countries. Evaluations, hence, depend upon whose welfare is being looked at.

Instruments of National Immigration Policy

Immigration policies in major destination countries, such as the United States, Canada, and

Australia, are characterized by a large degree of centralization (see for the following in detail Trebilcock 2003). State authorities conduct basic health, criminal record, and national security checks and rely on a quota system, which distinguishes between three primary classes of immigrants: independent applicants, family members, as well as refugees and asylum seekers. For each class and respective subclasses (e.g., workers with different skills), the states determine quotas limiting the number of respective immigrants. Moreover, the governments issue short-term visas for tourists, students, and temporary workers, who might become immigrants in the future. There are two major problems with this centralized approach of regulating immigration:

1. It is not guaranteed that only those applicants are excluded who are expected to create an unacceptable burden on the amenities of the welfare state. When competing for skilled migrants, burdensome procedures can be a large disadvantage. (With respect to the European Union, see Kocharov 2011.)
2. Since the quotas have to be set in advance, the centralized bureaucracy is confronted with the very difficult (if not impossible) task of predicting the future needs of the labor market.

A further question in this regard is the institutional design by which quotas should be implemented. The basic options are *ex ante* or *ex post* screening (Cox and Posner 2007). An *ex post* system evaluating post-entry conduct is said to provide more information and, hence, a more accurate screening than an *ex ante* system on the basis of pre-entry information. Overall, both have a number of strengths and weaknesses. A main concern with the *ex post* system is that immigrants live with the fear of deportation, an uncertainty, which may hamper their integration process to the detriment of the host country.

An alternative approach, which could avoid many of the problems regarding independent applicants and family members and which is to a large degree implemented in the European Union, relies on decentralized agreements between the parties concerned, such as between migrant

workers and employers. In the European Union, various regimes apply. Crucially, migrants' rights differ as to whether the migrant is a European national changing his place of residence from one European member state to another (see [International cooperation on migration](#)) or a so-called third-country national entering EU territory. The positions of the latter, furthermore, differ considerably depending on the country that they come from (Eisele 2014). The European Union's major goal is attracting highly skilled workers from non-European countries. (One policy instrument is the Council Directive 2009/50/EC of 25 May 2009 on the conditions of entry and residence of third-country nationals for the purposes of highly qualified employment ("Blue Card Directive").) Whereas the rights of EU nationals are generally aligned, newly acceding states to the European Union may face different types of immigration policies by the various EU member states during a transition period.

A more decentralized system is capable of discouraging irregular immigration. A decentralized system would induce welfare-enhancing and deter welfare-reducing migration, provided there are safeguards that undermine migrants' possibilities to externalize costs to the sending or the receiving country. With respect to the receiving country, the risk of negative externalities can be reduced by (still) centralized health, criminal record, and security checks as well as by a mandatory insurance which avoids that the immigrants become a burden to the welfare state (Trebilcock 2003, p. 296). It is suggested that opening the border would be desirable if at the same time countries made their social programs inaccessible to non-citizens (Sykes 1995). (In Germany, the Academic Advisory Board at the Federal Ministry of Finance made in 2001 the proposal of "delayed integration," i.e., immigrants should enjoy taxed financed social benefits for a transition period from their country of origin. See Sinn (2003, pp. 80–81).) With respect to the sending countries, the problem of "brain drain" may arise if skilled workers, who generate positive externalities in their countries of origin, emigrate. However, one has to take into account that there may be also benefits to the sending countries, such as

monetary remittances and positive externalities by those migrants, who improve their skills in the host country and return to the country of origin later on (Trebilcock 2003, p. 282; Boeri et al. 2012).

International Cooperation on Migration

There is no doubt that there are many more barriers to international migration than to international trade and investment. Whereas many barriers to trade have been removed by multilateral, regional, and bilateral agreements and foreign direct investment has been fostered by more than 3,000 bilateral investment treaties (BITs), a comparable degree of international cooperation with respect to migration is (still) missing (Hatton 2007; Trachtman 2009; Gordon 2010). What are the reasons for this mismatch?

There is broad consensus that there are considerable gains from freeing up international migration, even though the estimates differ a lot (Rodrik 2002; Hatton 2007, pp. 345–346). For three reasons, non-coordinated national immigration policies will deviate from the globally efficient policy and will typically lead to over-restrictive immigration policies (Sykes 2013a, p. 320): (1) *Terms-of-trade externalities*: A large receiving country, facing an upward-sloping curve of immigrant labor, may have an incentive to restrict immigration in order to lower the price for immigrant labor. As a consequence, foreign workers absorb some of the costs of the migration restriction. (2) *Enforcement externalities*: Sending countries may have no incentive to reveal information on individuals with contagious diseases, with a propensity to commit serious crimes, or with a poor employment record to the receiving countries, even though they have better access to this information. Without cooperation, all receiving countries face high enforcement cost and will respond with restrictive immigration rules. (3) *Externalities among receiving countries*: Since the immigration policy of one country affects the flow of immigrants to neighboring countries, uncoordinated immigration policy will lead to suboptimal results.

However, to induce international cooperation on immigration, it is not sufficient that gains from cooperation exist. Two other requirements have to

be met to induce cooperation in immigration policy: (1) Gains have to be distributed in a way that all cooperating countries or, more precisely, the relevant interest groups in these countries win and (2) cooperation must be self-enforcing, i.e., all parties should expect that deviations from the cooperative agreement will trigger sanctions by the others. Thus, one has to take into account potential obstacles to beneficial cooperation (Sykes 2013a, p. 327): (1) *The one-way problem*: Different from trade, where typically most countries are interested in exporting goods to and importing goods from the other countries, migration is typically a one-way-issue. People migrate from poor countries to rich countries. Thus, negotiating mutually beneficial agreements becomes more complex, since the parties concerned have to rely on “issue linkage,” i.e., in exchange for accepting immigrants from less developed countries, the developed countries have to be compensated in “another currency,” e.g., by granting their exporters or investors access to the markets in the less developed countries. (2) *Migration diversion*: In analogy to trade diversion, bilateral or regional agreements on immigration policy discriminate against immigrants from third countries, which may lower social welfare in the immigration countries. One could argue that migration diversion can be avoided if international cooperation adheres to an equivalent to the most-favored-nations obligations under GATT. However, since migration is not controlled by tariffs but by a variety of complex rules and regulations, this obligation would be much more difficult to enforce.

An Example of Far-Reaching Cooperation: The Free Movement of Persons in the European Union

The European concept of the internal market, as defined in the Treaty, is founded on the four fundamental freedoms: the free movement of goods, services, persons, and capital. The underlying idea is to overcome the fragmented goods and factor markets that used to be a characteristic element of Europe after World War II. The free movement of persons between the member states was originally restricted in several ways: (1) The right of free

movement was restricted to *nationals* of the member states. (2) The right of free movement was restricted to the *economically active population*, i.e., to workers, self-employed persons, as well as providers of services. (3) Free movement was interpreted as a *prohibition of (direct or indirect) discrimination on the grounds of nationality*. From the very beginning, the Treaty has included a number of explicit derogations from the free movement of persons (public policy, public security, and public health derogations) and a public service exception. In the last 50 years, this situation has been changed by secondary legislation, i.e., in particular directives and regulations and by clarifying judgments of the Court of Justice of the European Union: (1) The *link* between the right of free movement and *economic activity* has been *removed*, by granting this right also to tourists, students, and others. With the Treaty of Maastricht (1993), the decoupling of free movement from economic activity culminated in the recognition of the status of “citizen of the Union” for all nationals of the member states (now: Article 21 TFEU). (2) The right of free movement was extended to *family members who do not have the nationality of a member state*. (3) The prohibition of discrimination on the grounds of nationality was extended to a general *prohibition of obstacles* to the free movement of persons (Brücker and Eger 2012, p. 146).

The Citizens’ Rights Directive 2004/38 confirms and extends the right of free movement of European citizens in the following way:

- For *stays of less than 3 months*, the only requirement on Union citizens is that they possess a valid identity document or passport.
- The right of residence for *more than 3 months* remains *subject to certain conditions*: Applicants must be either engaged in an *economic activity* or they must have *sufficient resources and sickness insurance* to ensure that they do not become a burden on the social services of the host member state during their stay.
- *After a five-year period of uninterrupted legal residence*, Union citizens acquire the *right of permanent residence* in the host member state.

Even though the European Union has been removing obstacles to permanent migration between member states to an unprecedented extent, the level of internal migration in the European Union has been modest (see the numbers in Brücker and Eger 2012, p. 158). The reason is that the EU member states, in particular, the “old” 15 members before the accession of Eastern European countries starting in 2004, are characterized by relatively small differences in per capita income levels. On the other hand, language barriers and to some extent also cultural barriers determine migration costs that cannot be removed by legal reforms. But still, there is sufficient evidence that the removal of barriers to the free movement of persons triggered migration, which has contributed to a small but visible increase in the aggregate GDP in the entire European Union (Brücker et al. 2009). Finally, there is no evidence so far that the removal of impediments to the free movements of workers would have triggered a mass inflow of unskilled workers from the East to the West and would have led to the exploitation of the welfare state in the destination countries (Boeri and Monti 2009; Brücker et al. 2009).

Asylum Laws

A rather new field in the economic analysis of the law concerns asylum/refugee law. In order to grant asylum to an asylum seeker, it needs to be determined that he or she is in fact a refugee. The core international agreement in this regard is the 1951 Convention Relating to the Status of Refugees as modified by the 1967 UN Protocol. Importantly, it provides a general definition of “refugee” and stipulates in the so-called non-refoulement clause that a person cannot be forcibly returned to a territory where she may face the risk of persecution. Most countries (and all developed countries) are parties to this Convention. From an economic perspective, international cooperation in asylum matters can be viewed as an agreement among states to supply the global public good of refugee protection (Bubb et al. 2011). While the Convention sets some common ground,

national asylum laws also show differences. Lately, one of the key challenges faced by asylum policies is the need to distinguish between “real” refugees and economic migrants. In economic terms, states face the problem of asymmetric information when assessing an individual’s status. They have a screening problem. Gradually, some countries’ asylum policies have become stricter in an attempt to cope with the challenge of identifying the type of migrant. An economic effect of differing standards in various jurisdictions is that states with stricter policies regarding their admission criteria impose an externality on those countries that have more lenient policies. This, in fact, may induce all countries to raise their standards – a “race to the bottom” (Bubb et al. 2011; Monheim-Helstroffera and Obidzinskib 2010). It is being discussed if deeper integration can alleviate some of the problems resulting from differing standards. On the other hand, transfer systems, according to which wealthy states pay poor states to resettle refugees from other poor states, could create positive externalities on third countries (Bubb et al. 2011).

With a view to the effects of asylum policies, a recent empirical study on the asylum policies in developed countries finds that a tougher asylum policy – regarding entry requirements and conditions in the receiving country – had a deterrent effect on asylum applications (Hatton 2009). It accounted, however, only for one third of the decrease. In looking more closely at the provisions which deter potential asylum seeker, the author finds that the effect is due to those policies that limit access to territory and those that reduce the proportion of successful claims. Provisions, on the other hand, that diminish the socioeconomic conditions of refugees show to have only a low deterrent effect.

Research is also carried out from the perspective of the asylum seeker identifying which migration channel he or she would opt for (Djajić 2014). According to Djajić’s model, under current laws, relatively young, skilled, and wealthy asylum seekers, who have access to credit from the family network, are found to have a strong incentive to choose to enter a country with the aid of human smugglers without proper documentation.

Under the Common European Asylum System, Europe has set up a general legal framework for its asylum policy, gradually lifting up more competences to the EU level. In trying to determine the optimal degree of harmonization of Europe's asylum laws, Monheim-Helstroffera and Obidzinskib (2010) develop a regulatory competition model. With respect to the European Union, the authors take a critical stance on ongoing harmonization efforts: In a regulatory context like the European Union, those jurisdictions closest to the external border (peripheral jurisdictions) have an incentive to choose strict admission criteria for refugees. These states would lose most from harmonization efforts that would decrease their legislative discretion. For an EU-wide policy, the question is whether the benefits from harmonization outweigh the losses of the peripheral jurisdiction. From the refugees' point of view, flexibility is warranted as it increases their chances of being admitted to the European Union. While illustrated for the EU context, the model is also more generally applicable. The fear of a "race to the bottom" in the European Union was empirically rejected as national differences in application procedures continued to exist (up to the year 2010 Toshkov and de Haan 2013). The Common European Asylum System has recently undergone some changes; various directives and regulations have been revised and will enter into force in July 2015. Bottom line is that the procedure in the member states have been more aligned, e.g., procedures to apply for asylum have been approximated by the revised Asylum Procedures Directive (Directive 2013/32/EU 2013). In the revised Reception Conditions Directive, common rules have been adopted on the issue of detention of asylum seekers. (Directive 2013/33/EU of the European Parliament and of the Council of 26 June 2013 laying down standards for the reception of applicants for international protection OJ L 180, 29.6.2013, pp. 96–116.) The revised Dublin Regulation seeks to improve the system for determining the member states responsible for the examination of the asylum application by accelerating procedures and reiterating on the decisive criteria, such as recent possession of visa or

residence permit in a member state and regular or irregular entry to the European Union. (Commission Implementing Regulation (EU) No 118/2014 of 30 January 2014 amending Regulation (EC) No 1560/2003 laying down detailed rules for the application of Council Regulation (EC) No 343/2003 establishing the criteria and mechanisms for determining the member state responsible for examining an asylum application lodged in one of the member states by a third-country national OJ L 039, 8.02.14, pp. 1–43.)

Conclusions

In this entry, we provided some insights into relevant law and economics contributions on immigration laws and policies. In the light of different forms of migration, a focus on labor/economic migration seemed appropriate, as well as a short excerpt on research topics within asylum law. The relevant contributions illustrate how different types of migrants necessarily impact countries' welfare functions differently. Immigration law is a topic, which can never be a national matter only. Hence, cooperation is a must. We have alluded to some challenges to this cooperation on regional (European Union) or even international level, resulting primarily from potential negative externalities. Legal reforms in immigration law are ongoing, and the topic stimulates a lot of potential future research.

References

- Boeri T, Monti P (2009) The impact of labor mobility on public finances and social cohesion. IAB-Deliverable 5
- Boeri T, Brücker H, Docquier F, Rapoport H (eds) (2012) Brain drain and brain gain. The global competition to attract high-skilled migrants. Oxford University Press, Oxford
- Borjas GJ (2014) Immigration economics. Harvard University Press, Cambridge/Mass
- Brücker H, Eger T (2012) The law and economics of the free movement of persons in the European Union. In: Eger T, Schäfer H-B (eds) Research handbook on the economics of European Union law. Edward Elgar, Cheltenham, pp 146–179
- Brücker H, Jahn EH (2011) Migration and wage-setting — reassessing the labor market effects of migration. *Scand J Econ* 113(2):286–317

- Brücker H, Baas T, Beleva I, Bertoli S, Boeri T, Damelang A, Duval L, Hauptmann A, Fihel A, Huber P, Iara A, Ivlevs A, Jahn EJ, Kaczmarczyk P, Landesmann ME, Mackiewicz-Lyziak J, Makovec M, Monti P, Nowotny K, Okolski M, Richter S, Upward R, Vidovic H, Wolf K, Wolfel N, Wright P, Zaiga K, Zylicz A (2009) Labor mobility within the EU in the context of enlargement and the functioning of the transitional arrangements: final report. IAB-final report
- Bubb R, Kremer M, Levine DI (2011) The economics of international refugee law. *J Legal Stud* 40:372–373
- Chang HF (1997) Liberalized immigration as free trade: economic welfare and the optimal immigration policy. *Univ Penn Law Rev* 145(5):1147–1244
- Cox AB, Posner EA (2007) The second-order structure of immigration law. *Stanford Law Rev* 59(4):809–856
- Directive 2013/32/EU (2013) Directive 2013/32/EU of the European Parliament and of the Council of 26 June 2013 on common procedures for granting and withdrawing international protection Off J Eur Union L 180:60–95
- Djajić S (2014) Asylum seeking and irregular migration. *Int Rev Law Econ* 39:83–95
- Eger T, Weise P (1989) Participation and codetermination in a perfect and an imperfect world. In: Nutzinger HG, Backhaus J (eds) Codetermination. Springer, Berlin/Heidelberg/New York
- Eisele K (2014) The external dimension of the EU's migration policy: different legal positions of third-country nationals in the EU: a comparative perspective. Koninklijke Brill, Leiden
- Gordon J (2010) People are not bananas: how immigration differs from trade. *Northwestern Univ Law Rev* 104:1109–1145
- Hanson GH, Slaughter MJ (2002) Labor market adjustment in open economies: evidence from U.S. states. *J Int Econ* 57:3–29
- Hatton TJ (2007) Should we have a WTO for international migration? *Econ Policy* 22:339–383
- Hatton TJ (2009) The rise and fall of asylum: what happened and why? *Econ J* 119:183–213
- Hatton TJ (2014) The economics of international migration: a short history of the debate. *Labor Econ*. <https://doi.org/10.1016/j.labeco.2014.06.006>
- Hayter T (2000) Open borders: the case against immigration controls. Pluto Press, London
- Kocharov A (2011) Regulation that defies gravity – policy, economics and law of legal immigration in Europe. *Eur J Legal Stud* 4(2):9–43
- Monheim-Helstroffera J, Obidzinskib M (2010) Optimal discretion in asylum lawmaking. *Int Rev Law Econ* 30:86–97
- Ottaviano GIP, Peri G (2006) Rethinking the effects of immigration on wages. NBER working paper No. 12497
- Rodrik D (2002) Final Remarks. In: Boeri T, Hanson G, McCormick B (eds) Immigration policy and the welfare system. Oxford University Press, Oxford, pp 314–317
- Sinn HW (2003) The new systems competition. Blackwell, Oxford
- Sjaastad LA (1962) The costs and returns of human migration. *J Polit Econ* 70(4 Suppl):80–93
- Sykes AO (1995) The welfare economics of immigration law: a theoretical survey with an analysis of U.S. policy. In: Schwartz WF (ed) Justice in immigration. Cambridge University Press, Cambridge, pp 158–200
- Sykes AO (2013a) International cooperation on migration: theory and practice. *Univ Chicago Law Rev* 80:315–339
- Sykes AO (2013b) The inaugural Robert A. Kindler professorship of law lecture: when is international law useful. *Int Law Polit* 45:787–814
- Toshkov D, de Haan L (2013) The Europeanization of asylum policy: an assessment of the EU impact on asylum applications and recognitions rates. *J Eur Public Policy* 20(5):661–683
- Trachtman JP (2009) The international law of economic migration: toward the fourth freedom. W. E. Upjohn Inst. for Employment Research, Kalamazoo
- Trebilcock MJ (2003) The law and economics of immigration policy. *Am Law Econ Rev* 5:271–317

Impact Assessment

Marie-Helen Maras

John Jay College of Criminal Justice, City
University of New York, New York, NY, USA

Abstract

Impact assessments help improve the quality of proposed legislation. It enables European Community institutions to determine the economic, social, and/or environmental costs of proposed legislation. It further identifies available options to achieve the objectives of the proposed measure and the positive and negative consequences of each of the options.

Definition

An impact assessment is a tool used by European Community institutions to identify the main policy options available to achieve the objectives of proposed legislation and the intended and unintended costs and benefits of each option. In 2003, the Interinstitutional Agreement on Better Lawmaking set forth basic principles and common objectives for the European Commission, the

European Parliament, and the Council of the European Union for their cooperation during the legislative process and nonlegislative proposals and evaluation of these initiatives (European Parliament, Council and Commission, 2003). In the interest of better lawmaking, decision-makers must have knowledge of the potential economic, social, and environmental impacts (e.g., the internal market, competition, businesses, consumers, employment, human rights, public health, species, or habitats) of initiatives (European Commission, 2009). Since 2003, the European Commission conducted impact assessments to determine this (European Commission, 2006). Specifically, the European Commission conducts impact assessments to determine the economic, social, and environmental consequences of proposed major policy initiatives, including those presented in its Commission's Legislative and Work Programme (or its Annual Policy strategy) and/or non-legislative proposals with potential significant costs (European Parliament, Council and Commission, 2005). Impact assessments identify the main policy options that are available to achieve the objectives of the proposed initiative, and all relevant stakeholders are consulted during the process. These assessments also provide the costs and benefits of a proposed initiative. The intended and unintended impacts of a proposed initiative are evaluated in quantitative, qualitative, and monetary terms; if the quantification of the consequences cannot be completed, an explanation should be provided in the impact assessment.

The cost-benefit analysis seeks to ensure efficiency in the allocation of resources among competing objectives and proposed actions. What is determined in the impact assessment is whether each policy option represents a more efficient allocation of resources (on a cost-benefit basis) than if such resources were put to alternative use, that is, another policy option. Based on these identified costs and benefits, decision-makers can make informed judgments about proposed initiatives by identifying competing policy objectives and choosing best courses of action. For certain initiatives with potential far-reaching consequences, an extended impact assessment is

conducted to provide a more in-depth analysis of the potential impacts of proposals on society, the economy, and the environment.

The impact assessment helps improve the quality of proposals and ensure that they are evidence-based. This process thus assists political decision-making and ensures a comprehensive and transparent approach to the development of legislative and nonlegislative proposals. The Impact Assessment Board, which was created in 2006, is responsible for ensuring the quality of impact assessments (European Commission, 2006). To do so, it evaluates drafts of the European Commission's impact assessments. Once an opinion is issued by the Impact Assessment Board (replaced by the Regulatory Scrutiny Board on July 1, 2015), it accompanies the impact assessment during the European Commission's decision-making on the relevant proposed initiative (European Commission, 2015). The impact assessment is then considered by the European Parliament and the Council of the European Union when evaluating the European Commission's proposals. Pursuant to the 2003 Interinstitutional Agreement on Better Lawmaking, if the European Parliament and the Council of the European Union seek to amend the European Commission's proposal in a substantial manner, they must also conduct an impact assessment, using the European Commission's original impact assessment as a starting point.

References

- European Commission (2006) A strategic review of better regulation in the European Union. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, Brussels. COM 689 final. <http://ec.europa.eu/transparency/regdoc/rep/1/2006/EN/1-2006-689-EN-F1-1.Pdf>
- European Commission (2009) Part III: annex to impact assessment guidelines. http://ec.europa.eu/smart-regulation/impact/commission_guidelines/docs/iag_2009_annex_en.pdf
- European Commission (2015) Decision of the president of the European Commission on the establishment of an independent Regulatory Scrutiny Board. C 3263 final. http://ec.europa.eu/smart-regulation/better_regulation/documents/c_2015_3263_en.pdf

- European Parliament, Council and Commission (2003) Interinstitutional agreement on better lawmaking. OJ C 321. <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l10116>
- European Parliament, Council and Commission (2005) Inter-institutional common approach to impact assessment. http://ec.europa.eu/smart-regulation/impact/ia_in_other/docs/ii_common_approach_to_ia_en.pdf

Impracticability

Hüseyin Can Aksoy
Faculty of Law, Bilkent University, Bilkent,
Ankara, Turkey

Abstract

The principle of *pacta sunt servanda* requires that agreements must be kept. However such rule is not absolute. When performance of a contractual obligation becomes impracticable, i.e., considerably more burdensome (expensive) than originally contemplated –albeit physically possible– due to an unexpected event, this would lead to adaptation of the contract to the changed circumstances or to avoidance of the contract. In the law and economics literature, impracticability has been substantially studied to figure out who should bear the risk of impracticability; and what would be the efficient remedy for such breach of contract.

Synonyms

Change of circumstances; Economic impossibility; Hardship; Imprévision; Lapse of the contract basis; Wegfall der Geschäftsgrundlage

Definition

An unforeseeable change in the circumstances, arising after the formation of the contract due to an external event, which renders the performance considerably more burdensome albeit physically possible.

Introduction

In the mid-1970s, Westinghouse Electric Corporation sold nuclear reactors to electric companies and also agreed to supply uranium at a fixed price of \$8–10 per pound. However, while the company was still obliged to deliver around 70 million pounds of uranium to 27 different buyers, the market price of uranium increased to over \$30 per pound. On September 8, 1975, Westinghouse Corporation claimed that performance of its contractual obligation would result in a loss of \$2 billion and announced that it would not honor fixed price contracts to deliver uranium. Could the Westinghouse Corporation rely on such change in the circumstances and refrain from performing its contractual obligations?

The principle of *pacta sunt servanda* requires that agreements must be kept. Accordingly, non-performance of any contractual obligation constitutes breach and may cause liability of the debtor. However such rule is not absolute. For instance, in almost all legal systems, it is a well-established principle that when performance is rendered impossible by an event of *force majeure* (e.g., war, earthquake, hurricane, etc.), the debtor will not only be released from his obligation to perform, but this will also eliminate his/her liability to compensate the damages of the creditor. However, the issue is more controversial when performance of a contractual obligation becomes impracticable, i.e., considerably more burdensome (expensive) than originally contemplated – albeit physically possible – due to an unexpected event.

On a theoretical basis, the result of impracticability might be one of the following:

- (i) The debtor can be expected to perform despite the increase in costs of performance. Correspondingly, if the debtor fails to perform, he/she has to compensate the damages of the creditor. In such cases, the scope of the recoverable damages (reliance damages or expectation damages) of the creditor must also be determined.
- (ii) Impracticability can be regarded as an event, which excuses the debtor. In such cases, the

creditor can neither ask for performance nor compensation. In return, the creditor does not perform either. However it must still be decided if impracticability excuses the debtor *ipso facto* or it grants the debtor a right to avoid the contract.

- (iii) It can be argued that impracticability does not lead to an absolute excuse or the avoidance of the contract but only to the adaptation of the terms of the contract to the changed circumstances. In other words, either the parties themselves or the court adapts the contract to the changed circumstances and reestablishes the balance between the debtor's obligation and the creditor's counter-obligation.

Impracticability in Modern Legal Doctrine

The modern legal doctrine regards impracticability as an exception to the principle of *pacta sunt servanda* provided that the performance becomes excessively difficult for the debtor due to an external event, which could not be foreseen at the time of the conclusion of the contract, and under the new circumstances the debtor cannot reasonably be expected to perform as stipulated in the contract.

It must be emphasized that the border between impossibility and impracticability is hard to draw. Despite their similarities, these two concepts have important differences as well. Since impossibility is regarded as an objective and permanent obstacle to performance, it causes expiration of the primary obligation of the debtor. This follows from the fact that the debtor cannot be expected to perform what is impossible. On the other hand, in case of impracticability the obligation is – at least theoretically – still possible, however at enormous and unexpectedly high cost. Therefore, in such cases, the primary tendency is to preserve the contract, and it is generally ruled that the contractual balance must be restored through adaptation of the obligations of the parties to the changed circumstances. However, if such adaptation is not possible, the second option is the avoidance of the contract. This would result in

elimination of the obligations of both parties, including compensation liability.

When performance becomes more burdensome – albeit possible – some legal orders, for instance, the German Civil Code, distinguish between two case groups, which lead to different legal consequences. According to §275 II BGB, if there is gross disproportionality between the creditor's interest in performance and the debtor's interest in non-performance (i.e., required efforts and expenses to perform), the debtor may refuse to perform. The classical example of this case is the ring which falls into the sea. In this case, the legal consequence of impracticability is the same as that of impossibility: the debtor's obligation to perform the primary obligation extinguishes. A second group captures those cases, in which the cost of performance rises enormously, just as in the first group, after the conclusion of the contract, but the value of the contract (performance) rises too. In such cases, which fall under §313 BGB, there is no disproportionality, but the circumstances which became the basis of the contract significantly change. For example, an art dealer buys a painting of a famous painter from a gallery which the gallery must still buy on the market. But before the purchase by the gallery, a fire in a museum destroys many paintings of the particular painter, which increases the price of the bought painting by the factor 10. Here the event which makes the performance so expensive increases proportionately the interest of the creditor in performance. In this case, the debtor's obligation to perform the primary obligation would not extinguish but the legal consequence would be adaptation of the contract by raising the price. Revocation would be the secondary option, if adaptation is not possible or one party cannot reasonably be expected to accept adaptation.

The practical difference between the scopes of these provisions is the following: when the value of the performance increases, the creditor's interest in receiving performance increases proportionally as well; hence the performance which has become burdensome for the debtor is not disproportionate to the interest of the creditor. Therefore, an objective increase in the market price of the good is separated from the cases, where the

procurement of the good has become particularly costly for the specific debtor.

Economic Analysis of Impracticability

Impracticability has been substantially studied in the law and economics literature. In fact, economic considerations might play an important role in drawing the line between a bearable difficulty and “excessive difficulty” (impracticability). Moreover, economic reasoning might be useful in determining when the parties can resort to avoidance of the contract instead of adaptation to the changed circumstances.

Risk-Bearing Perspective

In the earlier stages, authors approached the issue from the economic theory of efficient risk bearing and imposed the entire results of impracticability on either of the parties (Posner and Rosenfield 1977; Joskow 1977). Within this stance, in their pioneering study, Posner and Rosenfield argue that a discharge question arises only if the contract has not assigned the risk in question (to either of the parties) and the event, giving rise to the discharge claim was not avoidable by any cost-justified precautions (of the debtor). Provided that these two conditions are met, the authors argue that the loss should be placed on the party who is the superior (the lower-cost) risk bearer. Accordingly, the superior risk bearer can be figured out by three factors: knowledge of the magnitude of the loss, knowledge of the probability that it will occur, and costs of self-insurance or market insurance. For instance, the party, who is in the position to prevent the materialization of the risk or to insure the risk at a lower cost, would be the superior risk bearer. In the end, if the debtor is found to be the superior risk bearer, he/she must perform despite the increase in the costs of performance. However should the creditor be in the position to bear the risk, the debtor would be discharged.

Efficient Remedy Perspective

Posner and Rosenfield’s all-or-nothing approach, which shifted the entire risk to either of the parties, was later on questioned by some authors, who

focused their studies on designing efficient remedies for breach of contract. Instead of deciding who should bear the entire risk, these authors associated impracticability with efficiency and focused on dividing the risk between the contracting parties. Within this context, considering the efficiency of the remedy in a given case, the result of impracticability could vary between expectation damages and no remedy.

With efficiency considerations, it is generally argued that except for some very rare cases, discharge of the debtor would not yield to efficient results (White 1988; Sykes 1990). As a matter of fact, associating impracticability with discharge of a contract will negatively affect the risk bearing of the parties. Moreover the availability of discharge as a remedy will create high breach incentives on the debtor.

Even if it is accepted that impracticability should not lead to discharge of the debtor except for exceptional cases, it is argued that neither expectation damages nor reliance damages must be covered in all cases of impracticability. In fact such decision must be rendered on a case-by-case analysis. Within this stance, some authors propose to focus on the risk aversion of the parties while making a choice between the imposition of expectation damages and reliance damages (Shavell 1980; Polinsky 1983) or even between expectation damages and no damages (Sykes 1990). Others argue that the parties’ expectancies at the time of the conclusion of the contract regarding the change of circumstances will be decisive on the choice between expectation and reliance damages (Eisenberg 2009).

Another group of authors argue that in cases of impracticability, the appropriate remedy would be adjustment of the contractually contemplated price (Speidel 1981; Trakman 1985; Trimarchi 1991). The main arguments of these authors is that in cases of absolute uncertainty (in systematic risks such as inflation or international crises, which affect the economy as a whole), the “superior risk bearer” criterion is inappropriate and insurance is not an adequate option (Trimarchi 1991). Moreover, it is asserted that the price adjustment of the court redresses the advantaged party’s opportunistic conduct as the advantaged

party has the duty to cooperate and bargain in good faith ex-post (Speidel 1981).

Another proposal is that the total surplus of the contract must be taken into consideration, when deciding on the result of impracticability (Aksoy and Schäfer 2012). Within this stance, following the increase in costs of performance, if the contract still generates positive surplus, there is no reason to avoid the contract. In this case, if the risk was very remote and both parties failed to consider the risk, the obligations of the parties can be adapted to the changed circumstances. However if the total surplus is obviously highly negative and if this is observable by a third party like the judge, the contract would be avoided. In this case, in addition to the excessive and low probability cost increase, the increase in relation to the interest of the buyer (consumer surplus) triggers avoidance of the contract.

Finally, it must be emphasized that the necessity of the excuse doctrines is disputed itself. For instance, there are some authors who question the well-recognized assumption that parties cannot allocate risks under uncertainty. According to Triantis (1992), the question is not whether a particular risk is allocated or not, but at what level it is allocated. The author argues that even if all risks are not “explicitly” allocated, they can be allocated under broader risk groups. For instance, an increase in the transportation costs due to increase in oil prices following a nuclear accident in the Middle East cannot be “explicitly” anticipated, but the parties may allocate the broader risk of a large increase in oil prices for any reason. Therefore, allocation of risks must be left to the parties and the role of the contract law should be restricted to interpretation and enforcement of the risk allocations of the parties.

References

- Aksoy HC, Schäfer HB (2012) Economic impossibility in Turkish contract law from the perspective of law and economics. *Eur J Law Econ* 34:105–126
- Eisenberg MA (2009) Impossibility, impracticability, and frustration. *J Legal Anal* 1:207–261
- Joskow PL (1977) Commercial impossibility: the uranium market and the Westinghouse case. *J Legal Stud* 6:119–176
- Polinsky AM (1983) Risk sharing through breach of contract remedies. *J Legal Stud* 12:427–444
- Posner R, Rosenfield A (1977) Impossibility and related doctrines in contract law: an economic analysis. *J Legal Stud* 6:83–118
- Shavell S (1980) Damage measures for breach of contract. *Bell J Econ* 11:466–490
- Speidel RE (1981) Court-imposed price adjustments under long-term supply contracts. *Northwest Univ Law Rev* 76:369–422
- Sykes AO (1990) The doctrine of commercial impracticability in a second best world. *J Legal Stud* 19:43–94
- Trakman LE (1985) Winner take some: loss sharing and commercial impracticability. *Minn Law Rev* 69:471–519
- Triantis GG (1992) Contractual allocations of unknown risks: a critique of the doctrine of commercial impracticability. *Univ Toronto Law J* 42:450–483
- Trimarchi P (1991) Commercial impracticability in contract law: an economic analysis. *Int Rev Law Econ* 11:63–82
- White MJ (1988) Contract breach and contract discharge due to impossibility: a unified theory. *J Legal Stud* 17:353–376
- Further Reading**
- Christopher B (1982) An economic analysis of the impossibility doctrine. *J Legal Stud* 11:311–332
- Goldberg V (1985) Price adjustment in long-term contracts. *Wisconsin Law Rev* 1985:527–543
- Gordley J (2004) Impossibility and changed and unforeseen circumstances. *Am J Comp Law* 52:513–530
- Perloff JM (1981) The effects of breaches of forward contracts due to unanticipated price changes. *J Legal Stud* 10:221–235
- Smythe DJ (2011) Impossibility and impracticability. In: De Geest G (ed) *Contract law and economics*, encyclopedia of law and economics, 2nd edn. Edward Elgar, Cheltenham/Northampton, pp 207–224
- Smythe DJ (2004) Bounded rationality, the doctrine of impracticability, and the governance of relational contracts. *S Calif Interdisc Law J* 13:227–267
- Triantis GG (1992) Contractual allocations of unknown risks: a critique of the doctrine of commercial impracticability. *Univ Toronto Law J* 42:450–483
- Wright AJ (2005) Rendered impracticable: behavioral economics and the impracticability doctrine. *Cardozo Law Rev* 26:2183–2215

Imprévision

► Impracticability

Imprisonment

► Prisons

Incarceration

► Prisons

Incomplete Contracts

Maria Alessandra Rossi
Department of Economics and Statistics,
University of Siena, Siena, Italy

Abstract

The notion of incomplete contracts refers to the circumstance that some aspect of contractual parties' payoff-relevant future behavior or some relevant payoff in future contingencies is unspecified in the contract and/or unverifiable by third parties. This may be attributed, by and large, to three different causes: high enforcement costs entailing unverifiability by third parties such as courts or arbitrators; the transaction costs that arise from uncertainty about future events, from the contractual parties' bounded rationality, and from judges' bounded rationality; and, finally, from asymmetric information. Different research programs in the economics of contracting explore the implications of these different sources of contractual incompleteness, providing insights addressing an extremely wide range of contractual issues, including the theory of the firm, the theory of corporate finance, the analysis of formal and informal institutions, regulation and public ownership, innovation and intellectual property, and international trade. The extent to which the notion of contractual incompleteness also has relevant normative implications for the law and economics of contract regulation is an issue currently debated.

Definition

Contracts can be considered incomplete, from an economic perspective, when some aspect of parties' payoff-relevant future behavior or some relevant payoff in future contingencies is unspecified in the contract and/or unverifiable by third parties. In other words, an incomplete contract is a contract that is insufficiently state-contingent, so that some or all of the parties to the transaction are unsure as to the effective payoffs they will receive in any future state of the world. When the cause of incompleteness is attributed entirely to the fact that some aspects of the transaction are observable to the parties but unverifiable by third parties, an "incomplete contract approach" is said to be adopted.

Introduction

The notion of incomplete contract belongs more to the realm of microeconomic theory than to the law and economics approach *stricto sensu*. The scholarly contributions that have explored the wide-ranging implications of the notion have indeed focused mostly on the positive analysis of contractual parties' behavior with regard to the choice of the contractual form that may best limit the negative effects of incompleteness and on the comparative analysis of the choice of the governance arrangement most apt to complement or substitute for the incomplete contract. Thus, most of the literature that attributes relevance to the notion of contractual incompleteness does not directly address the issue of the effects of legal rules on contractual parties' behavior that is the core concern of the law and economics analysis of contracts. In other words, the literature on incomplete contracts deals mostly with the mechanics of private contracting rather than with contract law. There is, nonetheless, some overlap, to the extent that the economic notion of incomplete contract and the associated research programs provide normative implications for contract interpretation and contract regulation.

This entry provides an overview of the concept, considering first its meaning according to the

different theories that have highlighted its relevance and then its many implications in a wide range of research domains. The insights drawn from the notion of incomplete contract have indeed been fruitfully applied to a wide number of specific issues, including the theory of the firm, the theory of corporate finance, the analysis of formal and informal institutions, regulation and public ownership, innovation and intellectual property, and international trade. The normative implications of contractual incompleteness for the law and economics of contract regulation are also briefly reviewed.

The Meaning of Contractual Incompleteness

A contract may be said to be literally incomplete if it contains a true “gap,” namely, if it does not contain provisions relating to some event or circumstance that may arise in the future (*unforeseen contingency*) and therefore does not define parties’ behavior in such circumstances. While this notion of contractual incompleteness has been adopted, implicitly or explicitly, in a number of law and economics analyses, the expression “incomplete contract” is more commonly understood to refer to the strictly economic version of the concept, which introduced into microeconomic theory the general insight that real-world contracts may significantly diverge from the perfect, fully state-contingent contracts depicted by the theory of perfectly competitive markets.

Thus, incomplete contracts may be defined, from an economic standpoint, as insufficiently state-contingent contracts (see, e.g., Schwartz 1998). The economic literature identifies three possible reasons for their existence, to which correspond, by and large, three research programs in the economics of contracting.

The first reason for contractual incompleteness is given by enforcement costs and is emphasized by the research program inaugurated by Sanford Grossman, Oliver Hart, and John Moore and defined as “incomplete contract theory,” “new property rights approach,” or GHM theory.

According to this perspective, contracts are incomplete because some aspects of the transaction are observable by the parties but not verifiable by a third party in charge of enforcement – a judge or an arbitrator. In other words, the origin of contractual incompleteness should be attributed to a problem of asymmetry of information between the contractual parties and a relevant outsider – a judge or an arbitrator – in charge of enforcing the contract.

The second source of contractual incompleteness – emphasized by new institutional and transaction cost economics – is given by the transaction costs that arise from uncertainty about future events, from the contractual parties’ bounded rationality, which limits their ability to account *ex ante* for all the future contingencies that may affect parties’ contractual payoffs, and from judges’ bounded rationality. The relevant transaction costs may be of at least four main types: (1) the cost of foreseeing all the possible states of the world that may materialize during the contractual relationship; (2) the cost of negotiating and finding an agreement about the appropriate course of action parties should take in any future state of the world; (3) the cost of describing *ex ante* in a contract the characteristics of what is traded and/or the parties’ effort for each possible contingency; and, finally, (4) enforcement costs (Williamson 1985; Hart 2008). According to this perspective, incompleteness may also be endogenous because parties rationally decide to save on contracting costs when transaction costs exceed the corresponding benefit.

A third cause of contractual incompleteness is asymmetric information between contractual parties, explored by principal-agent theory. This perspective entails that contracts are voluntarily left incomplete because more complete contracts would create opportunities for moral hazard or adverse selection. Making payoffs contingent on future states of world that cannot be symmetrically observed by parties may indeed allow the informed party to misrepresent the state of the world that has materialized so as to increase her payoffs. Similarly, in some instances, a complete contract may not be chosen because it reveals to the counterpart valuable private information. In

all of these cases, incompleteness is entirely endogenous as it emerges from the equilibrium choices of players whose actions are not constrained by bounded rationality or asymmetric information with respect to third parties. This third source of contractual incompleteness is worth mentioning for completeness, but it should be noted that the notion of “incomplete contract” is generally evoked by reference to transaction and/or enforcement costs.

Thus, there is some debate on the root causes of contractual incompleteness. The debate goes as far as to question the very existence of a theoretical foundation for the notion of incomplete contracts and to suggest the superiority of complete contracting approaches, related to implementation theory and mechanism design (Schmitz 2001). The main criticism to the notion of incomplete contracts is that the assumption of unverifiability by third parties appears weak if parties are symmetrically informed (Maskin and Tirole 1999). Payoff-relevant information is generally only partly unverifiable and parties may invest resources to increase verifiability. Moreover, symmetrically informed parties may adopt complex revelation mechanisms that can be negotiated ex ante and that ensure that both parties will have incentives to reveal their true preferences ex post, so as to ensure efficient outcomes. It should be noted, however, that these complex contracts are seldom observed in practice.

The Holdup Problem

Contractual incompleteness matters, from an economic standpoint, in so far as it affects the realization of Pareto-efficient exchanges. Most of the incomplete contracting literature focuses on circumstances when this is the case because the contract involves some form of specific investment and at least an opportunistic party (Williamson 1985).

Investments specific to particular assets or relationships are investments whose economic value is much higher within the relationship and provided that the investing party has access to the relevant assets rather than outside the relationship

or in absence of access to the assets. In other words, the ex post value of specific investments outside of the relevant transaction is much lower than their ex ante next best alternatives. Thus, once specific investments have been incurred, the contractual parties become to some extent locked into each other (what Oliver Williamson has famously dubbed the “fundamental transformation” of a standard transaction into a bilateral monopoly).

Agents who make specific investments are therefore exposed to the risk of counterparts’ opportunistic behavior because, to realize the surplus from their investments, they need the cooperation of their contractual counterparts and they require access to a particular set of assets. However, precisely because of this *lock in effect*, contractual counterparts may behave opportunistically and try to renegotiate the terms of the original contract so as to obtain a greater share of the total surplus (this is the substance of what is called the “holdup problem”). Since the level of specific investment is not verifiable ex post, the extent to which an individual will be able to appropriate the surplus from his investment cannot be determined ex ante via the original contract and depends on his ex post bargaining power. The source of inefficiency therefore lies in the fact that the threat of holdup constitutes a deterrent to the ex ante realization of unverifiable specific investments, so that some Pareto-efficient transactions may be inhibited.

Incomplete Contracts and the Theory of the Firm

The first and foremost application of the incomplete contracting framework concerns the theory of ownership and vertical integration developed by the incomplete contracts approach on the basis of insights first proposed by Oliver Williamson. This theory explains the very existence and the boundaries of the firm by emphasizing the important economic function of the allocation of property rights over physical or nonhuman assets in a situation of incomplete contractibility (Grossman and Hart 1986; Hart and Moore

1990). When contracts are incomplete, ownership of physical or nonhuman assets is indeed important because it influences parties' threat points in the ex post bargaining and therefore the division of ex post surplus. This is because "the owner of an asset has residual control rights over that asset: the right to decide all usages of the asset in any way not inconsistent with a prior contract, custom or law" (Hart 1995, p.30). Ownership ensures ex ante the owner that he will be able to dispose ex post of the asset and will not be excluded from its use. The increased bargaining power at the renegotiation stage provides the owner with a greater incentive to invest with respect to non-owners.

The incomplete contracts approach thus conceptualizes the firm as a collection of assets over which the owner has residual rights of control and proposes a theory of optimal ownership allocation according to which ownership rights over non-human assets should be assigned to the agents who value them the most, i.e., to the parties who have to make the most relevant and specific investments in human capital. Assuming that parties may efficiently bargain ex ante on the allocation of ownership rights, the theory also predicts that efficient ownership allocations will tend to emerge in equilibrium (for a survey of contributions developing this approach, see Aghion and Holden, 2011).

The solution envisaged represents, however, only a second-best solution and the theory has the merit of highlighting also the costs of ownership allocation as a mechanism to align incentives. The most relevant of these costs is that the incentive provided through the allocation of residual control operates only with respect to the owner, while incentive problems persist with regard to non-owners. When the number of agents required to make specific investments is high, the gap between the first-best and the second-best solutions will be particularly wide, and the allocation of ownership rights will therefore display a limited efficacy as an incentive mechanism.

The incomplete contracting approach provides a framework for the comparative analysis of governance structures that may explain vertical integration choices and the scope of the firm's boundaries. The theory is, however, subject to

a major criticism: there is some logical tension in assuming that agents may not sign complete contracts specifying required specific investments but that they can perfectly foresee ex post payoffs in order to bargain on the optimal ownership allocation.

The key role played by the allocation of residual rights of control in an incomplete contract framework has been explored by this literature also by a different angle. Rather than focusing on the allocation of ownership as a mechanism to align incentives, a large number of contributions has explored the design of contractual procedures that define the ex post renegotiation framework in a way that, by constraining renegotiation, suitably allocating bargaining power and defining default options allows to overcome the holdup problem (Chung 1991; Aghion et al. 1994). The mechanisms thus defined (*option contracts*) allow to achieve first-best outcomes but require higher degrees of verifiability than GHM-style models because the external enforcer is assumed to be able to recognize who is the opportunistic party (for a survey, see Schmidt 1998). This strand of research seeks to explain contractual behavior rather than firm's organizational choices. The main limit of this approach stressed by critics is that the theory's basic insights are very sensitive to underlying assumptions.

Incomplete Contracts, Transaction Costs, and Institutions

Starting from a somewhat stronger notion of contractual incompleteness than simple observability plus non-verifiability, the new institutional and transaction cost research program focuses on the comparative analysis of the institutional arrangements designed to mitigate the effects of incompleteness (see, e.g., Brousseau and Glachant 2002). Given that agents have limited cognitive resources and face strong (Knightian) uncertainty, according to this perspective, they devise their contractual relationships in ways that allow for the minimization of transaction costs. The solutions that emerge from transaction cost

minimization vary from bilateral contracts to formal and informal institutions, all of which may perform similar functions in terms of ensuring effective enforcement of the contractual relationship. The contractual relationship is conceptualized as a sophisticated object designed to provide safeguards for specific investments, promote parties' commitments, and guarantee performance through the definition of negotiation procedures, supervision mechanisms, and conflict-resolution tools. It implements what this literature calls a "private order" meant to discipline parties' behavior.

The core issue involved by contract design is given by the trade-off between opportunism and efficient adaptation (see, e.g., Nicita and Pagano 2005). Contracts may be long or short, detailed or open-ended. The choice among these features depends on their relative costs and benefits: when contract duration is long, a very detailed contract may be chosen with the purpose of minimizing the risks of opportunism at the cost of incurring higher ex ante transaction costs and lower ex post flexibility. At the same time, the broader and more open-ended the contract, the higher the flexibility allowed to the parties in order to efficiently react and adjust the mechanics of their relationship to unforeseen contingencies. This basic trade-off gives rise to the emergence of contracts with different durations and different degrees of ex ante specification of parties' obligations, according to the nature of the underlying relationship and to the characteristics of the institutional environment.

Institutions play a key role for contractual design because they provide the default rules of the game that parties may choose instead of elaborating more complex bilateral arrangements. Relevant institutions include, of course, public institutions in charge of enforcement such as the legal system and the judiciary but also a wide range of other institutional arrangements that have a private nature, both formal (codes of conduct, business associations, standard-setting organizations, etc.) and informal (customs, reputation, corporate culture, etc.) (North 1990). The nature of the institutional environment influences the choice of contractual terms and the ability of given contracts to implement efficient outcomes.

The logic of transaction cost minimization of the new institutional and transaction cost approach is meant to explain the emergence of different institutional arrangements as efficient responses to given transactional characteristics and features of the institutional environment. However, it does not easily lend itself to explain the persistence of inefficient institutions and private arrangements. The literature on institutional complementarities (Pagano and Rowthorn 1994; Aoki 2001) also starts from the notion of contractual incompleteness but provides an explanation for inefficiencies. It focuses on the existence of multiple equilibria across different choice domains: agents are unable to coordinate their choices across different domains, and therefore, their choices in one domain are influenced by the choices made in other domains, giving rise to a multiplicity of institutional arrangements, each of which constitutes a Nash equilibrium with self-reinforcing properties. Depending on initial conditions, inefficient equilibria may emerge and persist through time, unless exogenous shocks intervene to shift the system toward a different equilibrium.

Incomplete Contracts and Corporate Finance

The incomplete contract notion has shed new light also on the analysis of the determinants of the choice of a firm's financial structure, providing a perspective that highlights for the first time the importance of corporate finance choices for incentives. The incomplete contracts approach to corporate finance allows to overcome some of the limits of the so-called trade-off theory of the optimal capital structure of the firm according to which the optimum debt-equity ratio is chosen by equating the marginal benefit in terms of tax savings from the use of debt (which the tax treatment renders cheaper than equity after tax) with the expected marginal bankruptcy cost. This theory, prominent in corporate finance, does not address the issue of the impact of the capital structure on stakeholders' incentives to make firm-specific investments and therefore on firm performance.

Adopting an incomplete contracts approach allows, by contrast, to trace a link between corporate finance and the generation of a firm's cash flow by analyzing the effect of alternative financial structures on the allocation of control rights. It is when enforcement of both financial contracts and managers' incentive contracts is limited that the allocation of residual rights of control matters (Bolton 2013). This is because financing choices affect the distribution of decision powers in the event of unforeseen contingencies and therefore shape *ex ante* incentives.

This perspective thus provides an economic justification for the use of debt different from tax advantages: debt is a form of financing that ensures a flexible and contingent allocation of control rights between the investor and the entrepreneur (Aghion and Bolton 1992). Recourse to debt implies that control rests in the hand of the entrepreneur in case a good state of the world materializes and that it shifts in the hands of the financier in case of a bad state of the world. This provides safeguards to both parties: to the financier, who is shielded against excessive losses in the bad state, and to the entrepreneur, who is protected from the risk of losing control in the good state.

Incomplete Contracts and Innovation

Both contractual incompleteness and asset specificity appear especially pronounced in the context of innovative activities. Innovation is indeed a collective, cumulative, and highly uncertain process that involves specific investments by a large and varied number of firm stakeholders (financiers, managers, workers, etc.) who contribute financial, physical, and intellectual resources to a common endeavor whose final outcome is often to a large extent unpredictable and impossible to specify in a detailed contract. The risk of underinvestment associated to the holdup problem is thus magnified in this context, as the collective nature of the innovative process multiplies the instances of potential holdup.

Acknowledgement of these features has prompted a number of contributions exploring the

implications of incomplete contracts for firms' ability to innovate. This literature takes as a starting point the basic intuition of the incomplete contract approach, namely, that absent the possibility to write complete contingent contracts, the rules affecting the allocation of residual rights of control over the relevant assets deeply influence stakeholders' incentives to invest. Therefore, this stream of research focuses on the relationship between the institutions, market forces, and internal governance arrangements that jointly define firms' corporate governance structure and innovative performance. In an incomplete contracting framework, corporate governance rules matter because they affect the extent to which financial investors are guaranteed a return for their investments, and therefore the price at which there are willing to provide the firm with the funds necessary to undertake innovative projects, and because they affect the distribution of residual rights of control within the corporation, and therefore stakeholders' incentives to make specific investments in human capital. A rich theoretical and empirical literature has uncovered the many facets of firms' corporate governance that may have a bearing on innovation, including the degree of ownership concentration, owners' identity, firms' capital structure, the extent of workers involvement in firms' decision-making processes, and institutional factors such as the degree of unionization and the rules disciplining takeovers (for a survey, see Belloc 2012). A related strand of research adopts an incomplete contract framework to explore the costs and benefits of the allocation of intellectual property rights as an incentive mechanism (Aghion and Tirole 1994; Pagano and Rossi 2004).

Incomplete Contracts and International Trade

Another domain where it is natural to assume the presence of contractual incompleteness is international trade. International transactions occur in absence of effective enforcement systems, since uncertainties exist as to applicable laws when contracting parties reside in different countries,

existing international conventions and trade fora are limited at best, and implicit contracts based on repeated interactions are scarcely effective in curbing opportunism. The notion of incomplete contracts has indeed been fruitfully applied in open-economy environments to study the issue of firm organization and vertical integration in the international context so as to explain the determinants of multinational activity and the structure of international trade flows. In particular, attention has been devoted to explaining, on the basis of GHM-style arguments, why multinationals tend to integrate capital-intensive productions of intermediate inputs and to outsource labor-intensive productions (Antràs 2003). According to this literature, the choice reflects the allocation of property rights to the party who has to make the most important specific investment.

The other main strand of research exploring the implications of the concept of contractual incompleteness in an open-economy environment aims at explaining countries' comparative advantage in the production of goods requiring relationship-specific investments. Since countries differ in their ability to ensure contract enforcement, and since the effectiveness of enforcement affects incentives to make relationship-specific investments, countries with stronger contract enforcement will have a cost advantage in productions requiring relatively higher degrees of relationship-specific investments. This may explain patterns of international trade as well as firms' geographical location (see, e.g., Nunn 2007).

These two strands of research do not exhaust by any means the landscape of contributions linking incomplete contracts to international trade but nonetheless provide a useful introduction to the nature of the issue addressed. The literature on trade and institutions inspired by the incomplete contracts notion has, indeed, flourished both theoretically and empirically in recent years, providing insights for analyzing trade policy choices, understanding the role of power in international transactions, the financial structure of multinational firms, and many other globally relevant issues (for surveys, see Antràs 2013; Helpman 2006).

Incomplete Contracts and the Law and Economics of Contract Regulation

From a strict law and economics perspective, the economic problem of incomplete contracts matters because it is at the roots of the legal problem of contract regulation, which includes the issue of contract interpretation. Contractual incompleteness implies that there is a role for the State in filling the gaps left by parties by interpreting the contract, supplying a common framework and vocabulary to contracting parties, supplying default rules, and eventually regulating the contracting process through other means. What exactly this role should be has, however, so far not been fully defined by the incomplete contracts literature.

A relevant point of view in this regard holds that the notion of contractual incompleteness should be taken to suggest that the State should refrain from attempting to provide efficient default rules to "complete" contracts because it is highly unlikely to possess more accurate information or to incur lower transaction costs than the parties in drafting the relevant provisions (Hermalin and Katz 1993). Moreover, when contracts are endogenously incomplete, any attempt to provide default rules that modify parties' original contractual arrangement may undermine the latter by modifying the terms of the chosen renegotiation game. A further implication of this line of reasoning is that specific performance should be preferred to any attempt to regulate contracts by supplying default rules or imposing efficient solutions (Schwartz 1998).

A different viewpoint holds that the normative implications of the incomplete contracting literature are limited and inconclusive, mostly because the latter does not appear to successfully predict either contract content or interpret legal doctrines such as the penalty doctrine (Posner 2003).

Conclusion

The notion of incomplete contract has gained wide currency in both microeconomics and law and economics. The concept appears to some extent elusive: while most would agree that

non-verifiability of contractual terms by third parties and insufficient state-contingency of the contract are key elements of the definition, there is some theoretical debate on other relevant aspects and on the very meaning and foundations of the concept. In spite of these debates, however, the concept has proven to be rather versatile, as it underlies, in one form or another, a diverse collection of both theoretical and empirical studies. Starting from the basic application in explaining the size and the boundaries of the firm and in motivating comparative analysis of institutional arrangements, the notion of contractual incompleteness has spurred research in a wide range of more specific domains that has refined our understanding of firms and institutions. The appropriate way in which these insights should be incorporated into the analysis of the effects of legal rules on economic agents' behavior is, however, still somewhat controversial and deserves further study.

Cross-References

- [Governance](#)
- [Institutional Complementarity](#)
- [Institutional Economics](#)
- [Transaction Costs](#)

References

- Aghion P, Bolton P (1992) An incomplete contracts approach to financial contracting. *Rev Econ Stud* 593:473–494
- Aghion P, Holden R (2011) Incomplete contracts and the theory of the firm: what have we learned over the past 25 years? *J Econ Perspect* 252:181–197
- Aghion P, Dewatripont M, Rey P (1994) Renegotiation design with unverifiable information. *Econometrica* 62:257
- Aghion P, Tirole J (1994) The Management of Innovation. *The Quarterly Journal of Economics* Vol. 109(4), pp. 1185–1209
- Anràs P (2003) Firms, contracts, and trade structure. *Q J Econ* 118(4):1375–1418
- Anràs P (2013) Goes global: incomplete contracts, property rights, and the international organization of production. *J Law Econ Organ*. First published online 17 Feb 2013. <https://doi.org/10.1093/jleo/ews023>
- Aoki M (2001) *Toward a comparative institutional analysis*. MIT Press, Boston
- Belloc F (2012) Corporate governance and innovation. *J Econ Surv* 265:835–864
- Bolton P (2013) Corporate finance, incomplete contracts, and corporate control. *J Law Econ Organ*. First published online 9 Oct 2013. <https://doi.org/10.1093/jleo/ewt010>
- Brousseau E, Glachant J-M (2002) The economics of contracts and the renewal of economics. In: Brousseau E, Glachant J-M (eds) *The economics of contracts. Theories and applications*. Cambridge University Press, Cambridge
- Chung T-Y (1991) Incomplete contracts, specific investments and risk sharing. *Rev Econ Stud* 58:1031
- Grossman SJ, Hart OD (1986) The costs and benefits of ownership: a theory of vertical and lateral integration. *J Polit Econ* 944:691–719
- Hart OD (1995) *Firms, contracts, and financial structure*. Oxford University Press, Oxford
- Hart O (2008) Incomplete contracts. In: Durlauf SN, Blume LE (eds) *The new Palgrave dictionary of economics*, 2nd edn. Palgrave Macmillan, London
- Hart OD, Moore J (1990) Property rights and the nature of the firm. *J Polit Econ* 98:1119–1158
- Helpman E (2006) Trade, FDI, and the organization of firms. *J Econ Lit* 44(3):589–630
- Hermalin BE, Katz ML (1993) Judicial modification of contracts between sophisticated parties: a more complete view of incomplete contracts and their breach. *J Law Econ Organ* 9:230
- Maskin E, Tirole J (1999) Unforeseen contingencies, property rights, and incomplete contracts. *Rev Econ Stud* 66:83–114
- Nicita A, Pagano U (2005) Incomplete contracts and institutions. In: Backhaus A (ed) *Elgar companion to law and economics*. Edward Elgar, Cheltenham, pp 145–164
- North DC (1990) *Institutions, institutional change and economic performance*. Cambridge University Press, Cambridge
- Nunn N (2007) Relationship-specificity, incomplete contracts and the pattern of trade. *Q J Econ* 1222:569–600
- Pagano U, Rossi MA (2004) Intellectual property rights, incomplete contracts and institutional complementarities. *Eur J Law Econ* 182:55–76
- Pagano U, Rowthorn R (1994) Ownership, technology and institutional stability. *Struct Change Econ Dyn* 52:221–243
- Posner EA (2003) Economic analysis of contract law after three decades: success or failure? *Yale Law J* 112: 829–880
- Schmidt KM (1998) Contract renegotiation and option contracts. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*. Palgrave Macmillan, London
- Schmitz PW (2001) The hold-up problem and incomplete contracts: a survey of recent topics in contract theory. *Bull Econ Res* 53(1):1–17
- Schwartz A (1998) Incomplete contracts. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*. Macmillan/Stockton Press, London/New York
- Williamson OE (1985) *The economic institutions of capitalism*. The Free Press, New York

Independent Ethics Committee

► [Institutional Review Board](#)

Independent Judiciary

George Tridimas

Department of Accounting, Finance and Economics, Ulster University Business School, Belfast, Northern Ireland

Abstract

After describing the closely related concepts of judicial independence and independent judicial review of policy, this entry offers an overview of four issues: (1) Reasons for establishing an independent judiciary, including its ability to resolve problems of information asymmetry between citizens – principals and public officials – agents, transform constitutional declarations to credible commitments and provide a mechanism of political insurance; (2) mechanisms for appointing judges and the jurisdiction of courts; (3) modeling the role of the judiciary as an additional veto player in games of collective decision-making and policy implementation; and (4) the judiciary as an explanatory variable and its effect on economic variables of interest like economic growth and the size of the government.

Definition

Judicial independence means that courts enforce the law and resolve disputes without regard to the power and preferences of the parties appearing before them (La Porta et al. 2004). Its theoretical antecedents are traced to the Enlightenment, and its application in practice dates to the US Constitution. Judicial independence is an indispensable part of the rule of law. The rule of law requires that laws apply equally to both ordinary citizens and public officials and that they protect the rights of

individuals against the power of the state in both the political and economic spheres. In this respect the rule of law and judicial independence are inextricably linked with liberal democracy. The literature on the topic is enormous and cuts across different disciplines including law, economics, politics, and sociology. It is not possible to do justice to this scholarship in the confines of the present essay; rather its aim is to present a summary of the main issues. First, it considers the rationale of judicial independence and the closely related judicial review. Second, it looks at the institutional arrangements for judicial independence. Third, it considers how independent courts are modeled in the collective choice framework. Fourth, it discusses some evidence on the effects of judicial independence on economic variables of interest. These issues are analytically treated as separate but are best understood in relation to each other.

Rationale for Judicial Independence

Judicial Independence and Related Concepts

An independent judicial authority is necessary to resolve disputes and maintain the rule of law which are prerequisites for the functioning of a market economy and a free society. In general, two parties in dispute may resolve their differences by fighting violently against each other or by asking a third party to arbitrate. Realizing that fighting may result in serious inefficiency (destruction of life and property), they may ask a third party to adjudicate and agree to abide by its ruling. They will only do so, however, if they are reasonably confident that the adjudicating party is a neutral and unbiased referee. Disputes may emerge between private entities (citizens, companies, or other organizations), between private entities and the state which among other cases is always a litigant in cases of economic regulation and criminal acts, and between different state organizations (central government, local authorities, nationalized industries, and other public law bodies). Judges with the power to issue binding rulings must then be shielded from the threat of corruption and intimidation by both private litigants and the arms of the state.

However, the very act of referring a dispute to a mediator generates a new conflict: When the dispute resolver declares a winner and a loser, his legitimacy may be undermined leading to the collapse of the adjudication process and its benefits. The reason is that a ruling which obliges parties to behave in a particular way (take specific actions, pay damages, fines, sentence to prison) creates a two-against-one situation (winner and judge against the loser), which is resented by the loser. In order to overcome such problems, arbitrators base their rulings on generally accepted principles of justice and conduct as expressed in formal laws and informal norms and adopt rhetoric of normative justification. The two-against-one problem is even more pronounced in cases where the state is one of the disputants. A delicate balancing act then must be performed between the need to resolve disputes, protecting the independence of the judge, and ensuring that he is perceived as serving only notions of justice.

Closely related to judicial independence is the function of judicial review of policy, where courts may examine and subsequently ratify or annul laws and policy measures, passed by the legislature and enacted by the executive branch of government, for their compatibility with the constitution or other relevant statutes (like declarations of basic rights), and have been enacted according to the stipulated procedures (Stone Sweet 2002). Similarly to ruling in disputes between private parties, judicial review of policy is meaningless unless the reviewing judge is independent of the government. Two further related concepts are those of judicial activism and judicial discretion. Judicial activism is the propensity of courts to query the decisions of elected officials and range from “nullifying acts of the legislature, to abandoning neutral principles, to deciding cases in a politically ‘liberal’ or ‘conservative’ fashion” (Hanssen 2000, p. 538). Judicial discretion is understood as the degree to which the judiciary can implement rulings without being overruled by one of the other branches of government (Voigt 2008).

The interdisciplinary literature analyzing judicial independence can be divided into two strands. The first includes scholarship that treats a

politically independent judiciary as an endogenous variable and examines the reasons for its establishment and the characteristics and degree of its political independence. The second strand considers courts as an explanatory variable and seeks to understand how it affects other political and economic variables of interest, mainly but not exclusively economic growth.

Reasons for Judicial Delegation

A number of in truth complementary explanations of an independent judiciary and its review powers have been proposed in the literature. For a review complementary to the issues taken up in the present section, the interested reader is referred to Harnay (2005). In all cases the starting point is the constitution. Constitutions, written or unwritten, and other fundamental charters specify the rules by which collective decisions are made and the constraints set upon the government and the citizens. They contain declarations of general principles, procedures, organizational forms, and rights and obligations, which, in the absence of complete information and perfect foresight about future changes in tastes and technology, are rendered as incomplete contracts riddled with problems of interpretation and enforcement. The judiciary is the arm that interprets and enforces the constitution, all ordinary laws, and policy measures, and for this reason judicial independence and judicial review are often analyzed jointly.

Modern scholarship uses the insights of the economic analysis of institutions and game theory to examine the benefits of delegation to courts; see Law (2009) and Tiede (2006). Delegation of decision-making by uninformed principals, like the citizens or their political representatives, to the judiciary, a specialized agent, offers several benefits. (a) It resolves problems of information asymmetry as courts develop the relevant expertise in resolving disputes and interpreting and enforcing the law, which subsequently allows specialization of labor and increases welfare. (b) By taking resolution of constitutional disputes away from partisan politics and handing it to “politically disinterested” judges, judicial independence promotes the long-run interests of citizens.

Politicians are better informed than ordinary citizens and exercise discretionary actions which opens up the opportunity for abuse of power to pursue their own interests at the expense of the rest of the society. Citizens may be protected from such abuses by subjecting politicians to elections, which allows voters to confirm or reject politicians, and by setting up checks and balances, where decision-making is divided between different arms of the government and each arm can block the actions of the rest. An independent judiciary is part of the latter mechanism. As it does not need to pander to short-term shifts in public opinion, it may be trusted to look after the long-run interests of citizens and control politicians. Thus, judicial independence is a mechanism that transforms constitutional declarations to credible commitments. Specifically, citizens wish to protect certain individual rights when the cost suffered by someone who is denied that right is very large relative to the gain obtained by others when the right is denied and when those who grant the right are uncertain whether they will be protected or harmed by that right (Mueller 1991). For example, pronouncements of individual freedoms, property rights, protection of minorities, nondiscriminatory taxation, and the like can be trusted by the citizens, who safe in this knowledge will develop longer time horizons increasing investment, growth, and welfare (Maskin and Tirole 2004). Note that in the credibility rationale of judicial delegation, the independent judiciary protects the interests of citizens against a mighty government and political competition and judicial review are substitutes. This is based on an underlying conflict between on the one hand politicians (who irrespective of partisan ideologies pursue their personal interests against those of the citizens) and on the other hand all the citizens.

(c) Contrary to the credibility view, the political insurance view of an independent judiciary considers the conflict between political groups competing for office and focuses on independent courts as a mechanism of political insurance. This approach builds on the famous thesis of Landes and Posner (1975) that a judiciary independent of the current legislature adds permanence to the distributive gains secured by the original winning

political coalition. In essence the argument runs as follows (Stephenson 2003; Hanssen 2004a, b; Tridimas 2004, 2010). Constitutional judicial review implies that courts may prevent the election winner to implement his favored policy measures if found to violate the constitutional arrangements and the rights of citizens. However, in exchange for this constraint, when the same party is out of office, its opponent may also be prevented from implementing his favored policy. Hence, the losers of the political contest can use the review process as a mechanism to minimize the losses inflicted to them from the measures taken by the electoral winner. When political groups anticipate that they will not win every election and therefore they will be out of power, constitutional judicial review is a useful mechanism to restrain those in control of the government. Constitutional review by an independent judiciary lowers the risks associated with the uncertain outcomes of collective choice. On the above reasoning, one expects judicial review to be more pronounced in politics where political competition is strong and parties alternate in office. In the political insurance framework given the probability to win an election and the differences in the preferences of competing political groups, each political group is better off when an agent decides policy but prefers to delegate policy making to an “ally,” that is, an agent which has preferences similar to its own. See Cooter and Ginsburg (1996) and Hayo and Voigt (2007) for an empirical investigation of the determinants of judicial independence.

Note that delegating thorny issues to the judiciary may offer short-run benefits to politicians who this way shift blame for unpopular decisions to independent agents to escape electoral punishment (Fiorina 1986). However, such delegation to the courts decreases the ability to claim credit for policies with a favorable impact. When the expected gains from shifting the blame exceed the expected losses from foregoing credit, the politician will choose to refer policy making to the judiciary. Shifting of responsibility is easier if the judiciary is perceived by the electorate as independent of the other branches of government. However, the latter view loses its explanatory

power when voters cannot be fooled and recognize the politicians' play.

Institutional Arrangements for Judicial Independence

Judicial review of the acts of government is the most politicized aspect of the behavior of courts. Judicial involvement in the political process and collective choice raises a fundamental question: Decision-making by an independent but unelected judiciary may go against deep-seated notions of majority decision-making and electoral accountability. As soon as discretionary powers are granted to the judiciary, a new principal-agent problem arises: A judiciary which is strong enough to block the legislative majority is also strong enough to pronounce rulings to pursue its preferences. What guarantees are there that the independent judiciary will not pursue its own interests at the expense of the citizens that it is supposed to protect? This is the well-known problem of "who will guard the guards?" going back to the ancient Greek philosopher Plato and the Roman poet Juvenal. Note the reverse dilemma too: accountability of the judiciary to give reasons and explain their actions is hardly controversial. But accountability by holding judges responsible for their decisions may infringe their independence, for it cannot be precluded that measures which aim to strengthen the accountability of an agent may be abused and weaken its independence. The solution of this dilemma depends on the arrangements that in practice balance the demands for judicial independence and accountability of the judiciary; see also Cappelletti (1983) and Shapiro (2002) for the tension between democracy and judicial independence. We divide the arrangements for judicial independence under two broad categories, structure and jurisdiction.

Structure

The political independence of judges increases with the following: (a) The smaller the involvement of the government in the process of their appointment and the larger the legislative majorities needed for their confirmation; independence is

even higher when judges are nominated by the judiciary itself. (b) The longer their term of service; independence is also higher when judges serve for a single term only or do not seek reappointment. (c) The greater their financial autonomy which implies that salaries and budgets cannot be reduced by discretionary acts of the executive. (d) Transparency, the obligation to explain and justify rulings, enhances the independence of judges as it increases publicly available information, influences future courses of action, obliges nonelected judges to fully justify their decisions, and discourages politicians or other interested parties to intervene in judicial outcomes. However, there is no consensus regarding the optimal degree of transparency. For example, although full disclosure has a certain intuitive appeal, keeping secret the voting record of judges may also have some advantages in the specific circumstances of supranational judicial bodies like the European Court of the EU. The court does not disclose how individual judges have voted and, contrary to the US Supreme Court, dissenting opinions are not published. This secrecy protects judges against possible retribution from governments which lost their cases at the court. (e) The more difficult it is to discipline and dismiss judges. (f) The more difficult it is for the government to overturn judicial rulings it does not like. Rulings can be overturned by introducing new legislation or changing the status and power of courts. The latter is less likely when courts are bound to follow legal precedent and when their independence is declared in the constitution which, contrary to ordinary legislation, requires supermajorities to revise.

Jurisdiction

A range of issues are examined here – see Ginsburg (2002) for a detailed discussion of the structure and organization of the judiciary. (a) Whether ordinary courts can exercise judicial review of laws, as in the decentralized US system, or only specialized constitutional courts have such rights, as in the centralized system of the European countries following the model of the Austrian legal theorist H. Kelsen. (b) Whether judicial review is concrete or abstract. Under concrete the constitutionality of a law is checked in a case which is

actually litigated in front of a court, while under abstract a law may be examined without litigation. Related to this issue is whether review is carried out before or after the promulgation of a law. US courts practice concrete *ex post* review, while the French Constitutional Court offers an example of abstract *a priori* review. Abstract and *a priori* review is not based on a real case but on a hypothetical conflict and is conducted with less information about “facts” and as such is more limited but has the advantage that it can eliminate unconstitutional legislation before it actually does any harm. (c) In general, the power of the judiciary as an independent arm rises when individuals are granted more open access to the courts and *ceteris paribus* when courts are allowed to review more legislation, since under these circumstances the hold of the executive on policy making is weaker. Note however that easy access may encourage trivial applications for annulments frustrating the exercise of the will of the majority but also increasing the judicial workload and therefore the cost of the system. (d) The ability of court rulings to “make law.” Although judiciaries do not make laws in the sense that legislatures do, insofar as they interpret legislation, their rulings become a source of law and bind future rulings, their independence is greater than otherwise. Similarly, by annulling those acts and measures that they find incompatible with the constitution and fundamental charters, courts have a form of negative lawmaking power.

Modeling the Behavior of an Independent Judiciary

The behavior of the judiciary in collective decision-making is studied by applying spatial decision models and game theory to the process of policy making. The judiciary is modeled as a rational agent pursuing an objective function defined over one or more policy variables and is pitted against the executive arm and the legislature in a sequential game; see Ferejohn and Weingast (1992), Hanssen (2000), Vanberg (2001), Rogers (2001), Tsebelis (2002), and Stephenson (2004). Introducing the judiciary in the collective choice

game typically adds the highest court as a veto player, which affects the set of feasible alternatives against the status quo. In this policy game the executive and legislative branches are not only interested in the outcomes of their own decisions but also on whether their decisions will trigger the judiciary to move and reverse such decisions.

The literature makes two key assumptions about the preferences of the judges. First, their preferences are based on “deeply internalized” notions of justice, the rule of law, and respect for legal reasoning. Second, judges would like to see their rulings implemented. However, in modeling the utility function of judges, normative objectives are not specified. To a large extent this comes from the generality of the rule of law that eludes more specific normative specification. Application of the rule of law does not necessarily imply that a “good” law is applied. The law may privilege the interests of whomever the lawmakers wish to favor. As Shapiro (2002) put it, “The rule of law requires that the state’s preferences be achieved by general rules rather than by discretionary-arbitrary-treatment of individuals” (p. 166). For example, courts of the apartheid era in South Africa were upholding the law of the land, but to the black population, they could hardly appear as neutral and independent arbiters between the interests of different races.

Empirical Studies of the Economic Effects of Judicial Independence

Application of the rule of law promotes a just society, protects individual rights, and defends citizens against predatory governments, advancing in turn economic goals. Major research breakthroughs were made with the construction of indicators of the independence and the power of the judiciary as represented by supreme courts. Empirical research has found that countries with higher degrees of judicial independence enjoy higher economic performance (Henisz 2000), greater economic and political freedom (La Porta et al. 2004), and a lower share of taxes (Tridimas 2005). Most interestingly, Feld and Voigt (2003) distinguish between *de jure* independence, as

described in legal texts setting up the supreme court of a country, and *de facto* independence which is independence of the supreme court of a country as it is actually implemented in practice, and find that only *de facto* judicial independence is conducive to growth.

Regarding the effect of the method of selecting judges, appointment or election, on judicial outcomes, the literature notes a selection effect (the ideological preferences of judges who are elected may differ from the preferences of judges who are appointed) and an incentive effect (judges seeking reelection are more sensitive to the preferences of the electorate). It is found that appointed judges are more independent than elected ones, since elected judges are more sensitive to electoral considerations and may attach greater weight to the interests of litigants from groups who are presumed to have large electoral power; see Hanssen (1999).

Informative as these findings may be, several open questions remain. In the first instance, there is the perennial problem of reverse causality, that is, richer countries can afford good judicial institutions rather than good judicial institutions leading to higher income. Second, the judiciary is approximated by the highest constitutional court and the latter is treated as a single decision taker. This ignores that in reality the judiciary comprises a hierarchy of lower and higher courts. In addition, even though the constitutional court itself is a collective body comprising several justices, subject to the well-known problems of reaching a collective decision, these problems are assumed away and the court is treated as a single decision taker. Finally, from the viewpoint of policy advice, the creation of legal institutions conducive to economic success requires long gestation periods, which may be of little comfort to a government facing pressing short-run demands for growth-promoting policies. Significantly, the results from reforming the courts of developing economies to be politically independent and introducing statutes incorporating principles and procedures of codes found in advanced western countries have been underwhelming (Carothers 2006). It appears that the same deep-lying factors of institutional failure were left intact and prevented the legal reforms to function as intended.

Cross-References

- ▶ [Constitutional Political Economy](#)
- ▶ [Credibility](#)
- ▶ [Good Faith and Game Theory](#)
- ▶ [Incomplete Contracts](#)
- ▶ [Institutional Economics](#)
- ▶ [Political Economy](#)

References

- Cappelletti M (1983) Who watches the watchmen? A comparative study on judicial responsibility. *Am J Comp Law* 31:1–62
- Carothers T (2006) Promoting the rule of law abroad: in search of knowledge. Carnegie, Washington, DC
- Cooter RD, Ginsburg T (1996) Comparative judicial discretion: an empirical test of economic models. *Int Rev Law Econ* 16:295–313
- Feld PL, Voigt S (2003) Economic growth and judicial independence: cross country evidence using a new set of indicators. *Eur J Polit Econ* 19:497–527
- Ferejohn JA, Weingast BR (1992) A positive theory of statutory interpretation. *Int Rev Law Econ* 12:263–279
- Fiorina M (1986) Legislator uncertainty, legislative control and the delegation of legislative power. *J Law Econ* 2:33–51
- Ginsburg T (2002) Economic analysis and the design of constitutional courts. *Theor Inq Law* 3:49–85
- Hanssen FA (1999) The effect of judicial institutions on uncertainty and the rate litigation: the election versus appointment of State judges. *J Leg Stud* 28:205–232
- Hanssen FA (2000) Independent courts and administrative agencies: an empirical analysis of the States. *J Law Econ Org* 16:534–571
- Hanssen FA (2004a) Is there a politically optimal level of judicial independence? *Am Econ Rev* 94:712–799
- Hanssen FA (2004b) Learning about judicial independence: institutional change in the State courts. *J Leg Stud* 33:431–474
- Harnay S (2005) Judicial independence. In: Backhaus J (ed) *The Elgar companion to law and economics*, 2nd edn. Elgar, Cheltenham, pp 407–423
- Hayo B, Voigt S (2007) Explaining *de facto* judicial independence. *Int Rev Law Econ* 27:269–290
- Henisz W (2000) The institutional environment for economic growth. *Econ Polit* 12:1–31
- La Porta R, Lopez-de-Silanes F, Pop-Eleches C, Shleifer A (2004) Judicial checks and balances. *J Polit Econ* 112:445–470
- Landes W, Posner R (1975) The independent judiciary in an interest-group perspective. *J Law Econ* 18:875–911
- Law DS (2009) A theory of judicial power and judicial review. *Georget Law J* 97:723–801
- Maskin E, Tirole J (2004) The politician and the judge: accountability in government. *Am Econ Rev* 94:1034–1054

- Mueller DC (1991) Constitutional rights. *J Law Econ Organ* 7:313–333
- Rogers JR (2001) Information and judicial review: a signalling game of the legislative-judicial interaction. *Am J Polit Sci* 45:84–99
- Shapiro M (2002) The success of judicial review and democracy. In: Shapiro M, Stone Sweet A (eds) *On law, politics and judicialization*. Oxford University Press, Oxford, pp 149–183
- Stephenson MC (2003) When the devil turns. . . : the political foundations of independent judicial review. *J Leg Stud* 32:59–90
- Stephenson MC (2004) Court of public opinion: government accountability and judicial independence. *J Law Econ Organ* 20:379–399
- Stone Sweet A (2002) Constitutional courts and parliamentary democracy. *West Eur Polit* 25:77–100
- Tiede LB (2006) Judicial independence: often cited, rarely understood. *J Contemp Leg Issues* 15:129–261
- Tridimas G (2004) A political economy perspective of judicial review in the European Union. Judicial appointments rule, accessibility and jurisdiction of the European Court of Justice. *Eur J Law Econ* 18:99–116
- Tridimas G (2005) Judges and Taxes: judicial review, judicial independence and the size of government. *Const Polit Econ* 16:5–30
- Tridimas G (2010) Constitutional judicial review and political insurance. *Eur J Law Econ* 29:81–101
- Tsebelis G (2002) *Veto players: how political institutions work*. Princeton University Press, Princeton
- Vanberg G (2001) Legislative-judicial relations: a game theoretic approach to constitutional review. *Am J Polit Sci* 48:346–361
- Voigt S (2008) The economic effects of judicial accountability: cross-country evidence. *Eur J Law Econ* 25:95–123

Independent Regulatory Authorities

Régis Lanneau

Law School, CRDP, Université de Paris Nanterre, Nanterre, France

Abstract

Independent regulatory agencies are now considered to be the sign of modern economic regulatory systems. They proliferated since the 1980s, and it is believed that they are enhancing the efficiency of regulation. In this entry, I will use law and economics to develop some rationale to explain their diffusion and emergence.

Introduction

Independent regulatory agencies represent one of the key features of modern economic regulation. They have indeed proliferated by a factor three to six between the end of the 1980s and the beginning of the 2000s in OECD countries (Jacobzone 2005; see also Pollitt and Bouckaert (2000), considering that the spread of the new public management doctrine is a key to understand the rise of agencies). Almost all OECD countries now have at least financial regulator, energy regulator, environmental agency, and telecommunication authority. Developing countries are following a similar trend since the beginning of the 1990s. In Latin America for example (but the same trend is observed in Asia), less than 45 authorities existed in 1979 (and mostly regarding financial regulation); in 2002, this number was multiplied by three to reach 138 (Jordana and Levi-Faur 2005). Independent regulators are now established in major infrastructure and economic sectors as well as social and environmental arena.

These (administrative) agencies, established in general by legislative acts, are entrusted with substantial (but variable) regulatory power – from rulemaking to adjudication and sanctions – and granted a certain level of independence (especially regarding the executive branch). This independence materializes, quite often, with fixed terms, limits regarding reappointment, guarantees against removal (a “cause” is required), a staffing structure allowing a significant place to experts (and not politicians), and collegiality at its head. This independence is never absolute since, after all, agencies are agents, acting on behalf of a principal (but since agents, they could of course pursue their own agenda, Moe 1990). Independence cannot then be the absence of accountability (see also Çetin et al. 2016; Maggetti 2012).

This “rise of the non-elected” (Vibert 2007) – and more generally of the regulatory state (Glaeser and Shleifer 2003) – is striking. It would certainly be possible to notice the influence of the European Union regarding the liberalization of certain economic sectors (an independent energy regulator was for example required) or the role of the world bank and its market-oriented

regulatory arrangement (see also Gilardi 2005). Nevertheless, the rationale (and legitimacy) for their emergence and diffusion is often consequentialist and law and economics could then be used to assess critically the validity of these rationale. If independent regulatory agencies exist, it is because they are supposed to lead to more net benefits than if the regulation was in the hands of politicians, judges, or mere dependent regulatory agencies.

In this entry, we will not provide a law and economics explanation for the rise as such of independent regulatory agencies (for development regarding this dimension, see, for example, Glaeser and Shleifer 2003; Law and Kim 2011). These explanations are of course not the only one and cannot explain, as such, the diversity of design, power, and functioning of independent regulatory agencies (see, for example, Eberlein 2000 on the absence in Germany of a sectoral regulator for electricity before the influence of the energy directives of the European Union).

Independent Regulatory Agencies as a Shield Against Interest Groups

Independent regulatory agencies are supposed to reduce the amount of political rent-seeking and its adverse effects regarding the design of policies (Shapiro 1988). Indeed, elected regulators, according to a basic economic logic, are eager to maximize their support function in order to get elected or reelected; they might not then have all the incentives to enact welfare enhancing regulations to “buy” the support of some special interest groups (Stigler 1971; Posner 1974; Peltzman 1976). Reducing or removing the political factors in regulation could then be seen as a relevant strategy. A variation around this theme can be found in Glaeser and Schleifer (2003): for these authors, regulatory agencies emerged to face the problem of large, deep-pocket firms that were able to manipulate the courts. Rent-seeking being considered in this variation at the level of courts.

This logic suffers from three problems:

First, even if the political factor is removed, the members of the regulatory agencies might have

their own agenda which are not compatible with the “public interest” (i.e., obtaining a larger budget, extending their powers) and since they are “independent,” it might be difficult to incentivize them to do the “right” thing, especially since their goals are often varied, mixed, broad, and unclear. The dilemma between independence and accountability is clear at this level. Moreover, from a strictly legal point of view, the possibility to delegate regulatory power delegate (and the extent of the delegation) to “independent” agencies is sometimes unclear (Veljanovski is considering them as “constitutional anomalies,” Veljanovski 1991) and the necessity to coordinate the agency with other institution (e.g., for enforcement).

Second, the risk of capture and rent-seeking is not to be excluded by the mere fact of independence. Indeed, agencies’ members are often experts in a field and as such have had the opportunity of repeated interaction with some of the firms they are regulating. Moreover, when they cannot be reappointed, they could wind-up in a high paying job in one of the companies they were regulating, hence incentivizing them to develop regulations not always inspired by the public interest.

Third, this logic also disregards the fact that independent agencies (and their design) could be the result of rent-seeking activities (because, for example, some industries could consider that it would be easier to pressurize this type of agency).

Take for example the Interstate Commerce Commission (ICC) spawned by the Interstate Commercial Act (ICA) of 1887. It is often considered as the first Independent regulatory agency. For Stigler (1971), railroads were able to limit of interstate trucking through their influence over the ICC. For Mullin (2000), the ICC was captured by the shippers, not the railroads. Indeed, railroads were refused to raise their rates despite the evidence of increasing input costs. Stock market evidence is pointing toward a capture by long-haul railroads at the expense of short-haul railroads (Prager 1989). Even if it is difficult to precisely identify which interest group “captured” the ICC, evidence is not showing that an “independent” regulatory agency could fully solve the problem.

Judicial review of regulatory agencies could reduce the problem, but not entirely solve it, since it is difficult to assess when the margin of discretion has been trespassed. If judicial review exists, the Glaeser and Shleifer understanding of the rise of regulatory agencies should be reinterpreted.

Independent Regulatory Agencies as a Tool to Reduce Decision-Making Costs

The idea behind this logic is simple: since the field which is regulated is perceived as a technical field (high information requirements), only experts are required to ensure an efficient regulation. Politicians or judges are, in some domains, ill-equipped to deal with the complexity and technicity required to design efficient public policies. Taking advantage of agency expertise could then make sense. This logic leads to distinguish between technical fields (choosing the best means to achieve a specific end, providing that there is a technical solution) and political – or nontechnical – fields (choosing the ends or the means when they are requiring more than a mere expertise). Such a distinction is not new but is crucial to understand the rise of independent regulatory agencies. A variation of this line of reasoning will stress the possibility to use these agencies for “blame shifting”: if the domain is unpopular, delegating it to a regulatory agency could be an easy way to avoid the “blame” which could be the results of regulations (Epstein and O’Halloran 1999).

For this distinction to make sense, it is required to have the possibility to evaluate the output of these agencies. Indeed, since the problem is supposed to be technical, deviation should be easy to identify. Nevertheless, in the real world, this is far from being the case. For example, the Treaty on the Functioning of the European Union, in its article 127, states that: “The primary objective of the European System of Central Banks (hereinafter referred to as ‘the ESCB’) shall be to maintain price stability. Without prejudice to the objective of price stability, the ESCB shall support the general economic policies in the Union with a view to contributing to the achievement of the objectives of the Union as laid down

in Article 3 of the Treaty on European Union.” Regarding the first goal, the question of the possibility to use nonconventional tools like the Outright Monetary Transaction Program has been challenged (ECJ case C-62/14). Moreover, it is difficult to identify when a European central bank’s action has been undertaken to support “the general economic policies in the EU.” In other words, when a regulatory agency is assigned more than one mission, it is extremely difficult to assess its efficiency.

Independent Regulatory Agencies as a Way to Ensure Credibility of Long-Term Policy-Commitment

The third logic – temporal inconsistencies and credibility (Majone 1996) – was first pointed regarding monetary policy – and central banks – by Kyland and Prescott in 1977. As they specified, since “economic planning is not a game against nature but, rather, a game against rational economic agents” (Kyland and Prescott 1977), the problem is to ensure the credibility of announced rules (see also Elster 2000; Alesina and Tabellini 1988). If the government retains a discretionary power, it might not stick to a policy that was announced to further other agendas. Knowing this possibility, agent will react to the announced policy depending on its perceived credibility. In other terms, even if a policy has been designed to promote public interest, this policy might be rendered ineffective by rational actors who will anticipate some future move by policy makers. When this problem exists, one solution to ensure credibility would be to delegate the regulatory power to an independent regulatory agency (see also Majone 2001; Dixit 1996) whose purpose will only be to follow some established rules, reducing the probability of instrumental and political use. The rise of independent central banks is the best illustration of this “bootstrapping” logic. A variation of that logic is stressing the possibility to use regulatory agencies to avoid present politicians to bind future politicians and reduce uncertainties (Cukierman et al. 1992 for manipulation regarding the tax system).

Once again, the problem of the possibility to identify clearly the mandate of the institution to assess if it is trespassing its power remains. The possibility of a judicial review could help but not entirely solve the problem.

Conclusion: How to Insure Accountability of Independent Regulatory Agencies?

Ensuring the accountability of these nonelected bodies remains the central question in democratic societies. If it were possible to evaluate the performance of these agencies, accountability would be easy to assess. Nevertheless, this task is rarely undertaken (Gilardi and Maggetti 2010) and probably too difficult to be a viable solution. Of course, some papers are trying to assess the impact of these regulatory agencies (quite often only indirectly, for example, Jakee and Allen 1998) but it appears difficult to disentangle regulation (and their perceived efficiency) from regulatory agencies (and their specific role as an institution of regulation). Moreover, even if some players seemed to have benefited from a regulation enacted by regulatory agencies, it remains difficult to be fully certain that it leads to negative effects (considering the difficulty to identify what would be the best achievable). As Ronald Coase advocated, it is required to compare different institutional framework not in the abstraction of a model but with an eye on what is achievable (Coase 1960).

Independence and accountability could only be ensured through a system of a new separation of power which design remains to be identified (and game theory is thus required, Hägg 1997). Among the leads and good practices, a strict adherence to the rule of due process and a perfect transparency are certainly required, so is a system of appeals of their decisions (see the entry on ► [Regulatory Impact Assessment](#)). Moreover, a special attention should be paid to agencies coordination (at the level of the European Union, a system of energy regulator coordination was mandated considering the specific dimension of the electricity and gas market. It led to the creation of the

European Regulators' Group for Electricity and Gas. Moreover, a dialogue between specialized regulatory agencies and competition authorities is also crucial for efficient regulation). Nevertheless, designing a perfect system is a Herculean task requiring information on the full constraint system (both legal and social); the best we can do is only to reduce the probability of blatantly inefficient regulations.

Cross-References

- [Public Choice: The Virginia School](#)
- [Regulatory Impact Assessment](#)
- [Rent Seeking](#)
- [Separation of Power](#)

References

- Alesina A, Tabellini G (1988) Credibility and politics. *Eur Econ Rev* 32:542–550
- Çetin T, Sobaci MZ, Nargeleçekenler M (2016) Independence and accountability of independent regulatory agencies: the case of Turkey. *Eur J Law Econ* 41(3):601–620
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cukierman A, Edwards S, Tabellini G (1992) Seignorage and political instability. *Am Econ Rev* 82(3):537–555
- Dixit A (1996) The making of economic policy. A transaction-cost politics perspective. The MIT Press, Cambridge, MA
- Eberlein B (2000) Institutional change and continuity in German infrastructure management: the case of electricity reform. *Ger Polit* 9(3):81–104
- Elster J (2000) Ulysses and the Sirens. Studies in the rationality and irrationality. Cambridge University Press, Cambridge
- Epstein D, O'Halloran S (1999) Delegating powers: a transaction cost politics approach to policy making under separation of powers. Cambridge University Press, Cambridge
- Gilardi F (2005) The institutional foundations of regulatory capitalism: the diffusion of independent regulatory agencies in Western Europe. *Ann Am Acad Pol Soc Sci* 598:84–101
- Gilardi F, Maggetti M (2010) The independence of regulatory authorities. In: Levi-Faur D (ed) *Handbook of regulation*. Edward Elgar, Cheltenham, pp 201–214
- Glaeser EL, Shleifer A (2003) The rise of the regulatory state. *J Econ Lit* 41(2):401–442
- Hägg PG (1997) Theories on the economics of regulation: a survey of the literature from a European perspective. *Eur J Law Econ* 4(4):337–370

- Jacobzone S (2005) Independent regulatory authorities in OECD countries: an overview. In: OECD (2005) Designing independent and accountable regulatory authorities for high quality regulation. <https://www.oecd.org/gov/regulatory-policy/35028836.pdf>
- Jakee K, Allen L (1998) Destructive competition or competition destroyed? Regulatory theory and the history of Irish road transportation legislation. *Eur J Law Econ* 5(1):13–50
- Jordana J, Levi-Faur D (2005) The diffusion of regulatory capitalism in Latin America: sectoral and national channels in the making of a new order. *Ann Am Acad Pol Soc Sci* 598:102–124
- Kydland F, Prescott EC (1977) Rules rather than discretion: the inconsistency of optimal plans. *J Polit Econ* 85(1):73–491
- Law M, Kim S (2011) The rise of the American regulatory state: a view from the progressive era. In: Levi-Faur D (ed) *Handbook of regulation*. Edward Elgar, Cheltenham, pp 113–128
- Maggetti M (2012) The media accountability of independent regulatory agencies. *Eur Polit Sci Rev* 4(3):385–408
- Majone G (1996) Temporal consistency and policy credibility: why democracies need non-majoritarian institutions, EUI working paper, RSC no. 96/57, European University Institute, San Domenico di Fiesole
- Majone G (2001) Two logics of delegation. Agency and fiduciary relations in EU governance. *Eur Union Polit* 2(1):103–121
- Moe TM (1990) The politics of structural choice: towards a theory of public bureaucracy. In: Williamson OE (ed) *Organization theory: from Chester Barnard to the present and beyond*. Oxford University Press, New York, pp 116–153
- Mullin WP (2000) Railroad revisionists revisited: stock market evidence from the progressive era. *J Regul Econ* 17(1):25–47
- Peltzman S (1976) Toward a more general theory of regulation. *J Law Econ* 19:211–240
- Pollitt C, Bouckaert G (2000) *Public sector management reform: a comparative analysis*. Oxford University Press, Oxford
- Posner RA (1974) Theories of economic regulation. *Bell J Econ Manag Sci* 5:335–358
- Prager RA (1989) Using stock price data to measure the effects of railroad regulation: the Interstate Commerce Act and the railroad industry. *RAND J Econ* 20(2):280–290
- Shapiro M (1988) *Who guards the guardians? Judicial control of administration*. University of Georgia Press, Athens
- Stigler GJ (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 2:3–21
- Veljanovski C (1991) The regulation game. In: Veljanovski C (ed) *Regulation and the market*. IEA, London
- Vibert F (2007) *The rise of the non-elected, democracy and the new separation of power*. Cambridge University Press, Cambridge

Inference

► Rationality

Informal Economy

► Shadow Economy

Informal Law

► Customary Law

Informal Sector

Ozan Hatipoglu

Department of Economics, Bogazici University, Istanbul, Turkey

Abstract

I review and discuss definitions of informal sector introduced by social scientists over the last half century. I describe how informal institutions, informal markets, and their participants' activities form together the informal sector. I provide insight into the difficulties in defining informal sector from a judicial point of view. I discuss causes and consequences as well as economic costs and benefits of the informal sector. Finally, I provide a brief analysis of the existing econometrics methods in measuring its size.

Synonyms

[Hidden Economy](#); [Irregular Sector](#); [Parallel Economy](#); [Shadow Economy](#); [Underground Economy](#)

Definition

All income generating activities outside the regulatory framework of the state.

Definitions

A Brief History of Definitions

A formal meaning to informal sector (A plethora of adjectives are used to describe the informal sector in common language, e.g., shadow underground, subterranean, hidden, parallel, or irregular.) was not given until International Labor Organization (ILO) first officially introduced the term in its study of urban labor markets in developing countries (Hart 1973; Chaudhuri 2010). It was categorized as part of the urban labor force, which operates outside of formal labor regulations such as wage contracts or social security laws. The main emphasis in this conceptualization was on self-employed urban workers. ILO extended the definition of the term in 2002 to include also unregistered or unprotected labor working in formal sector firms. While this newer definition emphasizes labor markets, where most of the informal activity takes place, it is far from being authoritative or unique in its scope. In fact, a consensus on the definition of the informal sector is difficult to find because its coverage is nested in several branches of social sciences due to the multidimensional activities involved (Labor Economics, Anthropology, Sociology, Psychology, Finance, Macroeconomics, Criminology, and Statistics provide different methodological approaches to the analysis of the informal sector.) (Görxhani 2004).

Early works on informal sector have associated it with the existence of a dual labor market in developing countries (Hart 1973; De Soto 1989; Tokman 1972, 1978) and considered it as synonymous for small and self-employed. These were the key characteristics that distinguished it from formal sector where wage-employment and a well-regulated labor market existed together. This is hardly surprising given the dualistic nature of most of the world economies in the postwar and the postcolonialist era. A developed urban economy existed together with a subsistence economy based on agriculture. The rise of urban industries caused massive rural-urban migration that increased the available labor force in cities but at the same time failed to generate enough employment. The surplus labor was forced to generate its

own means of survival, and thereby it created a secondary urban employment sector that became later also an entry point for fresh immigrants (Mazumdar 1976) (Although wages in the informal sector are less than those in the formal sector in most cases, there is a wide diversity of earnings within the informal sector as ILO's Regional Employment Program, known by its Spanish acronym PREALC, reports (Tokman 1990; Mazumdar 1976; Lemieux et al. 1994)).

The dualist approach employed by researchers in treating informal sector as a separate entity received major criticism in early 1990's because of a buildup in empirical findings documenting close linkages between formal and informal sectors. Portes and Schauffler (1993) point to the existence of microproducers capable of producing with modern technology and capital accumulation, but also argue that this type of informal production is an exception in many developing countries including Latin America. Erase Harris (1990) and Sethuraman and Maldonado (1992) rejects both reject the dualist approach on the grounds that linkages between these two sectors play a greater role in explaining their formation than migration. This new strand of literature also suggests that entry to the informal sector might be due to individual choices, which are influenced by barriers to access such as heavy regulation, high taxes, or social factors.

A distinction between the legal and illegal nature of the informal work wasn't made until last two decades when authors from a variety of disciplines (Castells and Portes 1989; De Soto 1989; Feige 1990; Harding and Jenkins 1989) started to define informal sector commonly as that *all income generating activities outside the regulatory framework of the state*. This definition has now become the most widely used characterization of the informal sector in economics. Chen (2004) argues that during the 1990s, the emergence of the concepts of social capital and social networks led to questioning of the value of the concept of informality "even by its main proponents – and a significant decline in its use" (Klein 1999; Hart 1995; Portes 1994).

Since the beginning of the twenty-first century, the discussion on the definition of informal sector

has moved away from epistemological concerns to more practical issues such as how to transform informal sector concepts into instruments for statistical measurement and public policy purposes. Currently, the ILO terminology explains informality around three main notions: the *informal sector* refers to production and employment in unregistered enterprises; *informal employment* focuses on employment outside of the labor protection regulations of a given society, both in formal and informal firms; and *informal economy* covers all firms, workers, and institutions that operate outside the legal regulatory framework of society and their revenue-creating activities.

Labor Market-Based Definitions

Dual labor market theories (Doeringer and Piore 1971; Saint-Paul 1997) divide the labor market into primary and secondary or informal and illegal sectors. The primary sector consists of regular, wage jobs that are regulated and taxed and come with well-defined contracts and social security benefits. Secondary sector provides jobs with lower wages and less regulation, such as those in the service sector. Jobs in this sector are also sometimes referred to as pink-collar jobs. Informal sector consists of mostly self-employed who cannot access primary or secondary sectors. They work by themselves or create small unofficial business units in which workers are hired and transactions are made off-the-books and payments are cash-only. The final category is the illegal sector and it consists of criminal activities that generate income.

H. De Soto (1989) emphasizes the regulatory framework where the main distinction between informal and formal sector is the legal status of the activities. The legal status of the informal sector activities is a gray area in common law and a distinction is almost always necessary between what is legal and what is not. For example, a production process might be illegal in the sense that it avoids or circumvents work regulations while the output resulting from this process might be not.

Most studies on informal sector exclude criminal activity such as drug-trafficking or human trade from a labor market-based definition of the

informal sector. Unlike the criminal sector, the output in the informal sector is legal. Undeclared production of officially recorded firms is also generally considered as part of the informal sector. However, when the labor market-based definition is extended to include official firms, i.e., irregular sector, legal status of the informal sector is not straightforward. The production process in the irregular sector might be plagued with illegal procedures, such as tax evasion, avoidance of labor safety regulations, and social security fraud. Informal sector, therefore, can also be conceptualized as a market where illegal production of legal goods by unofficial firms takes place.

One can further refine this classification by distinguishing between activities that use monetary transactions versus those that use non-monetary transactions. Illegal informal activities that use monetary transaction include trade with stolen goods, drug dealing. Another common illegal informal activity is money laundering which is defined as the set of actions taken by individuals or organizations to cloak the source of funds that are created by criminal activity (Masciandar, forthcoming). Illegal informal activities that use nonmonetary transactions comprise of barter of drugs and stolen goods, theft for own use, as well as producing drugs for own use. Regular informal activities that use monetary transactions can be categorized into *tax evasion* such as unreported income from self-employment, wages, salaries, and assets from informal work and into *tax avoidance*, such as employee discounts and other benefits.

Study-Based Definitions

Since there is a substantial body of literature on informal sector descriptions, it is customary for researchers to define informal sector in accordance with the particular research question at hand. For example, studies trying to assess the size of the informal sector provide definitions of informal sector that are directly or indirectly measurable even though they might include activities that are nontaxable. Smith (1994) puts one such working definition forward as “market based production of goods and services, whether legal or illegal that escapes detection in the official

estimates of Gross Domestic Product (GDP).” Another instrumental definition with a taxability condition is provided by Schneider (2006) as “all income from monetary or barter transactions of legal goods and services that would be taxed if it were reported to tax authorities.”

Most statistical measurement studies exclude criminal activities because the magnitude of the transactions involved are very difficult to measure, but they are also broad enough to cover the diversity of ways in which the informal sector reveals itself in different countries. Studies concentrating on tax evasion or firm regulation focus on irregular sector rather than informal labor markets and take the informal sector to be a place of shadow activities of official firms.

In similar vein, the definition and the scope of the informal sector might also differ depending on whether the focus of the study is a developing or a developed country. In the former case, informal sector studies generally concentrate on labor-intensive informal work, whereas in the latter researchers have focused mostly on tax evasion and optimal regulation.

Informal Institutions, Informal Markets, Participants, and Activities

Informal institutions are governing arrangements that are created and sanctioned outside the regulatory reach of the state (Chen 2004). They are based on traditional power hierarchies or other socially accepted norms. Informal institutions as informal sector participants are more prevalent in developing countries when compared to developed countries.

Informal *markets* are organizational arenas in which factors of production and goods and services are traded.

Informal activity can refer to various informal dimensions of economic activity in an economic system. One way to describe this broad range of activities is to classify them with respect to market participants undertaking them. Individuals, small- or large-scale corporations, formal and informal institutions, financial and nonfinancial institutions, and governments all can engage in informal

activities in a variety of ways. In developed countries, formal institutions play a larger role in informal sector activities when compared to informal institutions. A corporation, for instance, can engage in informal sector activity by:

- (i) Evading taxes through hiding income from its legal activities
- (ii) Hiring informal workers
- (iii) Avoiding regulations with the aim of tax avoidance or reducing production costs to increase pretax profits
- (iv) Making hidden transactions with smaller informal firms that provide cheaper inputs in the lower end of the supply-chain.
- (v) Engaging in illegal activities such as bribery to evade taxes or to avoid workplace or environmental regulations

Individuals or households can also engage in informal activities. Individuals might work as self-employed without registration, or they might work for an informal employer such as own-account informal employers or informal own-family enterprises. Household nonmarket production is a type of informal activity that is not regulated by the state unless it is done for others. Since own household production does not create income it is generally left out from measurement studies and considered as a separate category outside the formal/informal sector divide.

Official state bodies such as governments, municipalities, or state departments can also engage in shadow activities in a variety of ways. They might extend work to informal employers without official tenders. State officials might accept bribery to facilitate procedures or to keep a blind eye on informal activities whether they are legal or not (Choi and Thum 2004; Dreher and Schneider 2006).

The formal financial institutions also engage in shadow activities in several ways. International Financial Stability Board (FSB) (2014) reports that shadow banking now accounts for a quarter of the global financial system. In many developing countries, the volume of informal finance lead by shadow banking exceeds formal finance in

terms of assets and volume of transactions. In developed countries, prior to the financial crisis of 2008, shadow banking mostly referred to legal structures that are used by official banks to keep complicated and risky securities off their balance sheets. With increased regulation on bank operations, informal banking now has become an independent sector and its definition is now extended as “credit intermediation involving entities and activities outside the regular banking system.”

Where formal banks have access to ample funds but are unable to control the use of credit, informal lenders can prevent nondiligent behavior but often lack the needed capital.

Formal Versus Informal: Causes and Consequences

In economics literature, taxes, social security burden, increased regulation, forced reduction in weekly working time, earlier retirement, unemployment, decline of civic virtue, and loyalty towards public institutions are counted among the main reasons for the existence of the informal sector (Frey and Hannelore 1983). Psychological studies offer factors such as tax morale and perceived fairness of the tax system and micro-sociological studies offer exclusion and other social barriers to entry as main causes of the informal sector.

In empirical studies, researchers have looked at the correlates of informal sector size by controlling other factors. Common significant factors found in these studies reflect the generally accepted aforementioned reasons, but some puzzles still exist (Hatipoglu and Ozbek (2011) report that a high informal sector size is associated with little redistribution and lower taxes especially in developing countries).

Monitoring and enforcement in the irregular sector are costly efforts for the regulator due to the high number of qualified personnel hours required. Stronger enforcement and heavier regulations can result statistically in less number of firms evading taxes, and a higher tax collection rate, but at the same time it may lead incumbent and newcomer firms to engage more in informal

activities thereby increasing informal sector size and reducing tax revenues.

The same argument applies to informal labor markets. When faced with higher taxes, individuals might opt out from formal sector and switch to informal sector to protect their after-tax income. As a result, the formal sector might shrink in size as well as in productivity as more skilled people switch to informal sector. Depending on its workforce's skill distribution, a country might have a large informal sector with plenty of skilled jobs or relatively small informal sector with relatively high overall unemployment. This trade-off is analyzed by Ihrig and Moe (2004) that shows how enforcement policies and tax rates interact to determine the size of the informal sector. They find even small changes in tax rates can significantly affect the size of informal employment. On the contrary, increased enforcement has negligible effect on informal employment. Their results suggest that modest reductions in tax rates combined with modest punishment for tax evasion are more effective than strong enforcement alone. The size of the informal sector and the level of redistribution is the result of the joint effects of distributional factors and how lucrative the informal sector is.

Many authors argue that reducing costs of entry to formal sector can reduce the size of the informal sector and improve labor market performance. Increasing enforcement may also reduce informality but can have negative effects on unemployment and welfare such that there is a trade-off between lower informal employment and higher unemployment rate. However, the trade-off disappears when one prefers policies that aim at reducing the costs of being formal, as opposed to policies that increase the costs of being informal.

Many scholars have emphasized the costs of operating in both sectors (Loayza 1996; De Soto 1989, Tokman 1990; Sarte 2000; Portes et al 1989). In the formal sector, in addition to having to pay direct or indirect taxes, formal firms need to comply with guidelines such as labor safety, environmental protection, consumer safety and quality control, as well as administrative procedures such as registration, and other

bureaucratic paper work. Wage regulations and project delays stemming from contract negotiations with organized unions also increase formal sector's operating costs. Whereas in the informal sector, there are costs to getting caught since the informal firm is penalized if detected. The major cost or lack of benefit associated with the informal sector is, however, that its participants cannot take advantage of state-provided public goods such as the police, the judicial system, as well as of business-related resources such as those provided by official trade organizations.

Working in the formal sector requires a minimum set of skills and entry costs. Guilds and business associations require their members to pay annual fees and comply with regulations. In return, individuals have access to resources and infrastructure that can help to improve their productivity. Depending on the tax burden, the intensity of regulations in the formal sector and wages in the informal sector, formally trained individuals such as doctors and lawyers might also choose to work in the informal sector.

Informal Sector: Costs Versus Benefits

The existence of the informal sector can be thought of as the expression of negative sentiments of individuals who are overburdened by the state and choose the exit option. If, in fact, the informal sector is caused by the burden of regulation, then an increase in informal sector size can cause a vicious cycle by eroding the tax base and thereby causing a further increase in tax rates (Olson 1982). A large informal sector size can create problems for both the voters and the policy-makers in taking the optimal decisions because official indicators on unemployment, income, and consumption become misleading.

Depending on how lucrative the informal sector is and how productive the formal workers are, people may choose to work in the informal sector to accommodate for their low levels of productivity. This in turns leads to less output in the formal sector with little to redistribute. This is consistent with the observation that the existence of a large informal sector coincides with less redistribution

in many developing countries. Hatipoglu and Ozbek (2011) show that informal sector may act as a redistributive mechanism in countries where the democratic rule is skewed towards the rich. Increasing the wages in the informal sector vis-à-vis the earnings in the formal sector leads to a decrease in subsidies and an increase in the informal sector size, which suggests that nontaxation is an indirect form of subsidizing incomes and that informal sector acts as an alternative way of redistribution.

Informal sector can also operate in a way to reduce the burden of taxation further by creating new jobs for the low skilled who otherwise would be dependent on the government. Empirical evidence from ILO and World Bank studies show that most of new jobs in the developing world over the past 15–20 years have been created in the informal sector. The main reasons are rapid growth of the labor force, insufficient formal sector job creation, and lack of social safety nets (Chen 2004; Blunch et al. 2001: 10; ILO 1972). In addition, most of the income created in informal sector is immediately spent in the formal sector; thus increased informal activity can also help the formal sector to grow.

Measuring the Size of the Informal Sector

Measurement of the informal sector size is one of the central themes of the informal sector literature. Correct measurement of the informal sector size is important for a variety of reasons. Firstly, it can identify the exact amount of distortion caused by informal sector activities in major national statistics such as GDP allowing both voters and policymakers to make more informed decisions. Secondly, causes and consequences of the existence of informal activities can be easier to identify and policy prescriptions can be made more efficiently.

Currently, there are a number of alternative estimation methodologies that differ in their approach in tackling the unobservable nature of the informal sector (See Frey and Hannelore (1984), Schneider (2006) and Schneider and

Enste (2000) for a more detailed presentation and critique of the available estimation methods).

Direct Methods

Direct approaches collect data directly from participants either through *microsurveys* or use the available data collected through *tax audits*. Since direct approaches depend on the respondents' willingness to comply it suffers from a statistical measurement error called sample bias. People who chose to respond are less likely to engage in informal activities and if they do, they tend to underreport those activities. This makes estimating the range of informal activities as well as translating it into monetary terms difficult. On the other hand, this approach provides detailed information about the structure of the informal sector that is difficult to obtain using indirect approaches.

In the *survey method*, a sample population thought to represent the actual population of a country is chosen to be respondents. The answers are statistically extended to the whole population. The *tax audit* method, on the other hand, measures the discrepancy between incomes declared for taxation and incomes measured by selective checks. This approach provides more precise estimates of the informal activities that selected respondents undertake, but it also suffers from a sample bias problem, since respondents are not selected randomly but based on their previous tax record and likelihood to evade taxes. Finally, except hidden income, most informal sector activities are poorly captured by this method.

There are two common deficiencies of the aforementioned direct approaches. Firstly, they only provide point estimates in time and do not provide information on how informal sector evolves over time in terms of size. Secondly, they provide also poor estimates of the composition of informal sector.

Indirect Methods

Indirect or *indicator* approaches use aggregate macroeconomic data or other "indicators" to infer about the size of the informal sector. These indicators are thought to capture the traces left by the informal sector activities in national statistics. One such trace is the *discrepancy between the*

official and the actual labor force. A decline in labor force participation is an indicator for increased activity in the informal sector. The main disadvantage of this method is that changes to labor participation rates might be due to other reasons than increased informal activity. Another indicator of the informal activity is the discrepancy *between GNP measures based on expenditure and income*. Since GNP based on expenditure and income approach should be equal, any discrepancy can signal the size of the informal sector. National statisticians are likely to minimize this discrepancy because they are caused by measurement error.

Currency demand approach assumes that informal sector transactions are mostly handled in cash, and therefore excess money holdings above what money demand regressions would predict can capture the size of the informal sector. It is one of the most commonly used methods in OECD countries. Since not all transactions are paid in cash, this method might underestimate the size of the informal sector. Another weakness of this method is that the rise in currency demand might be due to the slowdown in demand deposits instead of a rise in informal activity. This method also assumes the same velocity for money in both sectors, which might not be true (See Tanzi (1983) for an introduction and Thomas (1999) and Schneider (2002) for criticisms of the method. Giles (1999) and Bhattacharya (1999) address some of those criticisms.).

Physical input method assumes all kinds of production use a general-purpose, measurable input such as electricity. Since economic activity and electricity consumption are highly correlated in data, one can infer about the size of the informal sector by comparing the measured GDP and imputed GDP calculated using the measured inputs. The main criticism against this method is that some informal activities do not use electricity at all, e.g., personal services and many activities use energy sources other than electricity, e.g., oil and natural gas. Finally, the efficiency of electricity usage might change over time such that estimations over time become inconsistent.

Transactions method uses Fisher's quantity equation that relates the volume of transactions

to the velocity of money, prices, and money supply Feige (1986). Official GDP is subtracted from the total volume of transactions implied by the equation. Here, assumptions have to be made about the velocity of money being constant in both sectors as well as the existence of a base year with no informal sector, both of which might fail to hold.

Econometric Modeling Methods

This approach treats the informal sector as an unobservable variable and uses structural equation modeling to control for the causal relations between the informal sector size and its causes (burden of taxation, regulation, etc.) as well as the informal sector size and its effects (money demand, labor force participation rate, etc.). Schneider (2004), for example, uses dynamic multiple indicators multiple causes (DYMIMIC) method to estimate the size of the shadow economies around the world. This method provides only relative estimates; therefore results have to be combined with absolute measures from currency approach to compute absolute measures for all countries.

Informal Sector Sizes in the World

There are many studies which estimate the size of the informal sector for large sets of countries using aforementioned methods. A detailed survey is beyond the scope of this article. Interested readers can check Zilberfarb (1986), Lacko (1996), Tanzi (1999), Thomas (1999), Schneider and Enste (2000), Chatterjee et al. (2006), and Schneider (2009) among others in chronological order.

Conclusion and Future Directions

Works on the informal sector have concentrated on its definition, measurement, causes and consequences, as well as its links to other spheres in law and economics. Latest economic realities such as free trade and market friendly reforms seem to have contributed significantly to the growing prominence of the informal sector in economic

activities. The developmental history of third world countries as well as boom-bust cycles of the developed world have radically changed perceptions about the relationship between the formal and informal sectors and their roles in economic development. Researchers from a wide variety of backgrounds continue to explore links between the informal sector and other fields such as law, trade, finance, political economy, growth, and income distribution.

Within this realm, the role of informal finance has significantly grown during the last half-decade and especially after the financial crisis of 2008. One can expect to see more studies on the role of regulations in the financial industry in shadow banking as well as in other types of informal finance. Further contributions aside mainstream economics can be expected from network analysis, new institutional economics, micro-finance, legal pluralism studies, as well as Foucauldian studies on political economy.

References

- Bhattacharya DK (1999) On the economic rationale of estimating the hidden economy. *Econ J* 109:348–359
- Blunch N-H, Canagarajah S, Raju D (2001) The informal sector revisited: a synthesis across space and time. Social Protection Discussion Paper 0119. World Bank, Washington, DC
- Castells M, Portes A (1989) World underneath: the origins, dynamics, and effects of the informal economy. The informal economy: studies in advanced and less developed countries, 12
- Chatterjee S, Chaudhuri K, Schneider F (2006) The size and development of the Indian shadow economy and a comparison with other 18 Asian countries: an empirical investigation. *J Dev Econ* 80(2):428–443
- Chen M (2004) Rethinking the informal economy: linkages with the formal economy and the formal regulatory environment. Paper presented at the EGDI-WIDR Conference unleashing human potential: linking the informal and formal sectors, Helsinki
- Choi J, Thum M (2004) Corruption and the shadow economy. *Int Econ Rev* 12(4):308–342
- Doeringer PB, Piore MJ (1971) Internal labor markets and manpower analysis. Heath, Lexington
- Dreher A, Schneider F (2006) Corruption and shadow economy: an empirical analysis. Discussion Paper, Department of Economics, University of Linz
- Feige EL (1986) A re-examination of the “Underground Economy” in the United States. *IMF Staff Pap* 33(4):768–781

- Feige EL (1990) Defining and estimating underground and informal economies: the new institutional economics approach. *World Dev* 18(7):989–1002
- Frey BS, Hannelore W (1983) Bureaucracy and the shadow economy: a macro-approach. In: Hanusch H (ed) *Anatomy of government deficiencies*. Springer, Berlin, pp 89–109
- Frey BS, Hannelore W (1984) The hidden economy as an “unobserved” variable. *Eur Econ Rev* 26(1):33–53
- Gërkhani K (2004) The informal sector in developed and less developed countries: a literature survey. *Public Choice* 120(3–4):267–300, 09
- Giles D (1999) Measuring the hidden economy: implications for econometric modelling. *Econ J* 109(456): 370–380
- Harding P, Jenkins R (1989) *The myth of the hidden economy: towards a new understanding of informal economic activity*. Open University Press, Milton Keynes
- Hart K (1973) Informal income opportunities and urban employment in Ghana. *J Mod Afr Stud* 11(1):61–89
- Hart K (1995) L'entreprise africaine et l'économie informelle. *Reflexions autobiographiques*. In: Ellis S, Faure Y-A (eds) *Entreprises et entrepreneurs africains*. Karthala/ORSTOM, Paris
- Hatipoglu O, Ozbek G (2011) On the political economy of the informal sector and income redistribution. *Eur J Law Econ* 32(1):69–87
- Ihrig J, Moe K (2004) Lurking in the shadows: the informal sector and government policy. *J Dev Econ* 73:541–557
- International Labor Organization (ILO) (1972) *Employment, income and equality: a strategy for increasing productivity in Kenya*. International Labor Organization (ILO), Geneva
- Klein A (1999) The barracuda's tale: trawlers, the informal sector and a state of classificatory disorder off the Nigerian coast. *Africa* 69.04:555–574
- Lacko M (1996) Hidden economy in east European countries in international comparison. International Institute for applied systems analysis working paper
- Lemieux T, Fortin B, Frechette P (1994) The effect of taxes on labor supply in the underground economy. *Am Econ Rev* 84:231–254
- Loayza NV (1996) The economics of the informal sector: a simple model and some empirical evidence from Latin America. *Carn-Roch Conf Ser Public Policy* 45: 129–162
- Maldonado C, Sethuraman SV (1992) Technological capability in the Informal sector: metal manufacturing in developing countries, vol 22. International Labour Organization
- Mazumdar D (1976) The urban informal sector, world development. *Elsevier* 4(8):655–679
- Olson M (1982) *The rise and decline of nations*. Yale University Press, New Haven – Business & Economics – 273 p
- Portes A (1994) The informal economy and its paradoxes. In: Smelser NJ, Swedberg R (eds) *The handbook of economic sociology*. Princeton University Press, Princeton, pp 426–447
- Portes A, Castells M, Benton LA (eds) (1989) *The informal economy: studies in advanced and less developed Countries*. Johns Hopkins University Press, Baltimore
- Portes A, Schauffler R (1993) Competing perspectives on the Latin American informal sector. *Popul Dev Rev* 19:33–60
- Saint-Paul G (1997) Economic integration, factor mobility, and wage convergence. *Int Tax Public Finance* Springer 4(3):291–306
- Sarte PD (2000) Informality and rent-seeking bureaucracies in a model long-run growth. *J Monet Econ* 46:173–197
- Schneider F (2002) Size and measurement of the informal economy in 110 countries around the world. Working paper. Department of Economics, Johannes Kepler University of Linz
- Schneider F (2004) The size of the shadow economies of 145 countries all over the world: first results over the period 1999 to 2003. EZA discussion paper 1431
- Schneider F (2005) Shadow economies around the world: what do we really know? *Eur J Polit Econ* 21(3): 598–642
- Schneider F (2006) Shadow economies of 145 countries all over the world: what do we really know? Johannes Kepler University of Linz, Manuscript
- Schneider F (2009) Size and development of the shadow economy in Germany, Austria and Other OECD-countries. Some preliminary findings. *Revue économique Presses de Sciences-Po* 60(5):1079–1116
- Schneider F, Enste D (2000) Shadow economies: size, causes, and consequences. *J Econ Lit* 38(1):77–114
- Soto H (1989) *The other path: the invisible revolution in the Third World*. Harper Row, New York
- Smith P (1994) Assessing the size of the underground economy: the Canadian statistical perspectives. *Canadian Economic Observer* 3(11-010):16–33
- Tanzi V (1983) The underground economy in the United States: annual estimates, 1930–80. *Staff Papers, International Monetary Fund* 30:283–305
- Tanzi V (1999) Uses and abuses of estimates of the underground economy. *Econ J* 109(456):338–340
- Thomas JJ (1999) Quantifying the black economy: ‘Measurement without theory’ yet again? *Econ J* 109(456): 381–389
- Tokman V (ed) (1972) *Beyond regulation: the informal economy in Latin America*. Lynne Rienner Publishers, Boulder
- Tokman V (1978) An exploration into the nature of the informal-formal sector relationship. *World Dev* 6(9/10):1065–1075
- Tokman VE (1990) The informal sector in Latin America: fifteen years later. The informal sector revisited. OECD, Paris
- Zilberfarb B-Z (1986) Estimates of the underground economy in the United States, 1930–80. *IMF-Staff Papers*, 33/4, pp 790–798

Further Reading

Discussions on further aspects of the informal sector can be found in the following publications:

- Bajada C, Schneider F (2005) *Size, causes and consequences of the underground economy: an international perspective*. Ashgate Publishing Company, Aldershot
- Chaudhuri S (2010) *Revisiting the informal sector: a general equilibrium approach*. Springer, New York/London
- Thomas JJ (1992) *Informal economic activity*, LSE, *handbooks in economics*. Harvester Wheatsheaf, London

Information Deficiencies in Contract Enforcement

Ann-Sophie Vandenberghe
 Rotterdam Institute of Law and Economics
 (RILE), Erasmus School of Law, Erasmus
 University Rotterdam, Rotterdam, The
 Netherlands

Abstract

While contracts are often useful devices for achieving commitment, they can be imperfect devices for doing so when contract breach is unverifiable by third parties or unobservable by the parties themselves. This contribution focuses on the law and economics literature which explains particular features of contract law on the basis of problems of non-verifiability and non-observability. An example is the legal system's use of weaker or no sanctions for contract breach of specific types of contracts, like employment and marriage contracts. It also includes the use of the non-verifiability problem for the evaluation of the desirability of particular legal duties, such as the duty to renegotiate contracts when circumstances change unexpectedly.

Synonyms

Non-observability; Non-verifiability

Introduction

A contract is an exchange of goods and services. However, an exchange does not necessarily

require a contract. A contract is a legal instrument that puts legal pressure on the actions that parties have agreed to take at various times. Parties want their contracts to be enforced by courts to avoid the danger of opportunistic behavior. The danger of opportunism arises when performance of contractual obligations is nonsimultaneous. For then, in the absence of legal enforcement (and assuming pure self-interested behavior, no repeat play, no reputation sanction, and no taste for fairness), the last performing party has an incentive to opportunistically withhold or change his performance obligation. The problem of opportunistic behavior can be solved by drafting a contract that specifies the actions that parties are supposed to take at various times and that makes the nonperforming party subject to legal sanctions. If a party breaches the contract without a good excuse, the legal system has several choices. It can either oblige her to perform the contract as agreed (specific performance) or oblige her to pay compensation instead (damages). For example, a buyer of goods or services will often write down in a contract, sometimes in great detail, the seller's obligations. When the seller fails to satisfy these obligations, he breaches the contracts. The seller sues for damages. If the court decides that the seller did not satisfy his obligations, it will award damages. While contracts are often useful for achieving commitment, they can be imperfect devices for doing so for several reasons (Hermalin et al. 2007). First, the use of contracts to assure commitment is difficult, when due to uncertainty parties lack the information to specify the terms of their exchange in advance. This problem has been the focus of the relational contract literature (Goetz and Scott 1981). A legal requirement to force parties to specify their contracts in more detail in advance is considered not to be a good solution, as it may lead to a party's bankruptcy. Vertical integration and third-party governance have been proposed as possible solutions (Williamson 1975; Klein et al. 1978). Second, when the damages from breach are relatively low compared to the enforcement costs, enforcement may be incredible. Damage multipliers have been proposed as a possible solution in case the probability of a suit is low (Craswell 1996).

But even if legal commitment has been established and the means for its enforcement are available, a contract may be an imperfect method for assuring commitment (i) when breach of a contractual obligation cannot be verified by a third party, like a judge (non-verifiability problem), and (ii) when breach is unobservable by the parties themselves (non-observability problem). There is a standard distinction in economic theory between information that parties can observe and information that is verifiable by a third party. The distinction is drawn because the costs of proving to a third party (e.g., courts) that a particular state of the world existed or a particular action was taken can exceed the gains.

Unverifiable Breach

Even if a contract party can determine that there has been a breach, she may be unable to demonstrate that fact to a third-party enforcer at reasonable cost. For example, an employer may notice that an employee who promised to work hard is shirking his duties. Still, it may be difficult for an employer to prove in court that performance was substandard. The employer cannot act on observable but non-verifiable evidence. In many cases the most important element in the success of a business is cooperative effort, which depends heavily on attitude and morale. Yet these are often the most difficult elements to explain to an outsider (Epstein 1995). The problem of non-verifiability is particularly pressing when a debtor does not commit to achieve a specific result (“obligation de résultat”), but instead commits to provide his best efforts to realize a result (“an obligation de moyens”). In the latter case, the duties undertaken are directed toward a result, but the debtor is only forced to deploy certain methods, thereby meeting the required standard of behavior, such as best efforts. Creditors often have a hard time to prove with verifiable facts that a debtor did not use his best efforts. The simple fact that the result was not achieved is not a useful proxy for low effort, since the bad result may also be due to bad luck. For example, if a lawyer is hired to work in the best interest of his client, it

may be difficult to show that the quality of his pleadings is low – the fact that he lost the case may be due to factors beyond his control. The contract itself may sometimes require a debtor to prove with verifiable facts that sufficient steps were taken with a view to achieving a desirable result, but this system cannot capture all appropriate actions. Eric Posner (2000) points out that the use of expectation damages for breach of contract is problematic in cases in which the debtor’s ability to perform is dependent on non-verifiable actions of the other contracting party. A contract which forces the debtor to pay expectation damages in case she fails to perform gives good incentives to the debtor to perform the contract. But suppose that the creditor is expected to take actions which make it more likely that the debtor is able to perform, but these actions are non-verifiable. Expectation damages would then give the wrong incentives because they would reward the creditor who fails to provide best efforts.

Parties sometimes do contract on the basis of observable but unverifiable information. Such contracts are, in the economic literature, said to be “implicit” or self-enforcing. They help parties to coordinate their affairs and are performed as long as coordination is mutually beneficial (or as long as reputational concerns compel compliance). An implicit self-enforcing contract is one where opportunistic behavior is prevented by the threat of termination rather than by the threat of litigation (Klein 1980). It is left to the judgment of parties concerned to determine whether or not there has been a violation of the agreement. No third party intervenes to determine whether a violation has taken place or to estimate the damages that result from such violation. If one party violates the terms, then the only recourse of the other party is to terminate the agreement after he discovers the violation. Both parties continue to adhere to an agreement if and only if each gains more from adherence to, rather than violation of, its terms (Telser 1980). People keep promises because the prospective gains from doing so exceed the prospective losses. Still, the circumstances under which the threat of termination would be a sufficient deterrent are quite severe:

the value of the contract to both parties must either be expected to continue indefinitely or have a substantial positive probability of continuing in all future states (Rosen 1984). Otherwise, there are well-known tendencies toward opportunism as the end-period approaches, and the contract is no longer self-enforcing. This problem has been discussed in economics under the name of the “last-period” problem.

Unobservable Breach

Breach is unobservable when the beneficiary of a contractual duty is unable to determine whether the promise has been kept or broken. For example, an employer may be unable to determine whether an employee has kept his promise to treat customers in a friendly manner when the employee works at a distant location. Output-based pay instead of a fixed wage may be a way to give incentives to increase the quality of the input, but is not a viable solution in case the agent is risk averse. The challenge is then to design a “mixed contract” so that there are incentives to perform well, but without burdening the agent with too much risk. The multitasking problem is a further complication of output-based pay (Holmstrom and Milgrom 1991). If payment is made dependent on the performance of one task, the agent’s incentives to perform well with respect to other tasks would be severely impeded. Given the incentives and risk problems with output-based pay, many employees receive fixed wages instead. With monitoring being less than perfect, employees are to some extent free to decide how hard they will work. What motivates them to work hard when wages are fixed? What motivates them to do more than the observable and verifiable minimum standard of performance, like showing up for work during working hours? Intrinsic motivation, altruism, and an agent’s identification with the principal’s goals become important in this context. De Geest et al. (2001) have pointed out that the use of expectation damages for contract breach is problematic in cases in which intrinsic motivation is important for contract performance, since it may destroy such incentives.

Why the Legal System May Sometimes Use Weak or No Sanctions for Contract Breach

In general contract law, the standard damage measure for contract breach is the expectation measure. This measure awards compensation for both the reliance expenses (“negative contract interest”) and the expected profits (the “positive contract interest”). The threat of having to pay expectation damages in case of contract breach is a rather heavy dose of legal pressure on contractual obligations. Still, law and economics scholarship considers it as the optimal sanction because it assures that breach only occurs when it is efficient. But for specific types of contracts, modern legal systems adopt much weaker or no sanctions for contract breach. According to the US employment-at-will doctrine, an employment relationship without a definite duration can be freely terminated. Both the employer and the employee have the right to abrogate the relationship at any time, for any reason, without notice or compensation. Parties have no ground to challenge the termination in court. The implication is that employment is “voluntary,” meaning that parties perform only because they want it, not because otherwise they would be sanctioned for contract breach. Of course the absence of *legal* pressure to cooperate does not mean the absence of any pressure. *Informal* pressure may exist to trade, for example, because a party would otherwise lose a deferred benefit. Many European dismissal laws equally put little legal pressure on employment parties to cooperate with each other. In Europe, many employment relationships can be terminated without having to show just cause, provided that reasonable notice is given. Except in the limited cases where there is an abuse of rights or bad faith termination, parties have no ground to challenge the termination in court. An employee who quits early does not have to compensate the employer for the loss of expectancy, but only has to pay an amount of money in lieu of notice. Why is not more legal pressure put on employment parties to stay in their relationship? Dari-Mattiacci and De Geest (2005) explain the legal system’s use of weak sanction on the basis of

the problem of non-verifiability. In order for the legal system to put legal pressure on employment parties to perform, courts would have to find out in case of employment termination which one of the parties is responsible for the failure of the employment relationship and subject that party to legal sanctions. But courts have substantial problems in verifying who was (more) responsible for the failed employment relationship. Did the employee perform poorly? Did the employer promise a more interesting and more challenging job than he offered in reality? Was one of the parties constantly unfriendly and perhaps even responsible for an unpleasant working atmosphere? These facts may be very difficult for a third party to figure out. Of course courts could easily verify which one of the parties took the initiative for the separation and sanction the initiative taker accordingly. However, if sanctions were made to depend on whether the separation is a quit or a layoff, then the distinction can quickly become blurred (Milgrom and Roberts 1992). An employer can often make an employee's life so miserable at work that he or she just has to quit, or an employee can misbehave so badly that the employer sees no choice but to fire the offender, and yet third parties cannot tell who is blamed. Such a sanctioning system will induce parties to play a "you-quit-first game." Each party will try to push the other one to quit first in order to avoid being the one who gets sanctioned. What is the outcome of this game? Who has the best chance to win a you-quit-first game?: (a) the party with the highest ability to generate negative externalities (destroying work or making life unpleasant) and (b) the party who performs poorly. If party A keeps his promises but party B breaches, staying in the relationship is least attractive for party A. As a consequence, and somewhat paradoxically, the non-breacher is most likely to quit first. In order to avoid that the "wrong" party is sanctioned for the unsuccessful employment relationship, legal systems should be reluctant to sanction contract breach when there is a low degree of verifiability. If third parties don't know who the true breacher is, it may be better not to sanction anyone in order to avoid the risk that the innocent party is the one that effectively gets the sanction. Dari-Mattiacci and De Geest (2005)

also apply their framework to divorce laws, which regulate the sanctions in case of marriage contract termination. Here too the courts have difficulties in finding out which party could have done more to prevent marriage failure. Parties to a marriage contract promise to love and respect each other, but finding out which party failed first to provide best efforts to keep the obligation is very difficult. Instead, courts could use easily verifiable proxies, such as who was the party who first quit the house or started a new relationship, and sanction that party accordingly. But again this would lead to parties playing a you-quit-first game, and – as predicted by game theory – it is not necessarily the true breacher who quits first. Therefore, it may be better for the legal system not to sanction contract breach in case of divorce. A no-fault divorce system corresponds with this insight. Under no-fault divorce, neither spouse is required to prove "fault" or marital misconduct on the part of the other to obtain a divorce, and marital misconduct cannot be used to achieve a division of property favorable to the "innocent" spouse.

Should There Be a Duty to Renegotiate Contracts in Case of Unexpected Circumstances?

The 2009 Draft Common Frame of Reference (DCFR) is a European model code envisioned as a collection of "best solutions" for definitions, terminology, and substantive rules in European private law. DCFR III – 1:110 (3)(d) states that if the performance of a contractual obligation becomes so onerous because of an exceptional change of circumstances that it would be manifestly unjust to hold the debtor to the obligation, a court may vary the obligation or terminate the obligation, but only if the debtor has attempted, reasonably and in good faith, to achieve by negotiation a reasonable and equitable adjustment of the terms regulating the obligation. It has been seriously doubted whether there should be a legal duty to negotiate in good faith (De Geest 2010). There is no exact standard available to define the conduct that is required by the parties, and for courts it may be very hard to find out which party negotiated in good faith.

Because of this “non-verifiability,” courts may tend to look at signals, i.e., easily observable facts that are believed to be associated with the negative behavior they want to discourage. Two such observable facts are the amount of time spent to negotiating and which party stopped negotiating first. Yet these signals may unintentionally create a you-quit-first game. Each party may try to make the other party leave the negotiation table first, so that the other one is held responsible for the failure of the negotiations. A legal duty to renegotiate is likely to cause delay and strategic behavior. Moreover, a legal duty is not necessary since contracting parties have private incentives to renegotiate their contract because in that way, they can save litigation costs.

References

- Craswell R (1996) Damage multipliers in market relationships. *J Leg Stud* 25:463
- Dari-Mattiacci G, De Geest G (2005) The filtering effect of sharing rules. *J Leg Stud* 34:207–237
- De Geest G (2010) Specific performance, damages and unforeseen contingencies in the draft common frame of reference. In: Larouche P, Chirico F (eds) *Economic analysis of the DCFR*. Sellier, Munich, pp 123–132
- De Geest G, Siegers J, Vandenberghe A (2001) The expectation measure, labour contracts, and the incentive to work hard. *Int Rev Law Econ* 21:1–21
- Epstein R (1995) *Simple rules for a complex world*. Harvard University Press, Cambridge, MA
- Goetz C, Scott R (1981) Principles of relational contracts. *Va Law Rev* 67:1089–1150
- Hermalin B, Katz A, Craswell R (2007) Contract law. In: Polinsky A, Shavell S (eds) *Handbook of law and economics*, vol 1. Elsevier, Amsterdam, pp 3–138
- Holmstrom B, Milgrom P (1991) Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *J Law Econ Org* 7:24–52
- Klein B (1980) Transaction costs determinants of “unfair” contractual arrangements. *Am Econ Rev Pap Proc* 70:356–362
- Klein B, Crawford R, Alchian A (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J Law Econ* 21:298–326
- Milgrom P, Roberts J (1992) *Economics, organization and management*. Prentice-Hall, Englewood Cliffs, New Jersey
- Posner E (2000) *Agency models in law and economics*. University of Chicago Law School, John M. Olin law and economics working paper no. 92: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=204872
- Rosen S (1984) Commentary: in defense of the contract at will. *Univ Chicago Law Rev* 51:983–987
- Telser L (1980) The theory of self-enforcing agreements. *J Bus* 53:27–44
- Williamson O (1975) *Markets and hierarchies: analysis and antitrust implications*. Free Press, New York

Information Disclosure

Mika Pajarinen

The Research Institute of the Finnish Economy (ETLA), Helsinki, Finland

Abstract

This essay discusses information disclosure, i.e., making company related information accessible to interested parties, in regard to economic literature. The two main fields in which information disclosure is dealt with in this context are corporate finance and innovations. Short notes on other contexts in economics are also made. The essay excludes the descriptions of formal procedures and legal aspects of disclosing information.

Definition

To make (company related) information accessible to interested parties.

Information Disclosure in Corporate Finance

In corporate finance literature information disclosure relates closely to investment and growth opportunities. Firms whose financing needs exceed their internal resources may suffer from financial market imperfections due to asymmetric information and incentive problems between corporate insiders and outside investors (see, e.g., Berger and Udell 1998; Hubbard 1998; Petersen and Rajan 1994). The informational opacity of a firm may reduce the availability of external finance to the firm because the more opaque the firm the more scope there is for opportunistic

behavior by the firm's insiders (more moral hazard) and the harder it is for investors to determine the quality of the firm (more adverse selection). This may lead to higher interest rates demanded by investors and higher risk investment projects chosen by a firm than in a situation in which all relevant information on firm's financial status has been made accessible to potential investors. A conventional wisdom in the contemporary corporate finance literature argues the informational opacity of firms is in relation to firms' age and size. Recent entrants with short track record suffer more from the informational opacity than incumbent firms. Smaller firms are also typically more opaque than larger firms because their minimum requirements for information disclosure, e.g., in financial statements, are more scant than in larger, especially in the listed, firms.

Firms can reduce their informational opacity by voluntarily disclosing high quality information on their business activities over and above mandated disclosure. High quality disclosure is especially important for firms with lucrative growth prospects because for them standard disclosure is of too low quality. It is of too low quality in the sense that mandated disclosure often alleviates information asymmetry only to a limited extent. This kind of higher quality disclosure incurs, however, costs to firms. There can be many types of costs. Direct costs arise from producing and credibly disclosing information. In addition to more comprehensive accounting processes, this may include, for instance, acquiring prestigious auditors and the premiums charged by them. Furthermore, indirect costs can arise from various reasons, such as the pro-competitive effects of disclosure (Bhattacharya and Chiesa 1995; Healy and Palepu 2001). Disclosure may also reduce the incentives of outside investors to acquire information (Boot and Thakor 2001). Firms have thus to balance with the costs of high quality disclosure with returns of growth opportunities. Titman and Trueman (1986) show in a formal model that the firms that choose high quality disclosure are, in equilibrium, those with favorable information about the firm's future and its growth opportunities.

Information disclosure in corporate finance is related also to the interactions of the insiders of

firms, i.e., corporate governance system. Hermalin and Weisbach (2012) argue that more transparent disclosure policies can in fact aggravate agency problems between managers and shareholders and increase firms' costs due to higher rates of executives' turnover and compensation demanded by them. Efforts to be more transparent can lead to a situation where managers are more reluctant to increase firm value in the long run but rather boost short-term investments and other actions that affect reported figures sooner at the expense of longer term (riskier) value creation investments, such as R&D.

Information Disclosure and Innovations

Besides corporate finance, information disclosure relates in economic literature also to generation of innovations. When a firm has made an innovation, it can basically choose to keep the innovation secret or apply for a patent. If the firm applies for the patent, it discloses the discovery of invention and provides information about it to the public. In addition, the patent applicant is obliged to reveal all prior art that may be relevant to the patentability of the applicant's invention. This includes disclosing existing technological information on both documentary sources, such as patents and publications, and nondocumentary sources such as things known or used publicly, which is used to determine if an invention is novel and involves an inventive step (for further details on patenting process, see, e.g., European Patent Office's www-pages <http://www.epo.org>).

When a firm is granted a patent on an invention, it gets a temporary monopoly right in exchange for disclosure. The monopoly right lasts normally 10 years. After the period of patent protection, the invention is freely available to competitors and other potential users. Other parties than the inventor can also utilize the patented innovation during the life of the patent by licensing and other arrangement facilitating a market for technological exchange. Social costs of disclosing the invention by patenting arise if this technological exchange does not perform well. In this case, patented inventions may hold up subsequent research on related inventions and

may generate substantial transaction costs from costly legal challenges about possible infringement (see Gallini 2007 for a more detailed discussion of the role of disclosure in the case of patenting and Griliches 1990 for a review of patents as innovation and economic indicators).

Information Disclosure in Other Contexts

Besides corporate finance and innovations, the term information disclosure is used for instance in consumer economics in regard to how transparent information firms are willing to reveal on their products to consumers (see, e.g., Ghosh and Galbreth 2013; Polinsky and Shavell 2012).

Furthermore, in the field of competition analysis, information disclosure is used in the context of its influence on competition in the field of interest (see, e.g., Arya and Mittendorf 2013; Feltham et al. 1992; Hayes and Lundholm 1996). More detailed information disclosure reveals more data to competitors and can potentially weaken disclosing firm's position in the market. On the other hand, an incumbent firm may strategically choose to disclose some information to prevent new entries in the market. In addition to competing firms, individual firm's disclosures may have spillover effects also in non-competing firms by revealing information on technological trends, governance arrangements, best policy practices, etc.

Cross-References

- [Auditing](#)
- [Externalities](#)
- [Innovation](#)

References

- Arya A, Mittendorf B (2013) Discretionary disclosure in the presence of dual distribution channels. *J Account Econ* 55:168–182
- Bhattacharya S, Chiesa G (1995) Proprietary information, financial intermediation, and research incentives. *J Financ Intermed* 4:328–357

- Berger A, Udell G (1998) The economics of small business finance: the roles of private equity and debt markets in the financial growth cycle. *J Bank Financ* 22:613–673
- Boot A, Thakor A (2001) The many faces of information disclosure. *Rev Financ Stud* 14:1021–1057
- Feltham G, Gigler F, Hughes J (1992) The effects of line-of-business reporting on competition in oligopoly settings. *Contemp Account Res* 9:1–23
- Gallini N (2007) The economics of patents: lessons from recent U.S. patent reform. In: *Economics of intellectual property law*, volume 1. R. P. Merges, Elgar Reference Collection, 16. Elgar, Cheltenham/Northampton, pp 63–86
- Ghosh B, Galbreth M (2013) The impact of consumer attentiveness and search costs on firm quality disclosure: a competitive analysis. *Manag Sci* 59:2604–2621
- Griliches Z (1990) Patent statistics as economic indicators: a survey. *J Econ Lit* 28:1661–1707
- Hayes R, Lundholm R (1996) Segment reporting to the capital market in the presence of a competitor. *J Account Res* 34:261–279
- Healy P, Palepu K (2001) Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. *J Account Econ* 31:405–440
- Hermalin B, Weisbach M (2012) Information Disclosure and Corporate Governance. *J Financ* 67:195–233
- Hubbart G (1998) Capital-market imperfections and investment. *J Econ Lit* 36:193–225
- Petersen M, Rajan R (1994) The benefits of firm-creditor relationships: evidence from small business data. *J Financ* 49:3–37
- Polinsky A, Shavell S (2012) Mandatory versus voluntary disclosure of product risks. *J Law, Econ Org* 28:360–379
- Titman S, Trueman B (1986) Information quality and the valuation of new issues. *J Account Econ* 8:159–172

Information Privacy

- [Privacy](#)

Innovation

Tanja Benedict
InnovationLab GmbH, Legal Services and IP
Management, Heidelberg, Germany

Synonyms

[Creativity](#); [Novelty](#)

Definition

The process of devising a new idea or thing or improving an existing idea or thing (Sandefur 2008). Innovation is also defined as “the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations.” It differentiates between four types of innovations, namely, “product Innovation,” “process Innovation,” “marketing Innovation,” and “organizational Innovation” (OSLO Manual, OECD 2007).

The Term Innovation

Meaning of the Term

The introduction of something new; a new idea, device, or method (Latin: innovatio: renovation, replacement, change, novelty). Innovation in a general sense is understood as renovation and redesign of divisions and parts of economy, science or society, in specific functioning models or behavior patterns in relation to an existing system (economically or socially). Innovation is understood to be an improvement of existing procedures, materials, or devices or a better match to changed requirements of functionality. Innovation has to be distinguished from invention and diffusion. Invention focuses on the mental process of creating a new idea. The emphasis is lying in the moment of creation, the birth of something, that is brought into the material world. In contrast to the term invention, innovation requires also the aspect of the new idea being applied or performed in a way that matters within the context of the innovation. An innovation has to introduce relevant changes on the applied level. Not every invention is necessarily an innovation. In comparison to sustaining innovations, innovations with a high relevance of change are called *disruptive innovations* (see “Clayton M. Christensen: Disruptive Innovation”). Diffusion describes the following process of spreading and transferring the applied invention in the market. Historically the term has been used in French since the thirteenth century in its general

meaning, in English it was recorded since the fifteenth century. Up to the twentieth century, the term had a specific meaning merely in botanic sciences and legal procedures. After World War II, the term has spread through reception of macro-economic theories on the international level.

Innovation as a Term of Economics

Innovation as a term of economics has been introduced by the Austrian-American economist *Joseph Schumpeter* in his *Theory of innovations* first published in 1912. Innovation as a concept of implementation of new products, processes, or management ideas is strongly connected to Schumpeter’s idea of creative destruction. Creative destruction occurs, when innovations make existing products and services obsolete, thus freeing resources to be employed and used elsewhere. So, creative destruction leads to greater efficiency. The creative entrepreneur profits from return of investment by temporary monopoly through innovation instead of profiting from exploitation of price differences. According to Schumpeter, the process of innovation consists of three parts: invention (the creation of the new idea), the innovation itself (implementation of the invention), and the market diffusion (production and sale). A product or process is called innovative only if there is market diffusion, that makes the difference between the mere invention and the innovation.

Innovation as a Legal Term

Innovation is not regarded a technical legal term. Until 2014 there has not been a legal definition of innovation, although the frequency of the term in statutory European and national law is steadily increasing. European Directive 2014/24/EU on public procurement defines innovation in Art. 2 (22) as “the implementation of a new or significantly improved product, service or process, including but not limited to production, building or construction processes, a new marketing method, or a new organizational method in business practices, workplace organization or external relations inter alia with the purpose of helping to solve societal challenges or to support the Europe 2020 strategy for smart, sustainable and inclusive growth;”

The term innovation has been introduced by the European Commission already in 2010 within the programmatic term *Innovation Union*, which is part of the strategy Europe 2020. Innovation in this sense means more than being innovative. “The EU initiative [Innovation Union](#) focuses Europe’s efforts – and its cooperation with non-EU countries – on the big challenges of our time: energy, food security, climate change and our ageing population. It uses public sector intervention to stimulate the private sector and remove bottlenecks which prevent ideas from reaching the market – including lack of finance, fragmented research systems and markets, under-use of public procurement for innovation and slow standard-setting.” (Communication of the Commission [2010](#)) Here innovation is more of an action program following specific political aims. Whereas innovation as an economic concept describes the economic success of innovative products through efficiency in some relation.

Innovation as a Term in Social Sciences

What Schumpeter developed for economics, Friedrich August von Hayek developed in the political theory of the “Open Society,” where he stressed the importance of innovation for social values. Hayek believes in the existence of an order in society. “It would be no exaggeration to say that social theory begins with - and has an object only because of - the discovery that there exist orderly structures which are the product of the action of many men but are not the result of human design” (Hayek [1944](#)). He assumes that planned orders are inferior and will not produce prosperity. Still prosperity requires some kind of regulation by man, as the rule of law, to prevent innovation being steered in a parasitic direction. “Natural selection operates on mutations, making the path of natural selection unpredictable, regardless of how well we understand the underlying principles” (Hayek [1944](#)). To Hayek, social and cultural evolutions are much the same: driven by innovation, fashion, and various shocks that “mutate” people’s plans in unpredictable ways with unpredictable results. The system may be more or less logical. And by no means predictable.

Concept of Innovation in Economics

Joseph Schumpeter: Creative Destruction

Schumpeter defines innovation as a process of changes focused on the creation of something new. The combination of various economic factors in a new, “innovative” way has an impact on the economy and the society on the whole. According to Schumpeter, the process of technological change has three levels: invention (the creation of a new idea), innovation (setting up the elements for the implementation of the invention), and market diffusion. Innovation has an overall positive connotation. But as every human activity innovation has costs as well as benefits. Instead of bringing relieve, creating wealth and power, innovations can also disrupt the status quo. Schumpeter calls the process of creative destruction the process by which innovation causes resources to be set free and therefore introduced the term creative destruction. Creative destruction occurs when innovation makes production processes obsolete and frees resources that can be employed elsewhere, thus increasing efficiency. Creative destruction has always been feared as source of unemployment. The term creative destruction is sometimes also known as Schumpeter’s gale, which Schumpeter derived from Karl Marx and integrated it into his theory of economic innovation. According to Schumpeter, the “gale of creative destruction” describes the “process of industrial mutation that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one” (Schumpeter [1939](#)). In the earlier work of Marx, the idea of creative destruction or annihilation implies not only that capitalism destroys and reconfigures previous economic orders. It must also ceaselessly devalue existing wealth (e.g., through war, dereliction or economic crises) in order to clear the ground for the creation of new wealth. Schumpeter believes in the existence of an economic equilibrium, whereas Marx defines the boom as a consequence of a depression. The motor of innovation according to Schumpeter is the position of monopoly, which the creative entrepreneur will seize by introducing an innovation.

Clayton M. Christensen: Disruptive Innovation

Not every innovation is of the same kind: there are sustaining innovations made by incremental research, aiming at keeping up technologically with competing firms. Most inventions produce sustaining innovation. Disruptive innovation describes a process, where innovation creates a new market and value network that will eventually disrupt an already existing market and replace existing products. It is more the business model that creates the disruptive impact, than the existence of one high technology invention. Disruptive innovation refers rather to the evolution of a product or service than to a product or service at a certain point. Harvard Business School economist Clayton M. Christensen introduced the term disruptive technologies in his article *Disruptive Technologies: Catching the Wave* (1995) with his cowriter Joseph Bower and described the term further in *The Innovator's Dilemma*. Later he replaced the term by disruptive innovation when he recognized that few technologies are intrinsically disruptive or sustaining, but that it is rather the business model enabled by the technology that has a disruptive character. The evolution from a technological focus to a business modeling focus is a central part of his concept.

Christensen said: "The technological changes that damage established companies are usually not radically new or difficult from a technological point of view. They do, however, have two important characteristics: First, they typically present a different package of performance attributes—ones that, at least at the outset, are not valued by existing customers. Second, the performance attributes that existing customers do value improve at such a rapid rate that the new technology can later invade those established markets" (Christensen 1995). Christensen differentiates between low-end disruption and new-market disruption. The first is targeting at customers that do not need the full performance value; the second is targeting at new customer groups whose needs have not been served by existing incumbents. Christensen's theory of disruptive innovation had a strong influence on business and his book was chosen for one of the most important books in economics. Eventually the term disruptive innovation was said to

be overused and to have become a cliché among people, partly because they do not understand Christensen's concept.

Functionality of Law for Innovation

Innovation as a process is not intrinsically a legal matter. But innovation requires an atmosphere, where values and rights indispensable for innovation have to be protected by the law. Law may be used, to protect existing values in a society, but not to create them. In the same time law has by means of regulations to avoid "parasitic" directions to mention Hayek, if innovation shall create prosperity. "A primary role of law and (when necessary) legislation is to narrow people's options so as to limit opportunities to get rich at other people's expense. So long as the rule of law can internalize external cost and thereby steer innovation in mutually beneficial rather than parasitic directions, an evolving order will be an order of rising prosperity." Law therefore has at least two functions for innovation: protecting the basis for its requirements and steering it by regulation. Law can open up the field for innovation, support enabling innovators to act, set incentives, steer and regulate. The functionality of law therefore is more of an instrument. The innovation process describes different stages of economic development that require legal regulation in various areas of law. As innovation regularly implies inventions or creation of know-how innovation is above all strongly connected to the law of intellectual property. To allow something new to emerge and to keep the position of monopoly created by the implementation of an invention, innovation requires the legal regulation of rights that embody inventions or know-how and can be used as tradable carrier of that right. So, intellectual property law can be regarded as the legal core of innovation (but not the basis). Neighboring subjects as the law of employees' invention and contract law (licensing contracts, research & development contracts) contribute significantly to innovation too. Innovation requires confidentiality, protection of trade secrets, exclusivity and protection against competitors. But these are only

legal topics on the surface of innovation. Innovation requires not only new ideas but also entrepreneurship and capital. In legal terms, innovation therefore requires basic rights and freedoms, as the freedom to do research and to work in an open academic environment with scientific methods. Capital needs to be stable and mobile.

But law may also be used to trigger or facilitate innovation. So legal regulations on funding R & D, enabling public-private partnerships in R & D, company law that offers easy to handle formats for companies and facilitates fast registering, tax law privileging insurance companies offering risk capital to start ups, labor law regulations that enable flexibility in hiring people, or low rates of courts and offices in patent matters may have strong impact on innovation. Innovation is not one legal field, but rather a cross-cutting issue.

Legal Basis of Innovation - the Legal Side of Entrepreneurship

Entrepreneurship is understood as the ability and willingness to act creative. The ability to act creative requires a spirit of individualism and the development of scientific methods in a society. The law can support this spirit and capacities by protecting the basic rights and freedoms: freedom of scientific research, freedom of movement, freedom of establishment, and others. Innovation needs to create its own requirements. Innovation may be a rule of interpretation on the level of constitutional law when it comes to fundamental understanding and interpretation of laws and jurisdiction. The relation between innovation and stability has to be solved in favor of the development of constitutional law, rather than preservation and persistence. In federal states, procedural means to compensate differences in federal judiciary by examining decisions on the level of constitutional law regarding their congruence, as well as the introduction of actions for natural persons on the constitutional level are also ways to open law to innovation. Freedom in business means to provide for competition. In Hayek's words, "the law should not only recognize the principle of private property and freedom of contract, but the legal system should also give a precise definition

of these two principles in a way that promotes competition" (Hayek 1944). With the fast growth of high-tech markets the phenomenon of monopolies and dominant market positions have to be observed closely and if necessary regulated to conserve competition. But innovation has changed structures of markets too: "market shares move faster, barriers to entry the market tend to be much lower, and natural monopolies leave as fast as they come" (Schrepel 2014, 216). Antitrust law therefore has to be updated to these new market mechanisms. "A paradox clearly appears on this point. Indeed, very few persons argue that antitrust should promote perfect competition. However, not to take every single aspects of innovation into account, – for instance, by not including disruptive and permissionless innovations in our analyses – has for consequence to indirectly promote this model of perfect competition. For the reason, it is time to hold innovation as real antitrust standard" (Schrepel 2014, 216).

Confidentiality

Innovation exploits the fact that a product, material, service, or business model is new to the rest of the market and therefore enables a monopoly. Once the invention is to become an innovation, it has to be filed as a patent and implemented in the market as a product. To uphold the monopoly as long as possible, the entrepreneur needs to protect himself against competitors by a ban to use his invention, a patent, by keeping most of his technological competence secret, by defining know-how as trade secret of his business. In all cases, you need confidentiality, which usually is created by non-disclosure agreements, material transfer agreements, and obligations of confidentiality in labor contracts.

Confidentiality has to be limited in time and relevance. An invention kept secret absolutely can hardly have an innovative impact. Sometimes the position of a monopolist can be upheld by suppressing innovation that could disrupt the incumbent product. On the long run though, it is difficult to suppress innovation entirely. So on the

one hand confidentiality is required to create monopoly; on the other hand, there will be no profit without implementation, which means the invention has to be made public to some extent. As innovation implies inventions and invention imply novelty, you need a protection of your research and development results before you file them as a patent. As a result non-disclosure agreements regulate a graded confidentiality, limited to the relevant topics, limited in time and people, who need to know. They oblige their partners to keep information defined as confidential secret and not to use it in any other business relation. Limitation of the obligation here is fundamental, without it, entrepreneurs may end up in an information blockade that eventually destroys business opportunities. Attempts to regulate confidentiality on a European level have not been successful. As non-disclosure agreements are used to create trustful relations between partners, it might be better not to make them obsolete by legal regulation.

Intellectual Property Law

Intellectual property law describes legal rights designed to enable technology transfer and doing business with IP. Patents, utility models, designs, registered marks, as well as trade secrets and know-how are means to protect the innovation against competitors and to keep the monopolistic position as long as possible. The patent is used for the protection of inventions. In the global market, it was very important to have a common understanding and a similar regulation of the patent in the different countries. With the Patent Cooperation Treaty, the Paris Convention for the protection of Industrial property rights, the foundation of the World Intellectual Property Organization there is a basis for economical exploitation of inventions in the global market. A patent is the right to forbid anyone to use the invention that is filed for a patent in the geographical area of the patent for a maximum of 20 years. Global Players need to register in all countries, where they intend to use the patent, which is, where they produce and sell a product, based on the patent. This can result in high costs, depending on the number of countries the patent should be registered for. The patent is

the strongest intellectual property right, especially compared to know-how or utility models. An essential difference between regulations on patents between European and US law is the patentability of business models in US Law, whereas under European Law business models can only be protected by means of protecting its designs or registered marks used by the business model. In Europe, a patent can only be granted for the solution of technical problems. Considering that the so-called platform businesses profit more from their business model, than from a specific high technology invention, the scope of application of European patents should be rethought. The European initiative to introduce a common European Patent and a corresponding judiciary has been stopped for the time being – not by Brexit negotiations, but by court actions challenging the judicial independence of European Patent authorities, currently discussed vigorously.

Law as Stimulation and Facilitation

Law can stimulate innovation by facilitation of legal processes or introduction of completely new legal forms. For example, the Innovation Partnership, a special form of procurement procedure in European public procurement law, introduced in 2014, is a specific procurement procedure with the aim of developing an innovative product or service and then acquiring the resulting services. Due to the special European procurement law for the member states, the legal institution also works in its procurement law which enables flexibility within the European market. A characteristic of the procedure is a respective two-stage of award procedure and the subsequent contract model. Many public procurement regulations on European level as well as on national levels are directed to the promotion of innovation.

Legal Requirements during the Process of Innovation

The process of innovation usually describes the development starting with research and

ending with the revenue gained after commercialization of the innovative product, service, or business model. Depending on the affiliation of authors, the process of innovation may concentrate on the first part and be called Technology Transfer process or concentrate on the commercialization and start up business. Among the plenty and various models for the innovation process, most of them can be divided into three larger phases, subdivided by different steps.

Phase 1: Conception/invention/knowledge phase:

This includes idea generation and evaluation, pre-disclosure, and invention disclosure assessment protection. On the legal side this requires potentially non-disclosure agreements, labor contracts for scientists with paragraphs on intellectual property and confidentiality, invention disclosures, patent or utility filings, patent investigation, and more.

Phase 2: In the second phase, the implementation or adaption phase, the invention made in phase 1 is developed into a product, a prototype, or even more a pilot series, ready for being tested. On the legal side this is accompanied by product development agreements with the company that intend to develop a product. This can either be an existing company or a start-up. In the latter case, the process is completed by financing. Usually the different processes overlap. Once a pilot series has been concluded successfully, innovation is ready for production in phase 3.

Phase 3: It starts with the licensee agreement, production, and continues with the market launch, eventually aiming at market penetration. Legally the step from phase 2 to 3 is enormous, because the law treats production very different from research and development. Liability of manufacturer, quality management, technical data sheets, higher insurance rates, different general conditions are only some of the legal aspects, production and sale require. This is a reason, why usually production is located in a different legal entity to keep the higher risks separate from research and development.

Impact of Innovation on the Development of Law. New Legal Questions?

Law can influence innovation, as well as innovation has an impact on the development of law. Digitalization, artificial intelligence, and information technologies let us rediscover old problems in a new form, but sometimes also raise new questions. The relation between these innovative technological fields and law, the question of regulation of technology is a cross-section field: liability, traffic law, data protection law, and criminal law. Digitalization requires data protection and raises the question of information privacy, especially when digitalization is introduced in the public sector for e-government, as introduced by the European Union with the eGovernment action plan 2016–2020 and e-government laws on the national level as the German law in 2013. As mentioned before, on one hand law will have to facilitate innovation by enabling mobility of citizens and businesses by cross-border and facilitating digital interaction between administrations and citizens/businesses for high-quality public services. On the other hand, it will have to limit the amount of personal data to be collected and protect citizens against misuse of personal data.

Data protection law in Europe is criticized for fulfilling its purpose only imperfectly. In wide ranges, it is still based on the data protection concept of the 1970s, when central data storage on mainframes, limited storage capacities, and a relatively small circle of – mostly government – data processors were characteristic. Only slowly legislation can keep up with the needs of technical development. The European Union started a general data protection reform and issued the General Data Protection Regulation in 2016, according to which the Data Protection Directive of 1995 expires. On the national level, data protection law, e.g., in Germany, is considered “over-regulated, fragmented, confusing and contradictory” (Roßnagel et al. 2001).

Digitalization requires the regulation of issues raised by the increasing use of automated and autonomous devices. Especially the fact that we

allow machines to make decisions for us raises significant problems of liability, responsibility, and the nature of legal personality. For the self-driving car law, for example, we will have to decide, which risks will be carried by manufacturer's responsibility and what will lie within the driver's liability. The solution of dilemmas in emergency situations, to be decided by machines, may not be solvable by law, but rather by ethics, sociology of technics, and "interdisciplinary methods to be developed" (Valentiner 2016).

Summary

Innovation is the process of devising a new idea or thing or improving an existing idea or thing. In contrast to the term invention, innovation requires also the aspect of the new idea being applied or performed in a way that matters within the context of the innovation. Innovation as a term of economics has been introduced by *Joseph Schumpeter* and is strongly connected to his idea of creative destruction. Clayton Christensen coined the term destructive innovation. His theories had a strong influence on business administration. Innovation requires entrepreneurship and capital. The function of law is to protect the basis for innovation and to regulate and steer it. Digitalization, information technologies, and artificial intelligence are challenges to the law. After decades of rising numbers of legal regulations, innovation could take us to the limits of law.

Cross-References

- Creativity
- Entrepreneurship
- European Patent System
- Hayek, Friedrich August von
- Information Disclosure
- Patent Litigation
- Patent Opposition
- Rule of Law
- Trade Secrets Law

References

- Christensen Clayton (1995) Disruptive technologies. Catching the wave, Harvard business Review, January
- Communication of the Commission Mitt. der Kom. v. 2010 (KOM) (2010) „Leitinitiative der Strategie Europa 2020 Innovationsunion“
- Hayek F (1944) Road to serfdom, 2001 first published 1944. Routledge Classics, London
- Roßnagel Alexander, Pfitzmann Andreas, Garstka Hansjürgen (2001) Modernisierung des Datenschutzrechts. Gutachten im Auftrag des Bundesministeriums des Innern. Berlin
- Sandefur T (2008) Innovation. In: The concise encyclopedia of economics. Springer, New York
- Schumpeter JA (1939) Business cycles. A theoretical, historical, and statistical analysis of the capitalist process. MacGraw-Hill, New York
- Valentiner Dana (2016) Gesetzgeberische Herausforderungen der Technikregulierung – ein Aufriss, juwiss.de/43–2016/
- ## Further Reading
- Christensen Clayton, Raynor Michael, McDonald Rory (2015) What is disruptive innovation? Harvard Business Review, Dec 2015
- Djeffal Christian (2017) Leitlinien der Verwaltungsinnovation und das Internet der Dinge, Alexander von Humboldt, Institut für Internet und Gesellschaft
- Fritsch M, Werker C (1999) Innovation Systems in Transition. Innovation and Technological Change in Eastern Europe: pathways to industrial recovery. M. Fritsch and H. Brezinski. Cheltenham, UK, Edward Elgar
- Hayek F (1944) Road to serfdom. Routledge Press, London
- Hoffmann-Riem, Wolfgang (2016) Innovation und Recht – Recht und Innovation, Recht im Ensemble seiner Kontexte
- Kaschny M, Nolden M, Schreuder S (2015) Innovation management in SMEs: strategies, implementation, practical examples. Springer Gabler, Wiesbaden
- Kühling J, Seidel C (2015) Anastasios Sivridis: Datenschutzrecht, 3rd edn. Müller, Heidelberg
- Lepore Jill (2014) The disruption machine, The New Yorker. June 23
- Schmidt David (2016) Friedrich Hayek. In Edward N. Zalta (ed) The Stanford encyclopedia of philosophy (Winter 2016 edn), URL = <https://plato.stanford.edu/archives/win2016/entries/friedrich-hayek/>
- Schrepel T (2014) Friedrich Hayek's contribution to antitrust law and its modern application. Global Antitrust Review:199–216
- Schulz Wolfgang, Dankert Kevin (2016) Governance by things as a challenge to regulation by law, IPR 5
- Thurston Thomas, Crunch Network (2014) Christensen Vs. Lepore: a matter of fact, posted 30 Jun 2014 by Thomas Thurston (@thurstont)
- Tinnefeld Marie-Theres, Buchner Benedikt, Petri Thomas (2012) Introduction to data protection law. In: Data

protection and freedom of information in a European perspective. 5th edn. Oldenbourg
 von Hippel Eric The sources of innovation, Oxford University Press, New York 1988, Download courtesy of OUP at <http://web.mit.edu/evhippel/www>

Institution

► Organization

Institutional Change

Christopher Kingston
 Amherst College, Amherst, MA, USA

Definition

Institutions are durable; that is precisely what makes them meaningful and important. But institutions also sometimes change. This entry compares a variety of theoretical approaches to understanding the process of institutional change. Some authors treat institutional change as a centralized, collective-choice process in which rules are explicitly specified by a collective political entity, such as the community or “the state,” and individuals and organizations engage in collective action, conflict, and bargaining to try to change these rules for their own benefit. Others emphasize the “spontaneous” emergence of institutions as an evolutionary process, in which new institutional forms periodically emerge and undergo some kind of decentralized selection process as they compete against alternative institutions. Still others combine elements of evolution and design. We differentiate a variety of approaches to the interaction between formal and informal rules and explore the path-dependent nature of institutional change. We also discuss recent theories based on the “equilibrium view” of institutions, which emphasizes that the constraints that motivate individual behavior are ultimately derived from expectations about the behavior of other actors in various contingencies. In maximizing

their welfare subject to these constraints, agents choose strategies which, in the aggregate, give rise to expectations which reinforce the constraints on everyone else, so the effective “rules” of the game emerge endogenously as equilibrium outcomes, and exploring institutional change involves explaining changes in equilibrium behavior.

Theorizing Institutional Change

The scholarly literature on institutional change is voluminous, but diffuse and eclectic, plagued by ambiguity about the meaning of commonly used terms including even the meaning of the term “institutions” itself. In this entry, adapted from a broader paper on the subject (Kingston and Caballero 2009), I attempt to map out the major theoretical approaches to institutional change and to highlight the main possibilities that an empirical researcher might consider.

Before we can discuss institutional change, we must define what we mean by “institutions.” Unfortunately, the appropriate definition is far from a settled issue, and the different definitions used by different authors naturally influence their views of institutional change. Many authors, however, adopt some variant of Douglass North’s view that institutions “are the rules of the game in a society or more formally, are the humanly devised constraints that shape human interaction” and that they “reduce uncertainty by providing a structure to everyday life” (North 1990: 3). Fundamentally, then, institutions are viewed as “rules” that are “humanly devised.”

There are many different kinds of “rules”; however, most authors follow North in distinguishing between “formal” rules such as laws and constitutions and “informal” constraints such as conventions and norms. Even this basic distinction is far from straightforward. The term “formal” is often taken to mean that the rules are made explicit or written down, particularly if they are enforced by the state, whereas “informal” rules are implicit; another interpretation is that formal rules are enforced by actors with specialized roles, whereas informal codes of behavior are

enforced collectively by the members of the relevant group. “Informal constraints,” as North notes, “defy, for the most part, neat specification” (North 1990: 36) but include “socially sanctioned norms of behavior” as well as “extensions, elaborations, and modifications of formal rules” and “internally enforced standards of conduct” (North 1990: 40).

But while social norms enforced by a community, conventions, and internalized ethical codes such as religious beliefs can all be viewed as informal constraints, they are distinct phenomena which may change in different ways and may have different short-run and long-run effects on the broader pattern of institutional change. This ambiguity has created considerable confusion about the nature of informal constraints and their interaction with formal rules. For example, some authors have regarded informal constraints as essentially immutable, or that they change only slowly, but there have also been episodes when at least some kinds of informal constraints, such as social norms, can change rapidly.

For the moment, then, let us treat institutions as the (formal and informal) “rules of the game” in a society; we will discuss an alternative definition later. How do these rules change? Two main kinds of processes emerge from the literature. Some authors view institutional change as the outcome of purposeful (centralized) design, either by a single individual (such as when a king issues a royal decree) or by many individuals or groups interacting to create or change rules through some kind of collective-choice or political process. Other authors envision institutional change as a more gradual, evolutionary (decentralized) process, frequently involving competition among alternative institutional forms. In empirical settings, aspects of both kinds of processes are often present, and the question becomes how they interact. This raises a host of conceptual issues, including how we should think about the interaction between formal rules and informal constraints, the role of politics and collective action, the nature of “competition” between different institutional forms, the role of bounded rationality and learning, the exogenous and

endogenous causes of institutional change, and the role of history and the potential for path dependence.

Politics and Collective Choice

Many authors treat institutional change as a centralized, collective-choice process in which rules are explicitly specified by a collective political entity, such as the community or the state, and individuals and organizations engage in collective action, conflict, and bargaining to try to change these rules for their own benefit.

Ostrom (2005), for example, envisions a multilayer nested hierarchy of rules: “operational rules” that govern day-to-day interactions, “collective-choice rules” (rules for choosing operational rules), and “constitutional rules” (rules for choosing collective-choice rules). In order to analyze how rules are formed at one level, Ostrom temporarily treats the higher levels of rules as fixed (*ibid.*: 61). For example, when “operational rules” are being chosen, constitutional and collective-choice rules are treated for the moment as exogenous.

The impetus for institutional change arises when some group or individual perceives an opportunity to change rules in a way that benefits them. This could be because of an exogenous change in underlying parameters that change the perceived costs and benefits from an institutional change – for example, a change in technology or relative prices. But the cause might also be endogenous if, say, people’s choices under one set of rules gradually lead to changes in parameter values that alter the costs and benefits of institutional change and thereby undermine current institutions. For example, Acemoglu and Robinson (2005) argue that after 1500, some European countries experienced a substantial growth in Atlantic trade which increased the political power of merchant groups. The growing strength of merchant groups in these countries led to institutional changes that constrained the power of monarchs and led to the development of institutions that were more conducive to economic growth.

The process of institutional change, in Ostrom's framework, is this: each individual calculates their expected costs and benefits from an institutional change, and if a "minimum coalition" necessary to effect change agrees to it, the change is enacted. What constitutes a "minimum coalition" is determined by the higher-level rules; for example, in a dictatorship, the dictator alone might constitute a winning coalition; in a democracy, a majority would constitute a winning coalition. The outcome therefore depends on how decision-makers perceive the likely effects of a change in rules and on whether those that desire change are able to bring it about or whether those that expect to lose by the change are able to block it under the higher-level rules which frame the political (rule-making) contest. Powerful groups may be able to block beneficial change or impose inefficient change. Of course, some kind of compromise or partial institutional change is also a possible outcome.

Free-rider problems can impede collective action to change formal rules. Voting, protesting, joining political associations, and learning about the impact of potential policies may all be individually nonrational actions, even if the individual cares deeply about the result. Leaders can play a key role by offering a "vision": attempting to shift their supporters' perceptions of the costs of the existing system or the feasibility and desirability of some proposed alternative. Whether they seek wealth or power, or are driven by ethical or philosophical values, the fundamental challenge that leaders must overcome in order to achieve institutional change is that of winning support and overcoming collective action problems among their potential supporters.

The role of political actors, such as judges and politicians, can be envisaged in a variety of ways. For some purposes, politicians might be viewed as simply reflecting the interests of particular groups, so that the political process remains essentially a battleground in which interest groups compete to mold formal rules to their own advantage. Other theories, however, give political actors a more autonomous role. For example, in Kantor's (1998) framework, groups of constituents lobby politicians to change formal rules, and the

politicians have incentives to be responsive to their constituents' demands. However, the politicians also have their own objectives and face other political and constitutional constraints. Which of these perspectives on the role of political actors is most appropriate will depend on the configuration of interest groups and the political structure in a given context.

Another important set of barriers to institutional change arises when the beneficiaries of a reform cannot credibly commit to compensate the losers because of the lack of an external authority to enforce intertemporal bargains. These problems are exacerbated when there is uncertainty or when there are a large number of parties to a negotiation. Acemoglu and Robinson (2006) highlight the importance of commitment problems as a cause of institutional change. In their theory, disenfranchised groups can use violence to force constitutional change (seize power), but because of collective action problems, they can only organize violence during rare (and exogenous) moments of crisis. Violence is destructive, so in a crisis, there is an opportunity for a mutually beneficial bargain in which the incumbent ruling groups would agree to carry out reforms, in exchange for which the disenfranchised groups would refrain from carrying out a revolution. However, the disenfranchised groups' opportunity to use force is fleeting, and the ruling groups cannot credibly commit themselves to honor their commitments to reform after the moment of crisis is passed. Therefore, a revolution to effect institutional change, though destructive and costly, may be the only way for currently disenfranchised groups to credibly constrain the policy choices of future governments.

Many authors note that institutional change a "path-dependent" process – that is, the institutions observed at any point in time, may in part be a function not just of current technology but also of precedent institutions and technologies (David 1994, 2007). Frequently, existing institutions create groups with a vested interest in preserving the status quo, which can impede institutional change and enable inefficient institutions to persist. More generally, the resources, physical and human capital, skills, technologies, and

organizations accumulated under one set of institutions can gradually alter the set of technologically feasible institutions and thereby affect future institutional development. Furthermore, even when these impediments are overcome, institutional change is usually incremental since it is often easier to achieve consensus on small adjustments than to effect major changes to existing rules.

There are also important behavioral aspects to the issue. The way people process information, solve problems, and learn may be important for understanding institutional inertia and change. For example, people may systematically misperceive opportunities in a complex and changing environment or may be unaware of potentially beneficial institutional changes until the new institutions are “invented.” North (2005) presents a framework in which economic actors have “mental models” which reflect their understanding of the world and which they use to evaluate the desirability of particular rule changes. Over time, as they learn about the world, they revise their mental models and may alter their perceptions of the effects of alternative rules and of the set of possible alternatives, providing an impetus for institutional change. This suggests that a key to understanding institutional change is an understanding of how people learn and revise their “mental models.” Ongoing research in cognitive psychology and behavioral economics offers the promise of deepening our understanding of many aspects of institutional change.

Although viewing institutional change as the outcome of a deliberate, collective-choice process of rule creation yields many insights, it also leaves several important questions unanswered. In particular, it has difficulty explaining why, in many cases, formal rules are ignored or fail to produce their intended outcome. A key reason for this difficulty is the prevalence of “informal” constraints that are not a product of deliberate design and often vaguely defined. To clarify, let us distinguish three types of “informal rules.”

First, the term “informal” is sometimes simply used to indicate that the rules are not written down or are not enforced by the state. A related

interpretation emphasizes the importance of actors with specialized roles in enforcing formal rules (Milgrom et al. 1990).

Second, informal constraints are sometimes viewed as ethical codes or moral “norms” which are internalized and directly reflected in players’ preferences.

Third, an important category of “informal rules” includes conventions – viewed as self-enforcing solutions to multiplayer coordination games (Sugden 1989) – and “social norms” which use a multilateral reputation mechanism and a credible threat of punishment to generate trust among members of a community (see, e.g., Kandori 1992; Greif 2006). Of course, these constraints may eventually come to be followed without rational evaluation and socially experienced as moral, ideological, or “cultural” rather than purely strategic constraints. However, if all others are following such rules, then even fully rational strategic players may also be induced to follow them. This is crucial because, even if most people follow the norm without rational evaluation, it is unlikely that a norm could evolve or survive if a rational mutant could achieve a higher payoff by deviating from it.

The second and third categories of informal rules (moral norms, conventions, and social norms) do, sometimes, change over time, but they do not fit easily into the collective-choice models because they generally evolve in a decentralized, “spontaneous” manner, rather than being deliberately designed and agreed to. This may be a serious shortcoming in some contexts because the evolution of informal rules is frequently an important part of the story of institutional change. Internalized moral codes, for example, may to some extent evolve to be compatible with prevailing formal rules, but they can also impact how formal rules change. For example, certain proposed rules may not be adopted because they are perceived as “unfair.” This is another channel through which “history matters.” Similarly, it is not uncommon for informal “rules” to originate as voluntary patterns of behavior that develop within a community and are later “formalized.” Many aspects of commercial law, for example, derived from the codification of

merchant's practices which had evolved spontaneously (Milgrom et al. 1990; Kingston 2007).

Evolutionary Theories of Institutional Change

A large body of literature treats institutional change as an evolutionary process, in which new institutional forms periodically emerge and undergo some kind of decentralized selection process as they compete against alternative institutions. In the evolutionary theories, new rules or behaviors (mutations) may emerge from deliberate human actions (including learning, imitation, and experimentation), or they may develop "spontaneously" from the uncoordinated choices of many individuals. The key difference between evolutionary theories and the collective-choice theories discussed in the previous section has to do with the decentralized selection process which determines which rules ultimately become widely adopted. Those institutions that prove successful spread, by imitation or replication, while unsuccessful institutions die out. As a result, overall institutional change occurs "spontaneously," through the uncoordinated choices of many agents, rather than via a single, collective-choice or political mechanism.

Sometimes, an evolutionary model of institutional change is implicit. Williamson (2000)'s "Transactions cost economics," for example, *assumes* that the sets of rules ("governance structures") that can most efficiently govern any particular transaction (those that "minimize transactions costs") are those that will be observed. Implicitly, this rests on an assumption that competitive pressure would weed out inefficient forms of organization, as originally suggested by Alchian (1950), because more efficient organizational forms will yield higher profits and will therefore survive and be imitated. So, for example, if a change in production technology renders existing institutions inefficient, competitive pressures ensure that a new configuration of optimal institutions will emerge. This approach is an example of "functionalist" reasoning: to

explain the attributes of an institution, we need to only ask what function it serves.

Although a functionalist approach can successfully explain many aspects of observed institutions, it does best in situations in which competition among institutional forms can plausibly operate to weed out inefficient rules and leads to a unique, optimal equilibrium. It has difficulty explaining why countries with similar technologies may use different institutions to govern apparently similar transactions, why inefficient institutions often seem to persist, or why less successful societies often fail to adopt the institutional structure of more successful ones.

The essence of these difficulties is that evolutionary processes frequently exhibit multiple equilibria. For example, credit cards are widely used by consumers because many merchants accept them for payment; and they are widely accepted because they are widely used. The resulting equilibrium is associated with a set of formal rules (credit card agreements), norms, and behaviors (carrying little cash), but we can easily envisage many other alternative institutional configurations, any of which, once established, would generate stable expectations and behavior within the context of a different set of "rules." That, is there may be multiple possible sets of self-enforcing rules ("conventions"), and there is no guarantee that the most efficient will be observed (Sugden 1989).

The possibility of multiple evolutionarily stable equilibria has two important, closely related consequences. First, observed institutions are not necessarily "efficient." And second, institutional change may exhibit "path dependence," in the sense that initial conditions and historical events can have a lasting impact on the institutions which are ultimately observed. For example, an institutional structure which was previously optimal might become sub-optimal as circumstances change, but without a coordinating device, such as legislation or the appearance of a "political entrepreneur," to engineer a change in the rules, the economy might remain stuck in the (now) sub-optimal equilibrium. Young (1996) uses an evolutionary framework to argue that historical accidents could lead to the selection of particular

conventions and argues that in the long run, the pattern of institutional change will follow a “punctuated equilibrium” process in which rapid switches between conventions are interspersed with long periods of stability.

Brousseau and Raynaud (2011) argue that many institutional arrangements begin as “private” local experiments, in which participation is voluntary, but that over time, through competition for adherents, economies of scale, and network effects, some (not necessarily optimal) institutions spread and emerge as “winners” and become “solidified” and formalized as part of the institutional environment, while participation in them becomes increasingly widespread and mandatory. Thus, they argue, some kinds of informal institutions can gradually “climb the ladder” and metamorphose into formal and permanent rules.

Blending Evolution and Design

Both the evolutionary and design-based approaches to understanding institutional change are useful in particular settings, but both are incomplete. Theories which view institutional change as the outcome of a centralized collective-choice process have difficulty explaining changes in informal constraints (such as social norms) that evolve in a decentralized manner, while evolutionary theories tend to neglect the role of collective action and the political process. This is not meant as a criticism of these theories: they have been developed to study a variety of different situations, and the assumptions and conclusions naturally reflect these differences. But in many real-world settings, both evolutionary processes and intentional design are at work, and it will often be difficult to cleanly separate the two. For example, gradual underlying changes in parameters, beliefs, or knowledge, which result from the spontaneous evolution of existing institutions over time, may give rise to deliberate attempts to design and implement new institutions; and following such attempts, competition or other evolutionary processes may subsequently play a role in determining whether particular institutional innovations survive and spread.

The question therefore naturally arises as to how to integrate these theories. To a large extent, this turns on the interaction between formal rules, which are generally deliberately designed (although there may also be evolutionary processes underlying their creation, as in the case of the common law), and informal rules, which are “much more impervious to deliberate policies” (North 1990: 6) and therefore (usually) evolve spontaneously.

Here, again, it is useful to begin with North (1990)’s seminal contribution. In North’s account, formal rules change through a political process as a result of deliberate (though boundedly rational) actions by organizations and individual entrepreneurs, while informal rules evolve alongside, and as extensions of, formal rules. Informal rules play a key role in institutional change because they change slowly and cannot be changed deliberately. Following a change of formal rules, therefore, the informal rules which “had gradually evolved as extensions of previous formal rules” (ibid.: 91) survive the change, so that the result “tends to be a restructuring of the overall constraints – in both directions – to produce a new equilibrium that is far less revolutionary” (ibid.: 91). Essentially then, formal rules occupy the driving role in institutional change; informal constraints apply the brakes.

In general, there are multiple equilibria and no guarantee of an efficient outcome (North 1990: 80–81: 136). The process of institutional change is also path-dependent because individuals learn, organizations develop, and ideologies form in the context of a particular set of formal and informal rules (1990, chapter 9). These organizations then may attempt to change the formal rules to their benefit, and over time this in turn may (indirectly) affect the informal rules.

Roland (2004) distinguishes between “fast-moving” (political) institutions (akin to formal rules), which can be changed quickly and deliberately via the centralized political process, and “slow-moving” (cultural) institutions (akin to informal rules), which change slowly because change is continuous, evolutionary, and decentralized. He outlines his view of institutional change by analogy: tectonic pressures along fault

lines (changes in slow-moving institutions) build up continuously but slowly and then suddenly provoke an “earthquake” that causes abrupt and substantial changes in fast-moving institutions (i.e., formal rules). Thus, Roland’s theory is, in a sense, the inverse of North’s, in that changes in informal rules, rather than formal rules, are the main drivers of institutional change.

The “Equilibrium View” of Institutions

A growing body of recent research shifts the focus by identifying institutions with equilibrium patterns of behavior rather than the “rules” that induce the behavior (Greif and Kingston 2011). This “equilibrium perspective” emphasizes that the constraints that motivate individual behavior are ultimately derived from expectations about the behavior of other actors in various contingencies. In maximizing their welfare subject to these constraints, agents choose strategies which, in the aggregate, and perhaps unintentionally, give rise to expectations which reinforce the constraints on everyone else. Of course, these expectations may be summarized in the form of formal and informal “rules.” But the constraints that motivate behavior – the “true” rules of the game – emerge as endogenous equilibrium outcomes, reflecting a socially constructed reality.

From the equilibrium perspective, institutional change becomes fundamentally not about changing rules, but about changing expectations; the essential goal of introducing new “rules” is to help players coordinate on one of the many possible equilibria by coordinating their beliefs about each other’s expected behavior both on and off the path of play. In equilibrium, however, it is ultimately these behavioral expectations, rather than the prescriptive content of the rules themselves, that motivate compliance. And, for a variety of reasons, it is possible that an attempt to introduce new “rules” may fail to change these expectations and equilibrium patterns of behavior or may change them in unanticipated ways. The process of institutional change is consequential precisely because there are typically multiple equilibria.

A rule “forbidding” some behavior, for example, will be effective only if people generally expect others (including those charged with enforcing the rule) to act in a way which makes it effective.

A formal rule making one player a judge, president, or police officer does not change that player’s set of physically feasible actions, but it *may* systematically alter people’s perceptions about how those actions are to be interpreted and how other players will respond: *if* the rule is effective, then by virtue of her role, the judge can take actions and give orders which would not be followed if she were an ordinary citizen. In order for the rules to be effective, the behavior specified by the “rules” – including that of the “enforcers” of the rules – must correspond to an equilibrium in the underlying “game of nature.”

Theories that define institutions as rules tend to obscure these possibilities because they consider the enforcement of rules separately from their content; they cannot explain why some rules are followed and others are not. For example, the legal speed limit on highways in Massachusetts is 65 miles per hour, but this limit is widely ignored. This is not to say that there are no “rules,” however. Most cars travel around 70 mph and are (almost) never pulled over at this speed, whereas cars traveling at 80 mph sometimes are (and those traveling at 90 mph frequently are). What accounts for the difference between the behavior specified by the “formal rule” and the behavior actually observed? Asserting that there is an “informal rule” specifying the observed behavior is unsatisfactory; it merely assumes away what we would most like to explain. The equilibrium perspective offers a more satisfactory explanation: drivers’ and police officers’ behaviors constitute an overall equilibrium based on a broadly shared set of behavioral expectations. By making the “enforcement” of rules endogenous to the analysis, this approach enables the treatment of “formal” and “informal” constraints in a unified manner.

Both deliberate, centralized and evolutionary, decentralized institutional changes are compatible with the equilibrium view. Exogenous parameter shifts such as changes in technology or

preferences can disrupt an equilibrium, leading individuals and organizations to try to change the “formal rules” in order to achieve a coordinated shift of many players’ beliefs about each others’ strategies. Previous institutions also provide focal points which can affect equilibrium selection in novel situations (Sugden 1989). Alternatively, gradual changes in parameters might cause gradual adjustments to expectations and behavior. Since the formal rules remain unchanged, this kind of institutional change could, of course, be interpreted as changing “informal rules,” but that merely labels the phenomenon without explaining it. Fundamentally, what is changing is a pattern of equilibrium behavior.

Greif and Laitin (2004) refer to parameters which are exogenous in the short run but which gradually change as a result of the play of the game as “quasi-parameters.” Changes in quasi-parameters may either broaden the range of situations in which the existing pattern of behavior (institution) is an equilibrium or may undermine the existing institution, leading to an “institutional disequilibrium” and an impetus for institutional change. Institutional change, in this view, is highly path-dependent; as Greif and Laitin emphasize, knowledge, resource ownership, wealth distribution, and other “quasi-parameters” can all be affected by past institutions and affect both the future institutional choice set (the set of feasible equilibria) and the choice of institutions within that set.

Conclusion

The appropriate model for studying institutional change is largely a matter of context. For situations in which competition will tend to weed out inefficient institutional forms, functionalist explanations such as the “transaction cost” model are likely to be useful for explaining observed institutions. In situations in which changes in formal rules occur within a stable political context, and have relatively predictable effects on behavior, treating institutional change as an outcome of collective action and political maneuvering may be more suitable. However, this approach cannot

explain why some formal rules become effective and others do not and tend to neglect the role of informal rules. The equilibrium view of institutions provides a more complete theory by treating both informal and formal rules, and their enforcement, within an integrated framework, and is therefore useful as a broad conceptual framework for understanding institutional change. However, it may introduce unnecessary complexity in the many real-world cases in which formal rules are relatively straightforward and effectively enforced.

Cross-References

- ▶ [Constructivism, Cultural Evolution, and Spontaneous Order](#)
- ▶ [De Jure/De Facto Institutions](#)
- ▶ [Institutional Economics](#)
- ▶ [Path-Dependent Rule Evolution](#)
- ▶ [Political Economy](#)
- ▶ [Transaction Costs](#)

References

- Acemoglu D, Robinson JA (2005) The rise of Europe: Atlantic trade, institutional change, and economic growth. *Am Econ Rev* 95(3):546–579
- Acemoglu D, Robinson JA (2006) *Economic origins of dictatorship and democracy*. Cambridge University Press, Cambridge
- Alchian A (1950) Uncertainty, evolution and economic theory. *J Polit Econ* 58(3):211–221
- Brousseau E, Raynaud E (2011) Climbing the hierarchical ladders of rules: a life-cycle theory of institutional evolution. *J Econ Behav Organ* 79(1):65–79
- David PA (1994) Why are institutions the ‘carriers of history’. *Struct Chang Econ Dyn* 5(2):205–220
- David PA (2007) Path dependence, its critics and the quest for ‘historical economics’. In: Hodgson GM (ed) *The evolution of economic institutions: a critical reader*. Edward Elgar Press, Cheltenham, pp 120–142
- Greif A (2006) *Institutions and the path to the modern economy*. Cambridge University Press, Cambridge
- Greif A, Laitin D (2004) A theory of endogenous institutional change. *Am Polit Sci Rev* 98(4):633–652
- Kandori M (1992) Social norms and community enforcement. *Rev Econ Stud* 59:63–80
- Kantor SE (1998) *Politics and property rights: the closing of the open range in the postbellum south*. University of Chicago press, Chicago

- Kingston C (2007) Marine insurance in Britain and America, 1720–1844: a comparative institutional analysis. *J Econ Hist* 67(2):379–409
- Milgrom P, North D, Weingast B (1990) The role of institutions in the revival of trade: the law merchant, private judges, and the champagne fairs. *Econ Polit* 1:1–23
- North D (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
- North D (2005) Understanding the process of economic change. Princeton University Press, Princeton
- Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
- Roland G (2004) Understanding institutional change: fast-moving and slow-moving institutions. *Stud Comp Int Dev* 38(4):109–131
- Sugden R (1989) Spontaneous order. *J Econ Perspect* 3(4):85–97
- Williamson O (2000) The new institutional economics: taking stock, looking ahead. *J Econ Lit* 38:595–613
- Young HP (1996) The economics of convention. *J Econ Perspect* 10(2):105–122

Further Reading

- Aoki M (2001) Towards a comparative institutional analysis. MIT Press, Cambridge
- Kingston C, Caballero G (2009) “Comparing theories of institutional change”. *J Inst Econ* 5(2):151
- Greif A, Kingston C (2011) Institutions: rules or equilibria? In: Caballero G, Schofield N (eds) Political economy of institutions, democracy and voting. Springer, Berlin
- North D (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge

Institutional Complementarity

Fabio Landini¹ and Ugo Pagano^{2,3}

¹LUISS University, Rome, Italy

²University of Siena, Siena, Italy

³Central European University, Budapest, Hungary

Abstract

Institutional complementarity refers to situation of interdependence among institutions. The present article presents a formal representation of institutional complementarity and discusses several economic applications of this concept.

Definition

Even if several definitions have been proposed in the literature (Crouch et al. 2005) they share the idea that institutional complementarity refers to situations of interdependence among institutions. Institutional complementarity is frequently used to explain the degree of institutional diversity that can be observed across and within socioeconomic system and its consequences on economic performance. In particular, the concept of institutional complementarity has been used to illustrate why institutions are resistant to change and why introducing new institutions into a system often leads to unintended, sometimes suboptimal, consequences.

Institutional Complementarity

The canonical formal representation of the concept of institutional complementarity is due to Aoki (2001) and relies on the theory of supermodular games developed by Milgrom and Roberts (1990). The basic structure of the model takes the following form.

Let us consider a setting with two institutional domains, A and B , and two sets of agents, C and D that do not directly interact with each other. Nevertheless, an institution implemented in one domain parametrically affects the consequences of the actions taken in the other domain. For instance, A can be associated with the type of ownership structure prevailing in a given country and B with the structure of labor rights. For simplicity, we assume that the technological and natural environment is constant.

Suppose that the agents in domain A can choose a rule from two alternative options: A^1 and A^2 ; similarly, agents in domain B can choose a rule from either B^1 or B^2 . For simplicity, let us assume that all agents in each domain have an identical payoff function $u_i = u(i \in C)$ or $v_j = v(j \in D)$ defined on binary choice sets of their own, either $\{A^1, A^2\}$ or $\{B^1, B^2\}$, with another sets as the set of parameters. We say that an (endogenous) “rule” is institutionalized in a

domain when it is implemented as an equilibrium choice of agents in the relevant domains.

Suppose that the following conditions hold:

$$\begin{aligned} u(A^1; B^1) - u(A^2; B^1) \\ \geq u(A^1; B^2) - u(A^2; B^2) \end{aligned} \quad (1)$$

$$\begin{aligned} v(B^2; A^2) - v(B^1; A^2) \\ \geq v(B^2; A^1) - v(B^1; A^1) \end{aligned} \quad (2)$$

for all i and j . The latter are the so-called supermodular (complementarity) conditions. Equation 1 implies that the “incremental” benefit for the agents in A from choosing A^1 rather than A^2 increases as their institutional environment in B is B^1 rather than B^2 . Equation 2 implies that the “incremental” benefit for agents in B from choosing B^2 rather than B^1 increases if their institutional environment in A is A^2 rather than A^1 . Note that these conditions are concerned with the property of incremental payoffs with respect to a change in a parameter value. They do not exclude the possibility that the level of payoff of one rule is strictly higher than that of the other for the agents of one or both domain(s) regardless of the choice of rule in the other domain. In such a case the preferred rule(s) will be implemented autonomously in the relevant domain, while the agents in the other domain will choose the rule that maximizes their payoffs in response to their institutional environment. Then the equilibrium of the system comprised of A and B – and thus the institutional arrangement across them – is uniquely determined by preference (technology).

However, there can also be cases in which neither rule dominates the other in either domain in the sense described above. If so, the agents in both domains need to take into account which rule is institutionalized in the other domain. Under the supermodularity condition, there can then be two pure Nash equilibria (institutional arrangements) for the system comprised of A and B , namely, $(A^1; B^1)$ and $(A^2; B^2)$. When such multiple equilibria exist, we say that A^1 and B^1 , as well as A^2 and B^2 , are “institutional complements.”

If institutional complementarity exists, each institutional arrangement characterizes as a self-sustaining equilibrium where no agent has an incentive to deviate. In terms of welfare, it may be the case that possible overall institutional arrangements are not mutually Pareto comparable or that one of them could be even Pareto sub-optimal to the other. In these cases, history is the main force determining which type of institutional arrangements is likely to emerge, with the consequence that suboptimal outcomes are possible.

Suppose, for instance, that $(A^2; B^2)$ is a Pareto-superior institutional arrangement in which $u(A^2; B^2) > u(A^1; B^1)$ and $v(B^2; A^2) > v(B^1; A^1)$. However, for some historical reason A^1 is chosen in domain A and becomes an institutional environment for domain B . Faced with this institutional environment, agents in domain B will correspondingly react by choosing B^1 . Thus the Pareto-suboptimal institutional arrangement $(A^1; B^1)$ will result. This is an instance of coordination failure in the presence of indivisibility.

Obviously, there can also be cases where $u(A^2; B^2) > u(A^1; B^1)$ but $v(B^1; A^1) > v(B^2; A^2)$. This is an instance where the two viable institutional arrangements cannot be Pareto ranked. Agents exhibit conflicting interests in the two equilibria, and the emergence of one institutional arrangement as opposed to the other may depend on the distribution of decisional power. If for some reasons agents choosing in domain A have the power to select and enforce their preferred rule, arrangement $(A^2; B^2)$ is the most likely outcome. Alternatively, agents choosing in domain B will force the society to adopt $(B^1; A^1)$.

Pagano (1992) and Pagano and Rowthorn (1994) are two of the earliest analytical contributions to institutional complementarity. In their models, the technological choices take as parameters property rights arrangements whereas the latter are made on the basis of given technologies. The complementarities of technologies and property rights create two different sets of organizational equilibria. For instance, strong rights of the capital owners and a technology with a high intensity of specific and difficult to monitor capital are likely to be institutional complements and define one possible organizational equilibrium. However,

also strong workers' rights and a technology characterized by a high intensity of highly specific labor can be institutional complements and define an alternative organizational equilibrium. The organizational equilibria approach integrates the approaches *à la* Williamson (1985), which have pointed out the influence of technology on rights and safeguards, and the views of the radical economists (see, for instance, Braverman 1974), who have stressed the opposite direction of causation. The complementarities existing in the different organizational equilibria integrate both directions of causation in a single analytical framework. A similar approach has been used to explain organizational diversity in knowledge-intensive industries, such as software (Landini 2012, 2013).

Institutional complementarities characterize also the relations between intellectual property and human capital investments. Firms owning much intellectual property enjoy a protection for their investments in human capital, which in turn favors the acquisition of additional intellectual property. By contrast other firms may find themselves in a vicious circle where the lack of intellectual property inhibits the incentive to invest in human capital and low levels of human capital investments involve that little or no intellectual property is ever acquired (Pagano and Rossi 2004).

Less formal approaches to institutional complementarities have also been adopted. In their seminal contribution, Hall and Soskice (2001) develop a broad theoretical framework to study the institutional complementarities that characterize different varieties of capitalism. Having a specific focus on the institutions of the political economy, the authors develop an actor-centered approach for understanding the institutional similarities and differences among the developed economies. The varieties of capitalism approach has inspired a large number of applications to the political economy field. To give some examples, Franzese (2001) and Höpner (2005) investigate the implications for industrial relations; Estevez-Abe et al. (2001) use the approach to analyze social protection; Fioretos (2001) considers political relationships, international

negotiations, and national interests; Hall and Gingerich (2009) study the relationship among labor relations, corporate governance, and rates of growth; Amable (2000) analyzes the implications of institutional complementarity for social systems of innovation and production.

In addition to institutional variety, the notion of institutional complementarity has also motivated studies on institutional change (Hall and Thelen 2009; Boyer 2005). In these works institutional complementarity is often presented as a conservative factor, which increases the stability of the institutional equilibrium (Pagano 2011). In the presence of institutional complementarity change requires the simultaneous variation of different institutional domains, which in turn demands high coordination among the actors involved. Sometime, institutions themselves can act as resources for new courses of action that (incrementally) favor change (Deeg and Jackson 2007).

Alongside contributions on the distinct models of capitalism, the concept institutional complementarity has found application also in other domain of analysis. Aoki (1994), for instance, studies the role of institutional complementarity in contingent governance models of teams. Siems and Deakin (2010) rely on an institutional complementarity approach to investigate differences in the business laws governing in various countries. Gagliardi (2009) argue in favor of an institutional complementarity relationship between local banking institutions and cooperative firms in Italy. Bonaccorsi and Thuma (2007), finally, use the idea of institutional complementarity to investigate inventive performance in nano science and technology.

References

- Amable B (2000) Institutional complementarity and diversity of social systems of innovation and production. *Rev Int Polit Econ* 7(4):645–687
- Aoki M (1994) The contingent governance of teams: analysis of institutional complementarity. *Int Econ Rev* 35(3):657–676
- Aoki M (2001) *Toward a comparative institutional analysis*. MIT Press, Cambridge

- Bonaccorsi A, Thoma G (2007) Institutional complementarity and inventive performance in nano science and technology. *Res Policy* 36(6):813–831
- Boyer R (2005) Coherence, diversity, and the evolution of capitalisms: the institutional complementarity hypothesis. *Evol Inst Econ Rev* 2(1):43–80
- Braverman H (1974) Labor and monopoly capital: the degradation of work in the twentieth century. Monthly Review Press, New York
- Crouch C, Streek W, Boyer R, Amable B, Hall P, Jackson G (2005) Dialogue on “Institutional complementarity and political economy”. *Soc Econ Rev* 3:359–382
- Deeg R, Jackson G (2007) Towards a more dynamic theory of capitalist variety. *Soc Econ Rev* 5(1): 149–179
- Estevez-Abe M, Iversen T, Soskice D (2001) Social protection and the formation of skills: a reinterpretation of the welfare state. In: Hall PA, Soskice D (eds) *Varieties of capitalism: the institutional foundations of comparative advantage*. Oxford University Press, New York, pp 145–183
- Fioretos O (2001) The domestic sources of multilateral preferences: varieties of capitalism in the European Community. In: Hall PA, Soskice D (eds) *Varieties of capitalism: the institutional foundations of comparative advantage*. Oxford University Press, New York, pp 213–244
- Franzese RJ (2001) Institutional and sectoral interactions in monetary policy and wage/price-bargaining. In: Hall PA, Soskice D (eds) *Varieties of capitalism: the institutional foundations of comparative advantage*. Oxford University Press, New York, pp 104–144
- Gagliardi F (2009) Financial development and the growth of cooperative firms. *Small Bus Econ* 32(4): 439–464
- Hall PA, Gingerich DW (2009) Varieties of capitalism and institutional complementarities in the political economy: an empirical analysis. *Br J Polit Sci* 39(3):449–482
- Hall PA, Soskice D (2001) *Varieties of capitalism: the institutional foundations of comparative advantage*. Oxford University Press, New York
- Hall PA, Thelen K (2009) Institutional change in varieties of capitalism. *Soc Econ Rev* 7(1):7–34
- Höpner M (2005) What connects industrial relations and corporate governance? Explaining institutional complementarity. *Soc Econ Rev* 3(1):331–358
- Landini F (2012) Technology, property rights and organizational diversity in the software industry. *Struct Chang Econ Dyn* 23(2):137–150
- Landini F (2013) Institutional change and information production. *J Inst Econ* 9(3):257–284
- Milgrom P, Roberts J (1990) Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica* 58(6):1255–1277
- Pagano U (1992) Organizational equilibria and production efficiency. *Metroeconomica* 43:227–246
- Pagano U (2011) Interlocking complementarities and institutional change. *J Inst Econ* 7(3):373–392
- Pagano U, Rossi MA (2004) Incomplete contracts, intellectual property and institutional complementarities. *Eur J Law Econ* 18(1):55–67
- Pagano U, Rowthorn R (1994) Ownership, technology and institutional stability. *Struct Chang Econ Dyn* 5(2):221–242
- Siems M, Deakin S (2010) Comparative law and finance: past, present, and future research. *J Inst Theor Econ* 166(1):120–140
- Williamson OE (1985) *The economic institutions of capitalism*. The Free Press, New York

Institutional Economics

Ringa Raudla

Faculty of Social Sciences, Ragnar Nurkse School of Innovation and Governance, Tallinn University of Technology, Tallinn, Estonia

Abstract

Institutional economics is interested in the interactions between institutions and the economy: how institutions influence the functioning, performance, and development of the economy and, in turn, how changes in the economy influence the institutions. Institutional economics studies the impact of institutions on economy, how institutions evolve, and how they could be improved. In furthering institutional economics, more extensive exchanges between the communities of institutional economics and law and economics would be fruitful. Those interested in institutional economics should certainly be more aware of developments in law and economics and utilize these insights in their further research – and vice versa.

Definition

Institutional economics is interested in the interactions between institutions and the economy: how institutions influence the functioning, performance, and development of the economy and, in turn, how changes in the economy influence the institutions. Institutional economics

studies the impact of institutions on economy, how institutions evolve, and how they could be improved.

Introduction

For readers from outside the discipline of economics, the adjective “institutional” in front of “economics” in the title of this entry may appear somewhat redundant. It would seem obvious that in order to understand what is going on in the economy, it is necessary to examine the institutions that the economy is embedded in. Thus, one could ask: why it is necessary to emphasize the term “institutions” in “institutional economics” given that “economics” should, logically, already include the analysis of “institutions”?

The reason for including the term “institutional” in front of “economics” is that neoclassical economics, which has constituted the economic mainstream since the mid-twentieth century, has paid only very limited attention to institutions. This has, at least partly, been due to the fact that economics came to be defined by its *method* (i.e., formalistic analysis of rational choice), rather than its *object* of analysis (i.e., the economy).

Fortunately, despite the lure of the mainstream economics in which the method was given prominence over the subject matter, a sufficient number of economists have been interested in real-life economies. In studying the *economy* as a subject matter, however, institutions cannot be ignored for too long before the analysis becomes stalled. Thus, in recent decades, institutions have returned to economic analysis, with an increasing number of economists paying attention to the role of institutions when explaining what is going on in the economy.

The law and economics movement has certainly played an important role in bringing “institutions” (like law) back into economics. Indeed, institutional economics and law and economics are closely related: they have similar intellectual roots, common pioneers, and overlapping research agenda. Law is, obviously, one of the most important “institutions” that are studied in institutional economics. Some would even say

that “law and economics” could be viewed as one of the subfields or “building blocks” of institutional economics.

What Is Institutional Economics?

Most broadly speaking, institutional economics is interested in the interactions between institutions and the economy: how institutions influence the functioning, performance, and development of the economy and, in turn, how changes in the economy influence institutions. Institutional economics studies the impact of institutions on the economy, how institutions evolve, and how they could be improved.

When studying the economic system, institutional economics assumes that “institutions matter” since institutions are the key element of any economy and should hence be placed at the center of analysis. Various definitions of the term “institution” have been offered by different authors. Broadly speaking, institutions can be defined as formal and informal rules, including their enforcement mechanisms. Douglass North, who can be considered as one of the “bridge-builders” between the old and new institutional economics, has defined institutions as “the rules of the game in a society” (1990, p. 3) or, more specifically, “humanly devised constraints that structure political, economic and social interaction” (1991, p. 97). According to North, institutions consist of both “informal constraints” (e.g., customs, traditions, and codes of conduct) and “formal rules” (e.g., constitutions and laws).

Some analysts (especially in the new institutionalist camp) have considered it useful to distinguish between different levels of analysis in examining institutions (see, e.g., Williamson 2000; North 1990): (1) the “highest” or so-called “social embeddedness” level entails norms, customs, and traditions; (2) the next level – the “institutional environment” – comprises of constitutions, laws, and political institutions; and (3) the lowest level of analysis looks at “institutional arrangements” or “governance arrangements” (e.g., vertical integration, franchising) and “organizations” (e.g., political parties, firms,

and trade unions). Institutional economics considers all these different levels of analysis as worthy pursuits (including interactions between the different levels), though admittedly, a lot more progress has been made on the third level of analysis than on the first two, where a lot of research still remains to be done.

Roots of Institutional Economics

The roots of institutional economics go back (at least) to the German historical school (represented, e.g., by Gustav von Schmoller, Wilhelm Roscher, Werner Sombart, and Max Weber) and the American institutionalist school (represented, e.g., by John Commons, Thorstein Veblen, and Wesley Mitchell) (for more detailed discussions, see, e.g., Medema et al. 1999; Rutherford 1994). The historical school dominated the discipline of economics in Germany from 1840s to 1940s, whereas the American institutionalist school had its peak during the interwar period. While there were important differences between these two schools and also between the individual thinkers within them, what they had in common was the focus on the role of institutions in shaping economic outcomes and their critique of the neoclassical approach to economic analysis (especially the formalistic aspects of it). The German historical school underscored the importance of sensitivity to specific cultural and historical circumstances in economic analysis (resulting in their emphasis on using empirical data to ground economic theories). The American institutionalist school emphasized the relevance of institutions (including the role of the laws and the state) in analyzing the economy. Commons (1924), for example, examined the legal underpinnings of the capitalist economic system; he demonstrated how evolutionary changes in the economic domain (e.g., with regard to what types of activities were deemed reasonable) facilitated specific changes in laws (e.g., the transformation of how the concept of property was legally defined) and how these changes, in turn, facilitated specific forms of economic activity. The American institutionalist school was also skeptical of the notion of the fixed preferences of individuals and emphasized that individual preferences can be shaped by

the institutions that surround them (e.g., via forming habits, as argued by Thorstein Veblen). In addition, the institutionalists criticized the static approach of the neoclassical economics and emphasized the evolutionary nature of economy (and hence the focus on change, technology, and innovation in economic analysis) (see, e.g., Veblen 1898).

In the postwar period, the popularity of the institutionalist approaches waned and economics became dominated by neoclassical economics. However, the institutionalist traditions were carried on by economists like Clarence Ayres, Karl Polanyi, John Kenneth Galbraith, Allan Gruchy, Simon Kuznets, Gunnar Myrdal, Ragnar Nurkse, Joseph Schumpeter, and others. In law and economics, the institutionalist traditions have been carried on by Allan Schmid, Warren Samuels, Nicolas Mercuro, and Steven Medema.

Different Approaches Within the “Modern” Institutional Economics

Until a decade or so ago, authors writing about “institutional economics” considered it necessary to distinguish between “old” and “new” institutional economics (for a more detailed discussion of the differences between these two camps, see, e.g., Rutherford 1994). “Old” institutionalism referred to researchers (e.g., Wendell Gordon, Allan Gruchy, Philip Klein, Marc Tool, Warren Samuels, Allan Schmid) following the traditions of John Commons, Thorstein Veblen, Wesley Mitchell, and others. The “new” institutional economics referred to the developments in economics that started in 1960s (led by Ronald Coase, Douglass North, and Oliver Williamson) when institutions were (at least to some extent) “brought back in” to the economic mainstream. Many authors in the emerging new institutional economics camp (e.g., Robert Sugden, Andrew Schotter, Mancur Olson, and Richard Posner) attempted to use the analytical tools of neoclassical theory to explain the emergence and impacts of institutions, with a specific attention to transaction costs, property rights, and contractual relations. In more recent years, however, the academics in the new institutionalist camp have increasingly moved away from the assumptions of neoclassical economics.

While one can still observe differences between the “old” and “new” camps, one can also talk more generally about (modern) institutional economics. In the light of the recent convergences between the camps (see, e.g., Dequech 2002), it would be helpful to offer a more *synthesized* view of what institutional economics is about. Thus, this entry tries to delineate the “common” ground of different institutional approaches by focusing on issues that many (if not most) researchers involved in doing research in institutional economics would agree are important.

Still, it is worth keeping in mind that institutional economics entails a rather diverse (and to some extent also conflicting) set of approaches. These different research streams vary with regard to the substantive questions studied and the methodology applied. Thus, institutional economics is not a single, unified, all-embracing, and well-integrated theory, proceeding from a set of common assumptions and hypotheses. Instead, it consists of different “building blocks,” coming from different traditions. Furthermore, it is worth emphasizing that institutional economics is an openly interdisciplinary endeavor, which draws on other disciplines (like sociology, psychology, anthropology, history, political science, public administration, and law) in order to understand and explain the role of institutions in economic life.

Differences Between Institutional Economics and Neoclassical Economics

Although at least some of the topics that used to be the playground of institutional economics have gradually found their way into the economic mainstream and there is a growing consensus about the importance of institutions in influencing economic growth (see, e.g., Acemoglu et al. 2005), it is still too early to say that “we are all institutionalists now.” Hence, a few remarks on how the *institutional* approach in economics differs from the *neoclassical* approach are still necessary. It is worth keeping in mind, though, that the different institutionalist camps differ somewhat with regard to their “distance” from the orthodox neoclassical economics: those who are closer to the “old” institutionalist

traditions are further removed from the neoclassical assumptions than those in the “new” institutionalist camp.

In sum, the differences between the (mainstream) neoclassical economics and institutional economics are the following (for a more systematic comparison, see, e.g., Hodgson 1988; Medema et al. 1999):

First, while neoclassical economics proceeds from the assumption of “rational” individuals who maximize their utility (*homo oeconomicus*), the institutionalist approach takes a more realistic view of individual behavior: it regards individuals as being purposeful but only *boundedly rational*. Unlike orthodox economics, institutional economics emphasizes the importance of severe information problems (including uncertainty about the future) and the costs involved in obtaining necessary information. Proponents of institutional economics have hence criticized the orthodox approaches for simply “assuming away” the information problems and “assuming” perfect knowledge.

Second, in contrast to the neoclassical approach of treating the tastes and preferences of individuals as “given” or “fixed” (at least for the purposes of analysis), most institutional economists view the individuals as “social beings” and hence consider it necessary to proceed from the assumption that preferences are *malleable* and that *changes* in preferences should be analyzed as well, including the role of institutions in molding individual preferences and purposes (via changes in habits, as argued by Hodgson 1988, 1998).

Third, while neoclassical economics is concerned with states of *equilibria* and equilibrium-oriented theorizing (focusing on “mechanistic” optimization under static constraints), most institutional economists prefer to take a more evolutionary view of economic phenomena and also institutions. They emphasize that economic development has an *evolutionary* nature and hence prefer dynamic modes of theorizing, with a focus on longer-run processes of continuity and change, entailing path dependencies, transformations, and learning over time. Institutional economics is also interested in the

evolutionary nature of the interactions between institutions and the economy. Further, while the neoclassical economics treats technology as “given” (or exogenous), institutional economics emphasizes the importance of examining the role of technological changes (and their interactions with institutions) in economic development.

Fourth, while neoclassical economics tends to treat the use of *power* as given (and accept the power structure as it is), at least some institutional economists (especially those with closer ties to the “old” institutional economics) are concerned with how power is actually deployed in the economic, political, and societal settings. Power is deemed important because power relations influence who gets to shape the institutions (including legal rules), whose values dominate, and whose “interests” are to be regarded as “rights.” The allocation of rights, in turn, would influence the distribution of power in society (see, e.g., Acemoglu et al. 2005; Furubotn and Richter 1997; Medema et al. 1999).

Why and How Do Institutions Matter?

Most generally speaking, institutions “matter” for the economy because the structures entailed in the institutions influence the behavior of the economic actors, which in turn influences the functioning and performance of the economy. The influence from the “institutions” to the “economy,” however, is not unidirectional: changes in the economy can bring about changes in institutions as well and, hence, it would be more accurate to talk about mutual interactions between institutions and the economy (see, e.g., Medema et al. 1999). In other words, we should not take a deterministic view of institutions according to which institutions always *determine* the actions of individuals: the causal arrow can go in the other direction as well. “Actors and structure, although distinct, are thus connected in a circle of mutual interaction and interdependence” (Hodgson 1998, p. 181).

How do institutions influence the behavior of economic actors? There are different ways to answer the question. The answer closest to mainstream economics is that institutions shape the

“choice set” available to the economic actors and “structure the incentives” of the actors (hence making a certain course of action more attractive than other courses) and thus steer individual behavior via affecting the costs and benefits associated with different types of actions (including engaging in different types of economic activities) (see, e.g., Acemoglu et al. 2005; Eggertsson 1990; Furubotn and Richter 1997; North 1991). Other answers point to the more “sociological” and deeper “psychological” mechanisms and emphasize role of institutions in shaping the habits and, through that, also the preferences of individuals, which, in turn, would influence their choices and actions (see, e.g., Hodgson 1988, 1998).

At the most basic level – and this is something that all institutionalists agree with – institutions influence the interactions of economic actors by providing “order” and reducing uncertainty in exchange. Given that institutions outline the “rules of the game” (which provide boundaries, constraints, and patterns for behavior), they provide economic actors with information about the potential behavior of *other* actors and hence help to establish baseline conditions for interactions between economic agents. Hence, institutions allow individuals to make reliable predictions about what *other* economic agents are likely to do in any given circumstance, which allows them to proceed with decision-making, negotiations, and exchange with at least some level of certainty (see, e.g., Hodgson 1988; Kasper and Streit 1998; Nelson and Sampat 2001; North 1991). Thus, institutions can both *constrain* and *enable* individual actions, e.g., by providing information about the behavior of others, defining pathways for doing things, allowing coordinated actions, and limiting opportunistic and arbitrary behavior.

Many institutional economists have emphasized that institutions influence the size of *transactions costs* associated with exchange relations (see, e.g., Furubotn and Richter 1997; Kasper and Streit 1998; North 1991). Transaction costs involve different cost associated with exchange processes, including search and information costs (e.g., discovering what one wants to buy, who the sellers are, and what the prices are),

bargaining and decision costs (e.g., associated with drawing up a contract), and policing and enforcement costs (see, e.g., Coase 1960; Furubotn and Richter 1997; North 1991). For example, provisions of contract law can help to lower the costs of concluding and enforcing contracts (e.g., the possibility to turn to courts in case of a breach reduces the need undertake “private” safeguarding measures by the parties themselves) and to hence facilitate impersonal contracting between strangers.

It has to be emphasized, however, that the institutions that have evolved in any given country do not necessarily *guarantee* a well-functioning economy and fast economic development. The institutions that have emerged can also be highly inefficient and entail elements that clearly hinder technological advances and economic development and growth (Freeman and Perez 1988; North 1990, 1991; Nelson and Sampat 2001).

Some Examples of How Institutions Influence Economic Performance

While lot of research still needs to be undertaken in order to achieve fuller understanding of the role of institutions in economic development, a number of insightful studies have been conducted so far. A complete overview of these achievements is certainly beyond the scope of this entry. Thus, the examples below constitute only a small portion of the body of research in institutional economics and should be viewed as indicative rather than exhaustive.

Markets as Institutions

Mainstream economics tends to treat the market as some sort of a “natural” feature of a social domain, an aggregate of individual bargains, resulting from free exchange between economic agents – almost as “an ether in which individual and subjective preferences relate to each other, leading to the physical exchange of goods and services,” independent of institutions (Hodgson 1988, p. 178). In contrast, for most institutional economists, the market should be conceived of as

an institution (or a set of institutions), involving “social norms, customs, instituted exchange relations and – sometimes consciously organized – information networks” (Hodgson 1998, p. 181). Institutional economics emphasizes that market institutions (and the institutions the market is embedded in) can play an important role in lowering transaction costs and hence facilitate more exchange relations.

Institutionalist research has also examined the role of the state in creating (what mainstream economists call) “free markets.” Karl Polanyi (1944), for example, shows, in his study of the Industrial Revolution in Great Britain, that the development of free markets in the eighteenth and nineteenth centuries actually involved a significant increase in the activities of the government: more legislation was called for and more administrators needed to monitor and safeguard the free working of the market system. In other words, the creation of “free markets” implied an increase in the control, regulations, and intervention by the state. The same applies today: “every successful market economy is overseen by a panoply of regulatory institutions, regulating conduct in goods, services, labor, assets, and financial markets” (Rodrik 2000, p. 7). The “freer” the markets, the greater the vigilance that may be required from the regulatory institutions (e.g., in the field of antitrust, financial regulation, securities legislation, etc.) (ibid).

In the light of these findings, some institutionalists emphasize that the dichotomy between regulation vs deregulation (or intervention vs nonintervention or “more” vs “less government”) is false. Instead, one should ask which *type* of regulation and intervention the state is engaged in (and for what ends) and whose “interests” are protected as “rights” by the state (Hodgson 1988; Medema et al. 1999). For example, if the government relaxes regulations pertaining to workplace safety, it expands the set of rights of employers and narrows the set of rights of employees – and vice versa when these regulations are toughened. In either case, the government is “present” – via the legal framework and the mechanisms of enforcement (Medema et al. 1999).

Property Rights

An important set of institutions that has captured the attention of many institutional economists – both in the “old” and “new” camps, from Commons (1924) to North (1990) – involves *property rights*. Again, while neoclassical economics takes property rights as “given” (i.e., perfectly defined), institutional economists take a much closer look at the actual definition, delineation, allocation, and enforcement of property rights and how these influence exchange relations and other economic activities. Institutionalists from different traditions agree that the specific content of property rights (e.g., control rights over assets or resources) and the way they are enforced influence the allocation and use of resources. As Rodrik (2000, p. 6) puts it, “an entrepreneur would not have an incentive to accumulate and innovate unless s/he has adequate control over the return of the assets that are thereby produced or improved.” Thus, for example, when property rights are not credibly secured (e.g., when there is a threat of expropriation by the government or unilateral seizure by another private actor), entrepreneurs are less likely to adjust (efficiently) to changes in technology, to invest (e.g., in research and development, which facilitates technological change), and to innovate. In contrast, secure property rights encourage firms to make higher value-added investments with longer-term time horizons (Keefer and Knack 1997).

Institutional economists have also examined the role of the *state* in protecting property rights. On the one hand, it is emphasized that protection of individual property entails legal structures for recognizing, adjudicating, and enforcing these rights, which can be provided by the state (e.g., Sened 1997). On the other hand, it is argued (especially by the new institutionalists) that the state can also pose a danger to private property through expropriation (Furubotn and Richter 1997). Bringing these two arguments together, Douglass North (e.g., 1990, 1991) has argued that economic development is fostered by an institutional environment in which the state is sufficiently strong to protect the private parties from seizing each others’ property but can at the same time make a credible commitment not to

expropriate the very same property it is defending and securing. As Hodgson (2004) puts it, “For private property to be relatively secure, a particular form of state had to emerge, countered by powerful and multiple interest groups in civil society. This meant a pluralistic state with some separation of powers, backed up by a plurality of group interests in the community at large.” In his empirical studies, Douglass North has argued that the establishment of clear and secure property rights played a major role in the economic development and rise of the “West.” Establishing secure property rights (with a credible commitment by the state to respect and secure them) allowed, for example, the emergence of capital markets and the employment of technology necessary for industrial production (see, e.g., North 1990, 1991). Many other studies (e.g., Acemoglu et al. 2005; Keefer and Knack 1997) have confirmed that finding.

Another discussion concerning property rights pertains to the question of whether the policy instrument of more extensive allocation of property rights implies “more” state or “less” state. Some economists in the new institutionalist camp (e.g., Demsetz, Alchian) regard clearer definition and allocation of property rights (especially the extension of private property) as “solutions” to different types of market failures, hence allowing the “lessening” of the need for government intervention. Hodgson (1988, p. 152), in contrast, has pointed out that by expanding the domain of formal property rights, the state still remains engaged, but in a different way – through the extension of litigational activity. Chang (2007), among others, has also warned us of the dangers of using the institutional prescription of “private property rights” as the main “solution” for guaranteeing economic development (see also Rodrik 2000). Indeed, a whole stream of research examines the conditions under which different types of ownership – private property, common property, state property, and various hybrid forms (like the township and village enterprises in China) – lead to optimal use of resources. Elinor Ostrom (e.g., 1990), for example, has shown that in the case of common-pool resources, common property can (when

combined with suitable institutional arrangements) lead to a better use of natural resources (e.g., fish stock, forests, water) than either privatization or nationalization.

Political Institutions and Bureaucracy

One of the building blocks in the institutionalist literature looks at the impacts of specific features of *political institutions* (including constitutions) on economic performance. The starting point for many of these studies is that the balance and separation of powers and the number and power of veto players are likely to influence the general character of policy action (including the levels of decisiveness and credibility) and also specific features of policies and laws the governments adopt, which, in turn, can influence economic performance. For example, it has been examined how the level of democracy (and the level of inclusiveness in governance) but also specific constitutional features – like government type (presidential vs parliamentary), electoral rules (e.g., plurality vs proportional), the organization of the judiciary, and vertical separation of powers – influence economic performance (see, e.g., Acemoglu et al. 2005; Persson and Tabellini 2003; Rodrik 2000).

Yet another stream examines the role of administrative structure, public administration, and the characteristics of bureaucracy in the economic growth and development. Evans and Rauch (1999) show that economic growth is higher in those developing countries where the bureaucracies entail more Weberian elements (e.g., recruitment based on merit and predictable long-term career paths). They argue (drawing, e.g., on Weber [1904–1911] 1968 and also Polanyi 1944) that merit-based recruitment and long-term careers facilitate higher competence of public administrators, lower levels of corruption, and long-term orientation. These factors, in turn, facilitate the design and implementation of policies that can help to promote growth, e.g., the provision of long-term (public) investments that complement those made in the private sector and helping private sector actors to overcome coordination and information problems (see also Rodrik 2000; Wade 1990).

Concluding Remarks

Despite an increasing number of studies examining the links between institutional setting and economic performance, we are only at the beginning of the journey to understand the interrelations involved and which institutions constitute a “good fit” in different countries and contexts. As Chang (2007, p. 3) puts it: “We are still some way away from knowing exactly which institutions in exactly which norms are necessary, or least useful, for economic development in which contexts.”

As the experience of transition, emerging, and developing economies has demonstrated, in further studies it would be necessary to explore the effects of different *configurations* of (complementary) institutions rather than examining the effects of specific institutions (e.g., the establishment of private property rights or the adoption of a new contract law) in isolation.

Also, the relationships between informal and formal institutions still need to be examined further. It is clear that the effectiveness of “formal” institutions (e.g., laws and regulations) depends on whether they fit sufficiently well with the “informal” institutions (like norms and customs), which in turn influences, for example, how well institutional transplants (from one country to another or implementing the “best practice” blueprints) can work. However, we are still far from completely understanding how informal forms emerge and persist, how such informal norms interact with formal norms, and how that, in turn, influences specific economic activities in a given country.

Finally, we still have only limited knowledge of how institutions conducive to economic development in specific contexts can be “built.” These are all important arguments for undertaking more in-depth qualitative and comparative studies in the field.

In furthering institutional economics, more extensive exchanges between the communities of institutional economics and law and economics would be fruitful. Those interested in institutional economics should certainly be more aware of developments in law and economics and utilize these insights in their further research – and vice versa.

References

- Acemoglu D, Johnson S, Robinson JA (2005) Institutions as a fundamental cause of long-run growth. In: Aghion P, Durlauf SN (eds) *Handbook of economic growth*, vol 1A. Elsevier, Amsterdam, pp 385–472
- Chang H-J (ed) (2007) *Institutional change and economic development*. United Nations University Press, Tokyo
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Commons JR (1924) *Legal foundations of capitalism*. Macmillan, New York
- Dequech D (2002) The demarcation between the “old” and the “new” institutional economics: recent complications. *J Econ Issues* 36(2):565–572
- Eggertsson T (1990) *Economic behavior and institutions: principles of neoinstitutional economics*. Cambridge University Press, Cambridge
- Evans P, Rauch JE (1999) Bureaucracy and growth: a cross-national analysis of the effects of “Weberian” state structures on economic growth. *Am Sociol Rev* 65(5):748–765
- Freeman C, Perez C (1988) Structural crises of adjustment, business cycles, and investment behavior. In: Dosi G (ed) *Technical change and economic theory*. Pinter Press, London, pp 38–66
- Furubotn EG, Richter R (1997) *Institutions and economic theory: the contribution of the new institutional economics*. The University of Michigan Press, Ann Arbor
- Hodgson GM (1988) *Economics and institutions: a manifesto for a modern institutional economics*. University of Pennsylvania Press, Philadelphia
- Hodgson GM (1998) The approach of institutional economics. *J Econ Lit* 36(1):166–192
- Hodgson GM (2004) Institutional economics: from Menger and Veblen to Coase and North. In: Davis JB, Marciano A, Runde J. *The Elgar companion to economics and philosophy*, pp 84–101
- Kasper W, Streit ME (1998) *Institutional economics: social order and public policy*. Edward Elgar, Cheltenham
- Keefer P, Knack S (1997) Why don’t poor countries catch up? A cross-national test of an institutional explanation. *Econ Inq* 35(3):590–602
- Medema SG, Mercuro N, Samuels WJ (1999) Institutional law and economics. In: Boukaert B, De Geest G (eds) *Encyclopaedia of law and economics*. Edward Elgar/The University of Ghent, Cheltenham
- Nelson RR, Sampat BN (2001) Making sense of institutions as a factor shaping economic performance. *Revista Econ Inst* 3(5):17–51
- North DC (1990) *Institutions, institutional change and economic performance*. Cambridge University Press, Cambridge
- North DC (1991) Institutions. *J Econ Perspect* 5(1):97–112
- Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. University Press, Cambridge
- Persson T, Tabellini G (2003) *Economic effects of constitutions*. MIT Press, Cambridge
- Polanyi K (1944) *The great transformation*. Rinehart, New York
- Rodrik D (2000) Institutions for high-quality growth: what they are and how to acquire them. *Stud Comp Int Dev* 35(3):3–31
- Rutherford M (1994) *Institutions in economics: the old and new institutionalism*. Cambridge University Press, Cambridge
- Sened I (1997) *The political institution of private property*. Cambridge University Press, Cambridge
- Veblen TB (1898) Why is economics not an evolutionary science? *Q J Econ* 12:373–397
- Wade R (1990) *Governing the market: economic theory and the role of government in East Asian industrialization*. Princeton University Press, Princeton
- Weber M ([1904–1911] 1968) *Economy and society*. Bedminster, New York
- Williamson OE (2000) The new institutional economics: taking stock, looking ahead. *J Econ Lit* 38:595–613

Institutional Review Board

Roberto Ippoliti

Department of Management, University of Turin, Turin, Italy

Scientific Promotion, Hospital of Alessandria – “SS. Antonio e Biagio e Cesare Arrigo”, Alessandria, Italy

Abstract

Institutional Review Board (IRB) are independent institutions aimed at approving, monitoring and reviewing human experimentations in order to protect research subjects’ rights from the necessity to increase the current medical knowledge.

After a brief historical introduction of humans experimentation, this section presents American and European regulation of this specific institution, as well the main related issues in Law and Economics.

Synonyms

Ethical review board; Independent ethics committee

Definition

Institutional Review Board (IRB) is a committee designated to approve, monitor, and review human experimentations in order to protect research subjects' rights.

IRB and its Evolution

Institutional Review Board (IRB), also known as independent ethics committee or ethical review board, is a committee designated to approve, monitor, and review human experimentations in order to protect research subjects' rights.

The matter of research subjects' rights was first addressed after the end of World War II, when Nazi experiments on prisoners and Jews were discovered. Those researches were motivated by racial and political conflict and the idea of rights for those kept in Nazi concentration camps held no meaning. The subjects were coerced into participating and there was no informed consent about potential related expected and unexpected adverse events (e.g., their death). The Nazi experiments were aimed at supporting the Nazi racial ideology, as well as at improving the survival and rescue of German troops, by testing medical procedures and pharmaceuticals.

After that experience, society felt that it was its duty to prevent research involving people who were used as test subjects not of their own free will and to check the scientific aims of research studies. Thus, The Nuremberg Code of Ethical Human Subjects Research Conduct (1947) was drafted. This was the first international document on human experimentation and research subjects which highlighted the main right of patients: voluntary consent of human subjects involved in clinical trials. Obviously, this information concerns risk-benefit outcomes of experiments, i.e., both expected effectiveness and potential related expected/unexpected adverse events. However, other examples of inhuman clinical research were necessary to develop a systematic review of these studies by independent institutions, for

example, the *Tuskegee syphilis experiment*, which was a clinical study conducted by the US Public Health Service to study the natural progression of untreated syphilis, or the *Vipeholm experiments*, which was sponsored by the sugar industry and dentist community, in order to determine whether carbohydrates affected the formation of cavities. In the former study, research subjects were rural American men who thought they were receiving free health care from the US government, even if they were never told they had syphilis nor were they ever treated for it (see Thomas and Quinn (1991)). In the latter case, patients of a Swedish mental hospital were fed large amounts of sweets to provoke dental caries, violating the basics of medical ethics (see Gustafsson et al. 1954).

Over the years, that Code has been followed by other international documents, among which one of the most important was the Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects (1964). This international document was developed by the World Medical Association in 1964 as a means of governing international clinical research. It is mainly a set of guidelines for medical doctors conducting biomedical research involving human subjects, and it includes the principle that research protocols should be reviewed ex ante by an independent committee. According to its recommendations, a technical board must review each clinical trial before enrollment can start. In other words, a third party has the duty to guarantee the main principles established by the aforementioned international code: patient information and scientific validity of protocols. Another international guideline is proposed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), which is a project that brings together the regulatory authorities of Europe, Japan, and the USA, as well as experts from the pharmaceutical industry, to discuss scientific and technical aspects of pharmaceutical product registration. ICH provides good clinical practice (GCP), i.e., an international quality standard that governments can transpose

into regulations for clinical trials involving human subjects (ICH Topic E 6 (2002)). According to ICH-GCP, an IRB should safeguard the rights, safety, and well-being of research subjects, with a particular attention to trials that may include vulnerable subjects (e.g., pregnant women, children, prisoners). A European standard for clinical trials and patient protection is the ISO 14155, which is valid in the European Union as a harmonized standard good clinical practice (i.e., ISO-GCP).

According to international documents and guidelines, both in Europe and in the USA, a protection system has been developed through the years, based on this Institutional Review Board (IRB). In the United States (USA), Title 45 of the Code of Federal Regulations, part 46 (revised 2009) – which is the reference regulation for US IRB – identifies this board as an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects, with the authority to approve or disapprove research. The purpose of IRB review is to assure that appropriate steps are taken to protect the rights and welfare of humans involved as subjects in the clinical trials, reviewing the research protocol and related materials (e.g., informed consent documents and investigator brochures). According to Directive 2001/20/EC (2001), the European Union (EU) recognizes this board as an independent body in a member state, consisting of health-care professionals and nonmedical members, whose responsibility is to protect the rights, safety, and well-being of human subjects involved in a trial and to provide public assurance of that protection, by, among other things, expressing an opinion on the trial protocol, on the suitability of the investigators and the adequacy of facilities, and on the methods and documents to be used to inform trial subjects and obtain their informed consent.

Considering their composition, FDA's requirements is set in Title 21 of the Code of Federal Regulations, part 56 (revised 2013) – which is an additional regulation for US IRB that oversee clinical trials involved in new drug applications – suggesting that an IRB must have at least five

members with enough experience, expertise, and diversity to make an informed decision on whether the research is ethical, informed consent is sufficient, and appropriate safeguards have been put in place. Moreover, the regulation states that if a study that includes vulnerable populations is under investigation, the IRB should have members who are familiar with these groups (e.g., an IRB has to include an advocate for prisoners when considering research that involves them). The European Directive 2001/20/EC (2001) does not specify the composition of each IRB, which is regulated at the national level by each member state.

Several law and economics issues are related to IRBs and their activity. Just to cite a few, Calabresi (1969) focused on the generational conflict, suggesting that IRB should be an expression of the value of research that involves human subjects and how it is necessary to achieve an adequate balancing of present against future lives. Other researchers focus on the effectiveness of the IRB decision-making process, highlighting how these boards have exercised primary oversight responsibility for human research subject, continuing to be often incapable of reviewing complex research protocols effectively (Hoffman 2001; Coleman 2004). Finally, proposing the ideal “Market of Human Experimentation,” Ippoliti (2013a, b) analyzes the relation between IRB activity and the transaction costs to achieve an exchange of information for innovation, suggesting the main impact on countries' competitiveness.

Another interesting topic related to the IRB members, which should be deeply analyzed from a law and economic prospective, concerns the undue influence of pharmaceutical companies and the related economic interests involved in the ethical decisions. Just to give an idea of the issue, Campbell et al. (2006) analyze the relation between IRB members and industry, suggesting that 36 % of these members had had at least one relationship with for-profit institutions in the past year. Moreover, authors denote that of the respondents, 85.5 % said they never thought that the relationships that another IRB member had with the industry affected his or her

IRB-related decisions in an inappropriate way. The economic undue influence is even more important considering the US market, where some IRB reviews are conducted by *for-profit* organizations – i.e., commercial IRBs (Emanuel et al. 2006).

Cross-References

► [Human Experimentation](#)

References

- American regulation: title 45 of the Code of Federal Regulations – part 46 (2009) Title 21 of the Code of Federal Regulations– part 56 (revised 2013)
- Calabresi G (1969) Reflection on medical experimentation in human. *Daedalus* 98(2):387–405
- Campbell EG, Weissman JS, Vogeli C, Clarridge BR, Abraham M, Marder JE, Koski G (2006) Financial relationships between institutional review board members and industry. *N Engl J Med* 355:2321–2329. <https://doi.org/10.1056/NEJMSa061457>
- Coleman CH (2004) Rationalizing risk assessment in human subject research. *Arizona Law Rev* 46(1):1–51
- Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects (1964)
- European Union regulation: Directive 2001/20/EC (2001)
- Emanuel EJ, Lemmens T, Elliot C (2006) Should society allow research ethics boards to be run as for-profit enterprises? *PLoS Med* 3(7):e309. <https://doi.org/10.1371/journal.pmed.0030309>
- Gustafsson BE, Quensel CE, Lanke LS, Lundqvist C, Grahnen H, Bonow BE, Krasse B (1954) The Vipeholm dental caries study; the effect of different levels of carbohydrate intake on caries activity in 436 individuals observed for five years. *Acta Odontol Scand* 11(3–4):232–264
- Hoffman S (2001) Continued concern: human subject protection, the institutional review board, and continuing review. *Tenn Law Rev*. 2001 Summer 68(4):725–70
- ICH Topic E 6 (R1) Guideline for Good Clinical Practice (2002)
- Ippoliti R (2013a) Economic efficiency of countries' clinical review processes and competitiveness on the market of human experimentation. *Value Health* 16:148–154
- Ippoliti R (2013b) The market of human experimentation. *Eur J Law Econ* 35(1):61–85
- The Nuremberg Code of Ethical Human Subjects Research Conduct (1947)
- Thomas SB, Quinn SC (1991) The Tuskegee syphilis study, 1932–1972: implications for HIV education and AIDS risk programs in the black community. *Am J Public Health* 81:1503

Insurance Market Failures

Donatella Porrini

Department of Management, Economics, Mathematics and Statistics, University of Salento, Lecce, Italy

Abstract

Insurance market is characterized by failures that impose particular negative consequences; given the failures, different remedies may improve the market outcome. On one hand, the insurance market is characterized by asymmetric information, i.e. moral hazard and adverse selection, and to correct the consequent severe market failures, monitoring and risk classification can be implemented. On the other hand, the insolvency issue: given the enormous amounts of funds in the hands of insurance companies, their default would have an extreme impact, and regulation is necessary to guarantee the payback for policyholders and beneficiaries.

Definition

Insurance is an instrument to give protection against the risk resulting from various perils or hazards, such as health risks, invalidity or death, accidents, unemployment, theft, fire, and many more. Contracts are offered by insurance companies providing risk-sharing mechanisms that allow their customers to replace these risks. Insurance market is characterized by failures that impose particular negative consequences on one or both market sides: given the failures, different remedies may improve the market outcome.

Introduction

Insurance plays three economic functions: (i) the transfer of risk from a risk-averse individual to the risk-neutral insurer, (ii) the pooling of risk so that the “uncertainty” of each insured becomes the

“certainty” of the insurance companies that this risk will occur to a percentage of their customers, and (iii) the allocation of risk for which each insured pays a price that should reflect the risk he contributes (Abraham 1995).

For these three reasons, insurance contracts increase social welfare while at the same time induce people to have a preventive behavior and contribute to internalize damage. At a macro-economic level, decreasing the economic effects of risks, insurance encourages companies to operate in risky sector and to make investments that they would not make otherwise. Meanwhile, life insurance plays a role as a long-term investment and savings instrument (Shavell 2000).

Asymmetric Information

The insurance market is characterized by the fundamental problem of bilateral asymmetric information. On one hand, individuals do not have complete information in understanding complicated insurance contracts and lack the ability to assess the adequacy of premium to risk. On the other hand, insurers suffer from lack of information regarding the risk characteristics of individuals. This second asymmetry generates the two phenomena of moral hazard and adverse selection.

Moral hazard (hidden action) depends on the insurers’ impossibility to perfectly know the extent their customers’ behavior may affect the occurrence and/or the dimension of the loss.

To be precise, the term moral hazard refers to at least two different situations in which the insured’s behavior can affect the probability of the various outcomes: (i) situations when insurance may induce greater use of a service by an insured individual or cause the insured to exercise less care and (ii) situations when an insured individual purposely causes harm or otherwise falsifies loss in order to collect insurance benefits or to inflate the loss.

In the case of moral hazard, the insured’s behavior changes, and the insurer is unable to either predict this change in advance or to prevent

it by exempting such behavior from the insurance contract coverage (Shavell 1979).

In fact, after signing an insurance contract, the insured may have less incentive to act carefully or take preventive measures, influencing both the damage probability and/or the loss dimension.

Moral hazard, even if severe, does not cause a complete breakdown of markets, but it raises the cost of insurance and, consequently, reduces the degree of insurance coverage negatively affecting market outcomes (Tennyson and Warfel 2009).

In this case, a remedy is monitoring the insured’s behavior: *ex ante* the loss occurrence, to monitor the level of care in preventing the loss, insofar as the insured’s behavior has any influence over the risk; *ex post*, to monitor the amount of claim when loss occurs, beyond the services the claimant would purchase if not insured and assuming the insured individual can influence the magnitude of the claim. Also incentive schemes linking the price of insurance to observed past behavior (e.g., bonus/malus systems) can be used as devices to contain this problem (Derrig 2002).

Adverse selection (hidden information) refers to the inability of insurers to observe risk characteristics of their customers, leading to offer a contract based on the average risk of the entire group of customers. In this case, more high-risk individuals purchase insurance; higher payouts by insurance companies force them to raise rates which, in turn, makes the insurance less attractive to low-risk individuals.

As a consequence, this may reduce the stability of the market equilibrium, and the market may completely break down, such as the famous “market for lemons” (Akerlof 1970).

In the case of adverse selection, a remedy would be to use statistical data to separate different categories of risky individuals by classification instrument.

Theoretically in determining the premium to be charged, insurers should estimate the expected losses for each individual being insured. But given the informational asymmetry, the insurance companies apply risk classification trying to group the individuals in such a way that those with a similar loss probability are charged the

same rate. The risk classification systems are clearly supported by statistical data showing differences in the event rate in different groups (Porrini 2015).

In practice, insurers have to identify risks that are independent, uncorrelated, and equally valued and to aggregate them in order to reduce the total risk of the set. An efficient risk classification reduces adverse selection, because it makes insurance more attractive to the low-risk individuals.

Classifying insurance risks increases the efficiency of contracting in terms of asymmetric information. However, the benefits are conditional on general principles, such as the non-discrimination, and generally to consumer protection issue.

Insolvency

Generally, the default of a company generates economic damages, first to shareholders, but in many cases also to the customers. Particularly, in insurance market, the insured individuals may lose future benefits and insurance coverage with possibly precarious economic situation in many cases and imposing to rely on a public coverage of these losses.

Moreover, this is reinforced by the consequences of the so-called inversion of the production cycle that comes from the fact that insurance services are only delivered after they are purchased and in many cases years later. This creates the necessity to monitor the financial condition and solvency of insurance companies over an extended period of time.

This feature leads potentially to insufficient capitalization and suboptimal solvency levels, by giving insurers scope for hiding poor underwriting and under-reserving, and these are motivations for government interventions aimed at monitoring to improve management discipline.

Regulation for solvency dates to the nineteenth century, when insurance insolvencies in the USA and Europe led to the establishment of state regulatory authorities. Given the social role and the involvement of insurance in the systemic risk issue (Faure and Hartlief 2003), solvency

becomes the primary focus of insurance regulation worldwide. Regulatory tools include risk-based capital requirements, electronic auditing of accounts, and a wide variety of limits on the ways that companies can invest the funds held in reserve to pay claims.

Moreover, regulation can be used to steer capital into preferred fields, given that insurance is an institution for storing and accumulating capital, competing with banking and securities firms. Although banking, insurance, and securities have traditionally been subject to different regulatory regimes, there is recently a “convergence” in the financial services marketplace that places great strain on the existing regulatory institutions (i.e., the diffusion of the business model of bank insurance).

The most common instrument of regulation for solvency consists of technical provisions that correspond to the amount required by the insurer to fulfil its insurance obligations and settle all commitments to policyholders arising over the lifetime of the portfolio. Technical provisions can be divided into those that cover claims from insurance events which have already taken place at the date of reporting and those that should cover losses from insurance events which will take place in the future.

Most countries supplement the above requirements by regulating the portfolio choices of insurance firms with the aim to ensure that insurers invest and hold adequate and appropriate assets to cover capital requirements and technical provisions.

The main focus of numerous national regulation is to ensure that insurance companies are able to honor their payment obligations in a continuous way, the most important instrument being that of a generalized capital and reserve rules, possibly supplemented by additional supervisory rules, such as investment restrictions and provisions of regular inspection by supervisory authorities.

Conclusion

Because the insurance business has become highly important to society's development, it is

relevant to find remedies to the failures that can impede a correct functioning of the market.

On one hand, the insurance market is characterized by fundamental problems of asymmetric information. Moral hazard and adverse selection play a central role in the insurance market, and to correct the consequent severe market failures, monitoring and risk classification can be implemented.

On the other hand, the insolvency issue is justified by the enormous amounts of funds and investments in the hands of insurance companies; as such, their default would have an extreme impact, and regulation is necessary to guarantee the payback for policyholders and beneficiaries.

Reference

- Abraham K (1995) Efficiency and fairness in insurance risk classification. *Virginia Law Rev* 71:403–451
- Akerlof GA (1970) The market for lemons: quality uncertainty and the market mechanism. *Q J Econ* 84:488–500
- Derrig RA (2002) Insurance Fraud. *J Risk Insu* 69:271–287
- Faure M, Hartlief T (2003) Insurance and expanding systemic risks. OECD, Paris
- Porrini D (2015) Risk classification efficiency and the insurance market regulation. *Risks* 4:445–454
- Shavell S (1979) On moral hazard and insurance. *Q J Econ* 93:541–562
- Shavell S (2000) On the social function and regulation of liability insurance. *Geneva Pap Risk Insur* 25:166–179
- Tennyson S, Warfel WJ (2009) Law and economics of first party insurance bad faith liability. *Conn Insur Law J* 16(1):203–242

Intellectual Property: Economic Justification

Dennis W. K. Khong
Centre for Law and Technology, Faculty of Law,
Multimedia University, Melaka, Malaysia

Definition

Economic justification for intellectual property means the economic reason for establishing and

supporting a system of intellectual property laws. It begins with the characteristics of information as public goods (Arrow 1962): non-rivalry in consumption and non-excludability. Intellectual property rights confer upon the right holders an exclusive right to legally compel users to buy a genuine product from the right holders or to obtain a license from them. Intellectual property rights are to prevent a market failure due to free-riding activities of non-payers. In general, intellectual property rights can be grouped into two families according to the function of the information therein: information as a good and information as a signal. Examples of intellectual property rights in the form of information as a good are copyright, patents, industrial designs, layout-designs of integrated circuits, and confidential information, while examples of intellectual proprietary rights in the form of information as a signal are trademarks, the tort of passing off and unfair competition, and geographical indications.

Introduction

This essay examines the economic justification for intellectual property. The economic justification begins with the twin characteristics of information as public goods: non-rivalry in consumption and non-excludability. Market failure of free riding results from the non-excludability of public goods. Granting intellectual property rights is a solution to solving this public goods problem by conferring upon the creators of new information an exclusive right. In general, intellectual property rights can be grouped into two families according to the function of the information therein: information as a good and information as a signal. Macroeconomic justifications are also discussed. Exceptions and limitations to intellectual property rights are seen as solutions to curb the ill effect of a monopoly right granted by intellectual property. Finally we examine the use of the economic justification as a yardstick to assess proposals for new forms of intellectual property rights.

Intellectual property rights are intangible rights in information. The World Trade Organization's

Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) requires member states to have laws protecting the seven most common types of intellectual property rights, namely, copyright, trademarks, geographical indications, industrial designs, patents, layout-designs (topologies) of integrated circuits, and undisclosed (or confidential) information. In some countries such as in the European Union, database rights are also available.

Since intellectual property rights essentially create a limited form of monopoly, it is necessary to examine the economic justification for them.

Information as a Public Good

The economic justification for intellectual property rights begins with the characteristics of information. Information is said to possess the twin characteristics of public goods (Arrow 1962): non-rivalry in consumption and non-excludability (Sidgwick 1887; Pigou 1923; Samuelson 1954).

Non-rivalry in consumption refers to the characteristic of goods being not exhaustible by consumption. In other words, a public good does not decrease in both quantity and quality as more users consume the good or when the amount of consumption of that good increases. Typically, the cost structure of a public good is said to have a fixed cost, of which is usually considered high, and zero marginal cost, that is, the cost of additional usage is theoretically zero. This quality of non-rivalry in consumption is a good characteristic because it makes the good inexhaustible for consumption and is thus an infinite good.

The second characteristic of public goods is non-excludability. Non-excludability means that once the public goods have been released or made available to the public, it is extremely difficult, if not impossible, to exclude non-payers from consuming or enjoying the benefits, or to compel users to voluntarily pay for the public goods, because it can be enjoyed without payment. It is this characteristic which is the root cause of the public goods problem. Pigou (1923) explains non-excludability

as “instances in which marginal private net product [falling] short of marginal social net product, because incidental services are performed to third parties from whom it is technically difficult to exact payment.” Non-excludability is a bad characteristic because it often leads to the problem of free riding, i.e., consumers enjoying its benefits without paying for its costs.

Using the above two characteristics as the defining criteria for public goods, it can be shown that information can be classified as public goods. By information, we do not confine it to bits of information but also include knowledge and electronic data. At the same time, we exclude from the definition of information its medium of carrier or the physical manifestation of information. For example, a novel consists of the physical medium, namely, paper and ink, and the information. Paper and ink have a cost component and by definition are rivalrous in consumption. On the other hand, the informational content of a novel, the storyline, and the plot and character development are intangible in nature and are non-rivalrous in consumption. One person’s knowledge and appreciation of the novel’s content do not hinder another person’s knowledge and appreciation of the same.

Likewise a motor vehicle such as a car contains both the physical embodiment of information and its informational content. The metal, rubber, and other materials used in manufacturing a car are merely the physical embodiment of the informational content of the design and technology behind the car. Although in this case not many people would be able to fully comprehend the technical details in the design of a car, the same analysis holds in that this information does not decrease in both quantity and quality in the hands of a person who could fully comprehend it. Moreover it is non-excludable because short of never selling to the public the said car, once the car is available in the market, a technically competent competitor might be able to reverse-engineer the design of the car without paying the original inventor, unless the force of law compels him to do so.

In the case of digital music, the music files have no physical embodiment short of some electromagnetic configurations. Digital files can be

easily copied at negligible or no reproduction cost. And once the music is made available on a digital network such as the Internet, it is extremely difficult or impossible to limit access to it to payers only. Thus digital music also exhibits the characteristics of public goods.

Market Failures

Information as public goods suffers from market failure due to free riding. Free riding or freeloading is defined as enjoying a benefit without paying for its cost. Two types of market failures may occur as a result. First is sub-optimal level of investment in the creation of information because of not taking into account the social benefits of information. The second is no or little investment in the creation of new information due to the threat of free riding.

To illustrate the first, we can use the example of computer software. A lone programmer who has no means of preventing free riding if he releases a computer program he has written will typically not be able to create a software package as comprehensive and well-designed as a piece of commercial software done by a software firm, because to develop commercial software would require a more extensive manpower and financial investment than could be shouldered by a typical lone programmer. So in the absence of the ability to obtain property rights in his software to the exclusion of non-paying customers, the lone programmer would probably develop his software in a sufficiently simple way to satisfy his own personal requirement. In other words, investment of effort by the programmer would be to the extent of marginal cost equals to marginal private benefit. This can be considered as a sub-optimal level of investment as optimality would require the programmer to take into account the social benefit of his creation and thus expand the programming effort to the extent of marginal cost equals to marginal social benefit.

The second form of market failure exists in the form of sub-optimal level of new information being created. This market failure occurs when potential copiers and users wait out for someone

else to create the new information. Since everyone is waiting for someone else to take the first step in creating new information, no one or few will end up doing the work of creation. This is akin to the case of prisoner's dilemma (Gordon 1992) where a cooperative strategy would be the optimal outcome, but since everyone's dominant strategy is to wait to free ride, the resulting Nash equilibrium is that all players will choose the free-riding strategy and no new information is created.

Solutions to Public Goods Problem

The traditional solution to the public goods problem is state provision from taxation as a means of funding. Both state provision and taxation make sense in relation to public goods, the benefits of which are enjoyed by the population at large, e.g., national defense, public health, and machineries of the government. However, it is not given that public goods must be provided by the state alone nor to be funded solely through a system of taxation. Non-state entities can be paid to provide public goods, or funding can be made through some form of nonvoluntary payment imposed on users. For example, Coase (1974) reminded us that in the 19th century, some lighthouses in England were operated by private parties. Passing ships were required to pay compulsory light dues.

In the same way, instead of relying on the state to create or fund the creation of new information, intellectual property rights as a form of property rights which allow creators to collect a user license fee become the preferred way to solve the problem of information as public goods. Indeed, Pigou (1923) in his *magnum opus*, *The Economics of Welfare*, discussed the public goods problem of scientific knowledge and the role of patent law to mitigate the problem of public goods.

There are several advantages to using intellectual property rights as a solution. First is that there will be less cross-subsiding by nonusers for the benefits of users. Using a system of intellectual property rights would mean that only users of those intellectual properties will have to pay license fees or purchase original copies.

Another advantage is that different creators will find their own profit-maximizing strategy based on what they perceive as the market demand, their own abilities, and preferences. So there is no difficulty of requiring central planning and the need for the state to decide on what information to create for its people. It also allows minority creators outside the mainstream society to cater to the demands of the members of the society at the fringe. In other words, intellectual property rights, to a certain extent, promote liberty and freedom of speech and expression.

On the other hand, intellectual property rights do not preclude state funding in the creation of new information. Indeed it is essential for the state to fund research both for its own use, such as for education and policy, and for public consumption, such as in the case of scientific research.

Intellectual property rights as a form of intangible property rights have another advantage in that it is transferable. This means that the potential value in an intellectual property can be realized by the owner selling or exclusively licensing it to another. Creators may choose this route in order to focus on the act of creation and leave the business of commodification to others.

Two Families of Intellectual Property Rights

Traditional legal treatment of intellectual property law tends to lump all different types of intellectual property rights into a single conceptual basket called “intellectual property.” Rather, all these different types of intellectual property rights are treated as discrete forms with no apparent similarities among them. Although this approach has its advantage for its simplicity, it nevertheless obscures the true nature and functions of different types of intellectual property rights. In general, intellectual property rights can be grouped into two families according to the function of the information therein. The first is information as a good, and the second is information as a signal.

Examples of intellectual property rights in the form of information as a good are copyright, patents,

industrial designs, layout-designs of integrated circuits, and confidential information. The commonality among these types of intellectual property rights is that the information so protected is an object that can be consumed to increase one’s utility, such as a story, a picture, a piece of music or art, an invention, a design, or some proprietary knowledge.

Information as a signal on the other hand covers intellectual property rights such as trademarks, passing off, and geographical indications. The information therein is not an object of consumption per se and cannot be enjoyed directly. Instead it helps us to use the market better by providing indications as to the sources and quality of goods and services associated with those intellectual property rights. Typically, information as a signal comes in the form of a brand, a logo, or an image of a personality, which may convey the origin or endorsement of a product.

These two families of information suffer from different forms of free riding and thus merit separate investigation.

Free Riding in Information as a Good

Free riding in information as a good happens when a free-rider, without the right holder’s consent, uses or reproduces an intellectual property. Thus, intellectual property rights are used to correct this market failure by granting to its creators and owners an exclusive or monopoly right to exploit the information so created (Plant 1934). Typically this occurs in two ways.

The first way of exploitation is in the form of products embodying the intellectual property, such as books, digital content, and technological inventions. As intellectual property rights give an exclusive right to the owner, the owner can then sell his products at a profit-maximizing monopoly price which is higher than their marginal costs of production. It is this higher than market competitive price which hopefully will bring excess profit sufficient to cover the high cost of the creation of the proprietary information.

The second way of exploitation is by way of licensing so that profit is made by allowing others to exploit the protected intellectual property

instead of selling products by the intellectual property owners themselves.

Free Riding in Information as a Signal

Intellectual property rights are also used to correct the market failure from free riding of information as a signal. Examples of this type of intellectual proprietary rights are trademarks, the tort of passing off in common law jurisdictions (or unfair competition in others), and geographical indications (Ritzert 2009).

Traditionally trademarks and the tort of passing off are said to protect the trade origin of or goodwill associated with the products. What this means is that the protected trademark serves as a link between the product and its originator, i.e., the trademark owner.

Businesses apply trademarks to products and services in order to distinguish the products and services originating from them from those of competitors. In a competitive market, or even in a duopoly, businesses increase profit by selling more. One of the ways of doing so is to attract and retain customers by offering a superior product compare to the competitors'. Even for market segment with low-quality products, businesses retain customers by offering products of a consistent quality. Alternatively, businesses attract customers by way of advertising in order to get potential customers familiar with their products through an advertising campaign.

Both strategies of superior or consistent product quality and advertising are not costless. A free-riding competitor may steal a trademark owner's market share by selling counterfeit products bearing the right holder's trademark without incurring the above costs. In fact, counterfeiters can do better in the short run by selling cheap, low-quality counterfeit products if quality is costly.

Trademark rights therefore grant the trademark owners an exclusive right to use his trademark for the classes of products or services it is registered to. With this exclusive right, the trademark owner can legally prevent counterfeiters from using his trademark by way of criminal and civil enforcements.

Another economic justification for trademark protection is to prevent consumer confusion. Supposing that a consumer values a genuine product highly and pays a high price for this product, but the product sold is a low-quality counterfeit, it would mean that the consumer has suffered a disutility. Trademark protection thus can be justified to prevent this form of economic transactions.

Therefore the economic functions of trademarks as intellectual property rights are to encourage optimal investment in advertising, optimal investment in product quality on the part of producers, and optimal consumption on the part of consumers (Landes and Posner 1987; Economides 1988). Without trademark rights, counterfeiters will reduce a trademark owner's incentive to promote his products bearing his trademark because part of the potential sale will be diverted to the counterfeiters as consumers are misled into buying counterfeit goods.

In addition, low-quality counterfeit goods in the market will jeopardize the trademark owner's effort in ensuring high and consistent quality of products bearing his trademark. Like in the market for lemons (Akerlof 1970), low-quality counterfeit goods will crowd out high-quality goods by the trademark owner as consumers could not differentiate the quality of counterfeit goods from genuine goods by way of observation. Under the condition of uncertainty in the quality, consumers will treat the market of the trademark goods as having an average quality and will have a lower willingness to pay, leading to a low equilibrium price. In the end, without trademark protection, low-quality counterfeit goods will crowd out high-quality genuine goods, making high-quality goods unavailable in the market.

Macroeconomic Justifications

Intellectual property rights are also justified from a macroeconomic perspective, especially in relation to intellectual property rights being incorporated into international trade treaties.

One argument, in the context of technology transfer, is that foreign businesses are reluctant to introduce or bring in new technology or

know-how to a country which does not have sufficiently broad and strong intellectual property rights protection. Intellectual property protection is seen as a pull factor for foreign investment, be it technological transfer or sending the design of a product to be manufactured in the receiving country (Mansfield 1995; Fink and Maskus 2005). Hence, intellectual property is argued as an exogenous growth factor.

Another macroeconomic justification builds on the relationship between intellectual property rights and economic growth (Gould and Gruben 1996). For example, businesses may earn higher profits with trademark protection. In addition, businesses are seen as having strong incentive to develop new knowledge and technology with copyright and patent protections (Sweet and Maggio 2015). Nevertheless, evidence of positive correlation between intellectual property rights and economic growth is ambiguous in many sectors (Aziz 2003; Yueh 2007).

Economic Justifications Versus Motivation

Economic justifications for intellectual property right are theoretical reasons or, at most, plausible empirical support for a system of intellectual property law. Empirical studies have uncovered other reasons and personal motivations for creators to create new works and inventions. For example, it was found that computer programmers voluntarily contribute to open-source projects because of identification with a group, pragmatic motivation to improve one's software or career, social and political motivation to support a movement, and enjoyment derived from the process itself (Hertel et al. 2003).

Exceptions and Limitations to Intellectual Property Rights

Although propertization through intellectual property rights is the preferred method of solving the market failure in information problem, it on the other hand creates a different form of market

failure, i.e., monopoly from exclusive right. In order to contain and limit the ill effects of monopoly power from intellectual property rights, various legal strategies are used, some of which are discussed below.

The first strategy is to control the breath of the intellectual property rights, i.e., the extent it could prevent other similar creations from being lawfully used. Limiting the breath of protection will facilitate competitors' ability to produce close but differentiated substitutes in the marketplace and thereby foster a market characterized by monopolistic competition (Chamberlin 1933) or imperfect competition (Robinson 1933). For example, copyright law protects expressions and not ideas. So general ideas of fictional stories are not protected, and authors may write fictional stories of the same genres as long as the detailed plots and character names are not colorably similar. Similarly, different publishers may produce their own language dictionaries containing substantially similar word list but with their own definitions.

The second strategy is to limit the term or duration of protection. Almost all forms of intellectual property rights, except for trademarks and confidential information, have limited term of protection. Limiting the term of protection has the beneficial effect of curbing the static inefficiency from monopoly power and allows users and other producers to have unlimited use of the intellectual property when it lapses into the public domain upon the expiry of its term (Walterscheid 2000). Although trademarks may be renewed indefinitely, registration may be revoked for non-use, so as to prevent useful trademarks from being hoarded by first-moving trademark owners. Undisclosed confidential information gets protection so long as the information is kept secret and not disclosed to the public. It loses protection immediately when an independent party rediscovers the confidential information through independent effort or through reverse-engineering of a publicly available product.

The third strategy is to create situational exceptions to protection. An economic justification for such an exception is that the transaction cost of licensing for the use of the intellectual property would outweigh the benefit of using the

intellectual property. Potential users will be deterred from using an intellectual property if the only legal route is through licensing. Example of such an exception can be found in copyright's fair use doctrine (Gordon 1982; Gordon 2002). In jurisdictions with statutory exceptions in copyright law, situational exceptions for disadvantaged groups such as the accessibility exception for the disabled can be justified on the ground that not all publishers are willing or interested in selling products which caters to a small market, and thus allowing the disabled to use technological solutions, such as text-to-speech software, to solve their accessibility problem is socially desirable.

The fourth strategy is to allow compulsory licensing in exceptional circumstances. Compulsory licensing can be seen as a form of liability rule protection of intellectual property rights (Reichman 2009). When compulsory licensing is evoked, the intellectual property owner has no power to set the price but has to instead accept the price set by an authoritative body, which in some circumstances is below a monopoly price. It also dispenses with the potentially protracted negotiation process in a national emergency. Compulsory licensing provision can be found in patent law which allows the state to manufacture essential medicines when patent holders refuse to grant a license at a reasonable price. Despite its unpopularity, compulsory licensing is rarely used and most often acts as a threat to force patent holders to lower their price of medicine.

Economic Justifications as a Yardstick

Assuming that the proper justification for intellectual property rights is economic in nature, i.e., to solve a market failure problem, then this economic justification can also be used as a yardstick to assess whether new forms of information should be conferred an intellectual property rights protection.

Certain countries and communities are pressing the World Intellectual Property Organization (WIPO) to recognize and protect traditional

knowledge. Unlike patent which protects only new inventions, traditional knowledge protection, as currently defined by WIPO, protects "knowledge, know-how, skills and practices that are developed, sustained and passed on from generation to generation within a community" (Matsui 2015). It appears that since traditional knowledge has been around the relevant communities for a long time, it would not be capable of patent protection because it would not satisfy the novelty requirement of patent law. The fact that traditional knowledge already exists means that there is no lack of an incentive to create problem. Instead the usual justification for protecting traditional knowledge is to reward or compensate the community for developing and preserving the traditional knowledge. However, this justification is actually of a distributional concern and not of an efficiency concern. In other words, traditional knowledge protection is merely a mechanism to transfer some wealth from an often wealthy first-world exploiter to a usually, but not necessarily, poorer and less-developed community.

Another application of the economic justification can be seen in assessing the arguments for retroactive extension of intellectual property term. Opponents of retroactive term extension argue that since retroactive term extension does not incentivize the creation of new intellectual property, it is not a good legal policy as it contradicts the basis for intellectual property rights which is to solve the market failure problem of information (Akerlof et al. 2002).

Cross-References

- ▶ [Copyright](#)
- ▶ [Creative Commons and Culture](#)
- ▶ [Digital Piracy](#)
- ▶ [European Patent System](#)
- ▶ [Geographical Indications](#)
- ▶ [Lighthouses](#)
- ▶ [Patent Litigation](#)
- ▶ [Patent Opposition](#)
- ▶ [Public Goods](#)
- ▶ [Trade Secrets Law](#)
- ▶ [Trademark Dilution](#)

- [Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond](#)
- [TRIPS Agreement](#)

References

- Akerlof GA (1970) The market for “lemons”: quality uncertainty and the market mechanism. *Quarterly J Econ* 84:488–500
- Akerlof GA, Arrow KJ, Bresnahan TF, Buchanan JM, Coase RH, Cohen LR, Friedman M, Green JR, Hahn RW, Hazlett TW, Hemphill CS, Litan RE, Noll RG, Schmalensee R, Shavell S, Varian HR, Zechkauser RJ (2002) Brief as amici curiae in support of petitioners in *Eldred v Ashcroft* 537 US 186 (2003) (No 01–618)
- Arrow K (1962) Economic welfare and allocation of resources for inventions. In: Universities-National Bureau Committee for Economic Research C on EG of the SSRC (ed) *The rate and direction of inventive activity: Economic and social factors*. Princeton, New Jersey University Press, pp 609–626
- Aziz S (2003) Linking intellectual property rights in developing countries with research and development, technology transfer, and foreign direct investment policy: a case study of Egypt’s pharmaceutical industry. *ILSA J Int Comp Law* 10:1–34
- Chamberlin EH (1933) *The theory of monopolistic competition: a re-orientation of the theory of value*. Harvard University Press, Cambridge
- Coase RH (1974) The lighthouse in economics. *J Law Econ* 17:357–376
- Economides NS (1988) The economics of trademarks. *Trademark Report* 78:523–539
- Fink C, Maskus KE (eds) (2005) *Intellectual property and development: lessons from recent economic research*. World Bank and Oxford University Press, Washington, DC
- Gordon WJ (1992) Asymmetric market failure and prisoner’s dilemma in intellectual property. *Univ Dayt Law Rev* 17:853–870
- Gordon WJ (1982) Fair use as market failure: a structural and economic analysis of the Betamax case and its predecessors. *Columbia Law Rev* 82:1600–1657.
- Gordon WJ (2002) Excuse and justification in the law of fair use: transaction costs have always been part of the story. *J Copyr Soc USA* 50:149–198
- Gould DM, Gruben WC (1996) The role of intellectual property rights in economic growth. *J Dev Econ* 48:323–350
- Hertel G, Niedner S, Herrmann S (2003) Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Res Policy* 32:1159–1177
- Landes WM, Posner RA (1987) Trademark law: an economic perspective. *J Law Econ* 30:265–309
- Mansfield E (1995) *Intellectual property protection, direct investment, and technology transfer: Germany, Japan, and the United States*. Washington, DC
- Matsui K (2015) Problems of defining and validating traditional knowledge: a historical approach. *Int Indig Policy J* 6:art 2
- Pigou AC (1923) *The economics of welfare*, 3rd edn. Macmillan, London
- Plant A (1934) The economic aspects of copyright in books. *Economica* 1:167–195
- Reichman JH (2009) Compulsory licensing of patented pharmaceutical inventions: evaluating the options. *J Law, Medicine & Ethics* 37:247–263
- Ritzert M (2009) Champagne is from champagne: an economic justification for extending trademark-level protection to wine-related geographical indicators. *AIPLA Q J* 37:191–225
- Robinson J (1933) *The economics of imperfect competition*. Macmillan, London
- Samuelson PA (1954) The pure theory of public expenditure. *Rev Econ Statistics* 36:387–389
- Sidgwick H (1887) *The principles of political economy*. Macmillan, London
- Sweet CM, Maggio DSE (2015) Do stronger intellectual property rights increase innovation? *World Dev* 66:665–677
- Walterscheid EC (2000) Defining the patent and copyright term: term limits and the intellectual property clause. *J Intellect Prop Law* 7:315–394
- Yueh LY (2007) Global intellectual property rights and economic growth. *Northwest J Technol Intellect Prop* 5:436–448

Internal Motivations

- [Intrinsic and Extrinsic Motivation](#)

International Litigation and Arbitration

S. I. Strong
University of Missouri School of Law, Columbia,
MO, USA

Abstract

The world of international dispute resolution has changed significantly in the last ten to fifteen years. International actors now must consider a wide array of options regarding

how, when and where their legal disputes will be resolved. This entry discusses the various means of resolving international legal disputes and outlines how law and economics analysis has helped rationalize an increasingly chaotic field of law. In so doing, the discussion considers four core areas of concern: the effectiveness of international dispute resolution; competition between and within different dispute resolution mechanisms; choice of substantive and procedural law and conflict of laws; and third-party funding.

Definition

The field of international dispute resolution is densely populated and includes a variety of types of litigation and arbitration. Not only do these mechanisms differ from each other; they also vary significantly from domestic forms of litigation and arbitration. Scholars must take these distinctions into account if their research is to be credible.

International litigation can refer to the judicial resolution of claims in either national court (sometimes called *transnational litigation*) or international court. An international court may operate regionally (as does the European Court of Human Rights) or globally (as does the International Court of Justice). Disputes heard in international courts are governed by public international law. Although claims arising under public international law may also be heard in national courts, international litigation in national courts typically involves matters of private international law. As a result, international litigation in national courts is often resolved pursuant to domestic legal principles, with the only “international” element arising as a result of the nationalities of the parties.

International commercial arbitration refers to final and binding adjudication of a dispute by a private, neutral decision-maker (i.e., a single arbitrator or a three-person arbitral tribunal). The definition of “international” varies according to national legislation but typically involves parties from different jurisdictions or parties from the

same jurisdiction arbitrating in a foreign country. The definition of “commercial” also varies according to national law, although a number of legal systems simply require some sort of economic effect.

International commercial arbitration is characterized as a “private” form of dispute resolution because it only arises upon the consent of the parties. However, international commercial arbitration also includes a number of public features. For example, the international commercial arbitration regime is built upon an intricate web of international treaties and national legislation designed to support the enforcement of both arbitration agreements and arbitration awards across national borders. These legal instruments have been construed by courts from around the world on numerous occasions, and the data has been collected and published by public and private entities for over 50 years (Strong 2009). The United Nations Commission on International Trade Law (UNCITRAL) is one of the more important public providers, having created a freely accessible online source (CLOUT, or Case Law on UNCITRAL Texts) with a wide variety of national court decisions concerning the various model laws and conventions promulgated by UNCITRAL.

As a matter of practice, international commercial arbitration bears relatively little resemblance to domestic forms of arbitration, including consumer, employment, labor, or final offer arbitration. These latter mechanisms are often extremely informal and typically feature a decreased emphasis on legal authorities and argument. International commercial arbitration, on the other hand, is a very sophisticated and highly legalistic procedure that resembles complex commercial litigation much more than it does domestic arbitration. Awards rendered in international commercial arbitration are usually fully reasoned and are often published in denatured (anonymized) form, thus allowing the international legal community to evaluate and predict arbitral behavior despite the confidentiality of the proceedings themselves (Strong 2009).

Investment arbitration (also referred to as *investor-state arbitration* or *treaty-based arbitration*)

refers to arbitral proceedings that arise pursuant to a bilateral investment treaty (BIT), a multilateral investment treaty (MIT), a free trade agreement (FTA), or an investment protection agreement (IPA). Although *interstate arbitration* (meaning arbitration between two nation-states) may also be initiated under many of these instruments, such proceedings seldom occur. Instead, it is more likely that an individual investor will file proceedings seeking redress from injuries allegedly caused by the actions of the host state.

Investment arbitration can be distinguished from *contract-based arbitration* (which includes both international commercial arbitration and domestic forms of arbitration) in several ways. For example, consent to investment arbitration does not rely on contractual privity between the investor and the respondent state but instead arises as a result of a standing offer of arbitration from the respondent state (as reflected in the relevant treaty or other international agreement) to the individual investor. Investment arbitration also differs from international commercial arbitration in that the former primarily addresses questions of public international law. As a result, investment arbitration has been analogized to international courts in some key regards (Born 2012).

The procedural sophistication and formality of investment arbitration is similar to that of international commercial arbitration. In fact, investment arbitration is sometimes governed by procedural rules originally developed for use in international commercial arbitration. However, the quasi-public nature of investment arbitration has permitted or required the introduction of a number of procedural features not seen in other types of arbitration, including the possibility of third-party participation through the filing of *amicus*-type submissions. The need for transparency in investment arbitration has also led to the routine publication of investment awards, often in their original rather than denatured form. Investment awards are fully reasoned and may include dissenting opinions.

Although litigation and arbitration are the best-known methods of international dispute resolution, *international mediation* has received an increasing amount of attention in recent years in

both the international commercial and investment contexts. The surge of interest in international mediation is somewhat surprising, given the long-standing availability of *international conciliation*. However, international conciliation has never been as popular as international arbitration as a means of resolving cross-border conflicts.

Although considerable debate exists as to whether conciliation and mediation are identical in all regards, scholars agree that both mechanisms are nonbinding (unless and until the parties execute a binding settlement agreement) and cooperative rather than adjudicative. Like arbitration, mediation and conciliation are private dispute resolution mechanisms that arise through the consent of the parties. Although a full analysis of the law and economics literature concerning international mediation and conciliation is beyond the scope of the current discussion, a number of commentators have suggested that mediation and conciliation provide a more effective and cost-efficient means of resolving international disputes (Spain 2010; Welsh and Schneider 2013). Further research in this field would appear to be beneficial and could possibly build on the large and growing number of empirical and economic analyses concerning the efficacy of mediation and settlement in the domestic context. However, future work would need to specifically address the effectiveness of these mechanisms in the international realm.

Another area of inquiry that is not addressed in this entry involves interstate trade disputes, such as those heard by the World Trade Organization (WTO) or arising under certain FTAs. Although law and economics scholars have considered these types of concerns on occasion (Guzman and Simmons 2002; Symposium 2012), these matters are traditionally not considered under the rubric of international litigation or international arbitration *per se*. Instead, these disputes are generally analyzed as a matter of trade law.

Categories of Analysis

Domestic forms of arbitration and litigation have long been subject to analyses based on law and

economics (Benson 2000; Sanchirico 2012). Although some of these studies may be readily transferable to the international setting, caution must be exercised, given the numerous distinctions between international and domestic forms of dispute resolution.

At one time, scholars hesitated to use law and economics to consider international legal concerns (Dunoff and Trachtman 1999). However, the last 15 years has seen a significant increase in economic analyses relating to international law. Some of these studies have focused on institutional concerns, while others have emphasized features relating to the allocation of jurisdictional authority in matters involving regulatory overlap (Guzman 2008a; Trachtman 2008; Danielsen 2011). A significant amount of work has also been done in matters concerning international litigation and arbitration, facilitated, in large part, by the ever-increasing amount of empirical data concerning international dispute resolution (Drahozal 2006; Franck 2007a; Van Harten 2012).

At this point, the law and economics literature on international litigation and arbitration appears to fall into four different categories: effectiveness of international dispute resolution, competition between and within different dispute resolution mechanisms, choice of substantive and procedural law and conflict of laws, and third-party funding for international litigation and arbitration. Each of these matters is addressed separately below.

Effectiveness of International Dispute Resolution

Domestic forms of dispute resolution are typically predicated on assumptions about the coercive power of the state. However, some forms of international dispute resolution involve states as defendants or respondents, thus raising questions about whether, why, and to what extent such mechanisms are effective, given the absence of traditional means of coercion. These issues have been analyzed from several perspectives, including that of rational choice theory (Guzman 2008b), game theory (Ginsburg and McAdams 2004), and behavioral economics (Van Aaken 2014).

The premises of the individual studies can vary significantly. In some cases, researchers evaluate the effectiveness of different international tribunals so as to help improve the design of future international dispute systems and thereby maximize the possibility that the legal system in question will achieve its stated goals. Other scholars study effectiveness in order to identify the limitations of the relevant court or tribunal. In all cases, the ultimate aim is to rationalize a field that has traditionally existed outside the realm of empirical and scientific analysis. When considering these issues, commentators have discussed bodies as diverse as the International Court of Justice, the European Court of Human Rights, the International Tribunal for the Law of the Sea, and various sorts of arbitral tribunals, including those operating under the Convention on the Settlement of Investment Disputes Between States and Nationals of Other States (ICSID Convention).

Competition Between and Within Different Dispute Resolution Mechanisms

International dispute resolution has become increasingly fragmented in the last few decades. Although international courts and tribunals may have exclusive jurisdiction over some types of legal injuries, parties can often bring the same or similar claim in another venue. As a result, questions regarding jurisdiction can be framed in terms of both commons and anticommons.

The existence of jurisdictional choice has resulted in a great deal of competition between and within different dispute resolution mechanisms, which has led to scholars in law and economics to consider these sorts of issues from a variety of perspectives (Ramsmeyer 2000; Trachtman 2008). Perhaps the most extensive work has been done on the relative benefits of international commercial arbitration over international litigation in national courts. One natural area of inquiry involves the question of transaction costs, which in this field can include costs associated with the dispute resolution process itself (often referred to as “litigation costs”) as well as costs associated with negotiation of the underlying dispute resolution provision.

Most scholarship in this field has focused on litigation costs, which makes sense in light of the traditional understanding of international commercial arbitration as a less expensive option than international litigation in national courts. However, the cost-effectiveness of arbitration has recently been called into question, at least with respect to direct costs such as those relating to attorneys' fees and arbitrators' fees. As a result, some observers have suggested that arbitration has lost its competitive advantage over litigation.

Although the increase in direct costs is disturbing to many in the international community, international commercial arbitration may nevertheless retain its superiority over litigation because arbitration allows parties to reap certain indirect benefits. For example, international commercial arbitration is often said to reduce transaction costs and provide a more efficient means of resolving international commercial disputes because it offers more predictability with respect to the forum, procedure, substantive law, and enforceability of the resulting decision (Ramsmeyer 2000; Strong 2014). Although widespread adoption of the Hague Convention on Choice of Court Agreements could apply some pressure with respect to these features, that instrument has not yet come into force. As a result, international litigation does not seem to be closing the gap with arbitration in terms of procedural or substantive predictability.

To the contrary, studies appear to suggest the continuing existence of a "home court advantage" in national courts (Bhattacharya et al. 2007). Furthermore, the law and practice relating to the recognition and enforcement of foreign judgments around the world remains highly unpredictable (Rotem 2010). As a result, the most rational course of action for international commercial actors is to avoid international litigation in either party's national court.

Although this data suggests that international commercial arbitration is preferable to international litigation, law and economics scholars have nevertheless considered whether particular arbitral practices can be improved upon. One area of interest involves the pre-hearing exchange of

documents and information (referred to as either disclosure or discovery). Recent years have seen an increase in the amount of information that parties have sought on a pre-hearing basis. However, economic analysis has suggested a rational actor should prefer the more limited form of document production that characterized early forms of international commercial arbitration rather than the more expansive type of disclosure that is increasingly becoming the norm (Rojas Elgueta 2011).

Some commentators have analyzed arbitral procedures on a more comprehensive basis and have factored in whether and to what extent arbitration is capable of providing ancillary assistance (such as the ability to issue anti-suit injunctions or freeze assets) and increasing the enforceability of the resulting decision (Rutledge 2012). These studies suggest that, rather than adapting to make arbitral procedures more like judicial procedures, international commercial arbitration should retain its distinctiveness so as to retain its competitive advantage. Indeed, some scholars believe the comparative benefits of international commercial arbitration are so significant that arbitration should be made the default position in international commercial disputes (Cuniberti 2009). Some commentators distinguish between the efficiency of international commercial arbitration (Posner 1999; Kovacs 2012) and that of investment arbitration (Franck 2011).

One issue that the literature is addressing with increasing frequency involves the role that dispute resolution mechanisms play in the development of international regulatory law. In many cases, the absence of a single political actor with international regulatory authority has required parties who have suffered multijurisdictional legal injuries to seek relief from national courts. Although some judges have occasionally been willing to fill these sorts of international regulatory gaps, national courts are typically limited in their ability to exercise jurisdiction over foreign parties or apply domestic law extraterritorially (Buxbaum 2006). Furthermore, it can be difficult, if not impossible, for national courts to coordinate their regulatory standards and mechanisms, thereby resulting in market inefficiencies.

Scholarship in this field offers a number of different insights and conclusions. For example, some observers suggest that national courts should be allowed to exert a broad jurisdiction over international regulatory concerns so as to increase efficiency in this area (Mehra 2004). This approach is supported by studies suggesting that private litigants may be particularly efficient enforcers of public regulatory aims in cases involving public goods or commons constellations (Van Aaken 2010). However, other commentators have expressed concern about overgeneralization in this area of law and have suggested that the varying nature of different public goods requires case-by-case analysis (Symposium 2012).

One way to avoid some of the problems associated with regulation via litigation in national courts is through arbitration, since arbitration does not experience the same kinds of problems as litigation does with respect to jurisdiction, enforcement, and application of foreign law (Strong 2013). However, it is unclear whether and to what extent it is appropriate for private tribunals to adjudicate matters of public concern. Some commentators have suggested that questions of public law should be arbitrable in developing but not developed countries, since that approach maximizes economic development and international commercial activity (McConaughay 2001).

Other scholars evaluating regulatory concerns have focused not on the nature of the claim or the quality of the governing law, but instead on the parties themselves. For example, questions have arisen as to whether state actors may be considered proper participants in international commercial (as opposed to international investment) arbitration. Authors approaching the issue from a law and economics perspective have answered that question positively, based on analyses demonstrating that states operate as firm-like economic organizations (Guevera-Bernal 2004).

Although states sometimes appear as parties in international commercial arbitration, they may also be named as respondents in international investment arbitration. Investment arbitration is a risky and costly endeavor, and states and commentators are continually reassessing the

economic value of participating in a treaty regime that exposes them to such a high degree of financial liability (Franck 2007b; Trakman 2012; United Kingdom Department for Business, Innovation and Skills 2013). One of the issues that is constantly under scrutiny is whether and to what extent participation in the investment treaty regime actually increases foreign direct investment in the signatory state (Franck 2007b).

States are not the only parties that must evaluate the value of investment arbitration. Investors who believe that they have suffered a legal injury typically undertake sophisticated cost-benefit analyses to determine whether to pursue a claim in investment arbitration or in another possible forum (Bjorklund 2007). Empirical and economic studies help investors make these decisions by illuminating the likelihood of success in a particular venue (Franck 2007a) and the costs associated with seeking a certain type of legal redress (Franck 2011).

Competition not only exists between different types of international dispute resolution, it also exists within each category of procedures. For example, arbitral institutions from around the world are constantly working to optimize their attractiveness to prospective parties and distinguishing themselves from competitor organizations (Dezalay and Garth 1995; Drahozal 2000a; Ramsmeier 2000; Choi 2003; Rogers 2006). In so doing, arbitral institutions must identify those procedural trends that should be followed and those procedural innovations that should be resisted. Arbitral institutions are now able to differentiate themselves in a variety of ways, ranging from transparency of proceedings to the availability of expedited procedures. Economic analyses consider not only whether and to what extent individual innovations improve the efficiency of the arbitral process but also how external market forces drive competition between the different organizations.

Competition also exists between individual cities and countries that want to position themselves as potential seats for arbitration. Rent-seeking lawyers may undertake a variety of efforts to position their city or state favorably vis-à-vis any actual or perceived competitors (Drahozal 2000a,

2004; O'Connor and Rutledge 2014). Economic analyses in this area consider the effectiveness of these various means of attracting arbitration business to a particular region, including standardization versus innovation in the national arbitration law, improvement of judicial procedures relating to arbitration, liberalization of the underlying substantive law, and relaxation of rules regarding the unauthorized practice of law.

Competition also exists with respect to individual arbitrators seeking to be named to various public or private tribunals (Dezalay and Garth 1995; Drahozal 2000a; Rogers 2005, 2006; Franck 2009). Commentators have suggested that informational imperfections regarding the qualifications and availability of potential arbitrators and the performance of arbitrators create numerous inefficiencies in the process of selecting arbitrators. Furthermore, the market for international arbitrators appears to reflect a number of barriers to entry. Indeed, empirical research clearly demonstrates that the market for international arbitrators is extremely constrained in terms of age, nationality, race, gender, education, and professional qualifications.

Choice of Law and Conflict of Laws

Another area of inquiry that has benefitted from the application of economic analysis involves choice of law in international disputes. A number of studies focus on issues of substantive law, based on the premise that substantive inefficiencies have a detrimental effect on global economic welfare (Drahozal 2000a; Rühl 2006; Whytock 2009; Ller 2011; O'Connor and Rutledge 2014). When testing this hypothesis, some researchers focus on the relative benefits of different national laws and non-state law, such as the International Institute for the Unification of Private Law (UNIDROIT) Principles of International Commercial Contracts and the United Nations Commission on International Trade Law (UNCITRAL) Convention on the International Sale of Goods (CISG), while other commentators consider the transactional costs associated with creating individualized contracts relating to choice of law. Other analyses look at state-imposed conflict of laws rules to determine

whether and to what extent these default regimes can be considered either efficient or reflective of the parties' presumed intent. In each of these cases, the methodological approach and motivating principles are similar to those seen in studies of the domestic law market, although the analysis is much more challenging at the international level, since the relatively high degree of substantive and procedural autonomy associated with international dispute resolution increases the number of variables that must be considered.

One area of particular interest to law and economics scholars is the *lex mercatoria*, also known as the international law merchant. The *lex mercatoria* is said to embody certain customary international legal norms concerning trade practices and usages and therefore provides various insights into the nature of contract law and corporations (Symposium 2004; Oman 2005). Other studies have focused on whether and to what extent arbitrators actually rely on the *lex mercatoria* and have often concluded that this body of law is applied much less frequently in practice than in theory (Drahozal 2000a; Symposium 2004).

Although most studies relating to substantive choice of law are global in nature, some focus on the particular pressures that arise within the European Union as a result of regional integration of certain political and economic matters (Watt 2003; Michaels and Jansen 2006). Specialists in European law have also considered the extent to which US-style law and economics theory has really taken hold in European legal thinking regarding conflict of laws questions and have concluded, contrary to certain conventional wisdom, that there is indeed evidence of a cross-Atlantic jurisprudential dialogue (Nicola 2008).

Although the literature on law and economics is rife with discussions relating to choice of substantive law, procedural choice of law is also an issue of interest in the field. Thus, a number of studies consider the transaction costs associated with customizing arbitration provisions (Choi 2003) or litigation procedures (Drahozal and O'Connor 2014; Hoffman 2014; Strong 2014). Although parties often have a great deal of scope in which to exercise their procedural autonomy,

several of these studies suggest that parties seldom do so, preferring instead to rely on default provisions established as a matter of state law or arbitral custom.

Other inquiries focus more on systemic issues. Thus, some commentators consider the economic benefits associated with procedural harmonization across national boundaries (Visscher 2012), while other observers focus on questions relating to why procedures promoting settlement are present in some jurisdictions but absent in others (Cortés 2013). These types of analyses may be of particular relevance to legislators seeking to improve the design of their national dispute resolution regimes.

Legislators may also be interested in studies focusing on the efficacy of specific rules of procedure. Thus, economic analyses have been applied to questions involving the exercise of extraterritorial jurisdiction (Trachtman 2001, 2008), enforcement of foreign judgments (Rotem 2010; Rühl 2006), and enforcement of foreign arbitral awards (Drahozal 2000b). These studies are quite diverse in nature and consider everything from the role of externalities in the decision-making process and how international jurisdiction is both a commons and an anticommons to the role of informational asymmetries and general issues relating to regulatory competition. However, some critics contend that law and economics is not the proper lens through which to view problems of international procedure (Kessedjian 2005).

The literature also encompasses a number of macro-level analyses, such as those considering the differences, if any, between hard and soft forms of international law (Shaffer and Pollack 2010) and between the common law and civil law traditions (Mattei 1997; Parisi 2002; Arruñada and Andonova 2008). These studies can be useful in determining the cost-effectiveness of various types of reform, both as a substantive and procedural matter.

Third-Party Funding

As international dispute resolution has become more expensive, parties and legal systems have begun to explore alternate ways of funding the

various proceedings. One particularly intriguing mechanism involves third-party funding, which is sometimes referred to as “third-party litigation funding.” This device involves an unaffiliated entity choosing to pay for the legal costs of one of the parties in exchange for a certain percentage of the final award. Third-party funding has been used in both litigation and arbitration and is coming under increased economic scrutiny. Some studies are comparative in nature (Barker 2012), whereas others are either national (Hylton 2012) or international (Steinitz 2011) in scope. Few conclusions have yet been reached, since the mechanism is still in its early stages, but early analyses suggest that third-party funding will fundamentally change the economics of dispute resolution.

Cross-References

- [Alternative Dispute Resolution](#)
- [Conflict of Laws](#)
- [Globalization](#)
- [Lex Mercatoria](#)

References

General

- Arruñada B, Andonova V (2008) Common law and civil law as pro-market adaptations. *Wash Univ J Law Policy* 26:81–130
- Bjorklund AK (2007) Private rights and public international law: why competition among international economic law tribunals is not working. *Hastings Law J* 59:241–307
- Born G (2012) A new generation of international adjudication. *Duke Law J* 61:775–879
- Danielsen D (2011) Economic approaches to global regulation: expanding the international law and economics paradigm. *J Int Bus Law* 10:23–89
- Dezalay Y, Garth B (1995) Merchants of law as moral entrepreneurs: constructing international justice from the competition for transnational business disputes. *Law Soc Rev* 29:27–62
- Dunoff JL, Trachtman JP (1999) Economic analysis of international law. *Yale J Int Law* 24:1–59
- Ginsburg T, McAdams RH (2004) Adjudicating in anarchy: an expressive theory of international dispute resolution. *William Mary Law Rev* 45:1229–1331
- Guzman A (2008a) How international law works: a rational choice theory. Oxford University Press, Oxford
- Guzman A (2008b) International tribunals: a rational choice analysis. *Univ Pa Law Rev* 157:171–235

- Guzman A, Simmons BA (2002) To settle or empanel? An empirical analysis of litigation and settlement at the World Trade Organization. *J Legal Stud* 31:205–227
- Kessedjian C (2005) Dispute resolution in a complex international society. *Melb Univ Law Rev* 29:765–808
- Ller HE (2011) The transnational law market, regulatory competition, and transnational corporations. *Indiana J Glob Legal Stud* 18:707–749
- Mattei U (1997) Comparative law and economics. University of Michigan Press, Ann Arbor
- McConaughay PJ (2001) The scope of autonomy in international contracts and its relation to economic regulation and development. *Columbia J Transnatl Law* 39:595–656
- Michaels R, Jansen N (2006) Private law beyond the state? Europeanization, globalization, privatization. *Am J Comp Law* 54:843–890
- Oman N (2005) Corporations and autonomy theories of contract: a critique of the new *lex mercatoria*. *Denver Univ Law Rev* 83:101–145
- Rutledge PB (2012) Convergence and divergence in international dispute resolution. *J Dispute Resol* 2012:49–61
- Rühl G (2006) Methods and approaches in choice of law: an economic perspective. *Berkeley J Int Law* 24:801–841
- Sanchirico CW (2012) 8 Encyclopedia of law and economics: procedural law and economics. Edward Elgar, Cheltenham
- Shaffer GC, Pollack MA (2010) Hard vs. soft law: alternatives, complements, and antagonists in international governance. *Minn Law Rev* 94:706–798
- Spain A (2010) Integration matters: rethinking the architecture of international dispute resolution. *Univ Pa J Int L* 23:1–55
- Symposium (2004) The empirical and theoretical underpinnings of the law merchant. *Chic J Int Law* 5:1–190
- Symposium (2012) Global public goods and the plurality of legal orders. *Eur J Int Law* 23(3):643–791
- Trachtman J (2008) The economic structure of international law. Harvard University Press, Cambridge, MA
- Van Aaken A (2010) Trust, verify, or incentivize? Effectuating public international law regulating public goods through market mechanisms. *Am Soc Int Law Proc* 104:153–156
- Van Aaken A (2014) Behavioral international law and economics. *Harv J Int Law* 55
- Watt HM (2003) Choice of law in integrated and interconnected markets: a matter of political economy. *Columbia J Eur Law* 9:383–409
- Whytock C (2009) Myth of mess? International choice of law in action. *NY Univ Law Rev* 84:719–790
- Choi SJ (2003) The problem with arbitration agreements. *Vanderbilt J Transnatl Law* 36:1233–1240
- Cuniberti G (2009) Beyond contract – the case for default arbitration in international commercial disputes. *Fordham Int’Law J* 32:417–488
- Drahozal CR (2000a) Commercial norms, commercial codes, and international commercial arbitration. *Vanderbilt J Transnatl Law* 33:79–146
- Drahozal CR (2000b) Enforcing vacated international arbitration awards: an economic approach. *Am Rev Int Arb* 11:451–479
- Drahozal CR (2004) Regulatory competition and the location of international arbitration proceedings. *Int Rev Law Econ* 24:371–383
- Drahozal CR (2006) Arbitration by the numbers: the state of empirical research on international commercial arbitration. *Arb Int* 22:291–307
- Franck SD (2007a) Empirically evaluating claims about investment treaty arbitration. *N C Law Rev* 86:1–87
- Franck SD (2007b) Foreign direct investment, investment treaty arbitration and the rule of law. *McGeorge Glob Bus Develop Law J* 19:337–373
- Franck SD (2009) Development and outcomes of investment treaty arbitration. *Harv Int Law J* 50: 435–489
- Franck SD (2011) Rationalizing costs in investment treaty arbitration. *Wash Univ Law Rev* 88:769–852
- Guevera-Bernal I (2004) The validity of state contracts arbitration: a ‘law and economics’ perspective. *Revista de la Maestría en Derecho Económico* 2:7–20
- Kovacs RB (2012) Efficiency in international arbitration: an economic approach. *Am Rev Int Arb* 23: 155–174
- O’Connor EO, Rutledge P (2014) Arbitration, the law market, and the new law of lawyering. *Int Rev Law Econ* 38:87–106
- Posner E (1999) Arbitration and the harmonization of international commercial law: a defense of *Mitsubishi*. *Va J Int Law* 39:647–670
- Ramsmeyer JM (2000) International dispute resolution: Law and economics. In: Hamada K et al (eds) *Dreams and dilemmas: economic friction and dispute resolution in the Asia-Pacific*. Institute of Southeast Asian Studies, Singapore, pp 464–477
- Rogers CA (2005) The vocation of the international arbitrator. *Am Univ Int Law Rev* 20:957–1020
- Rogers CA (2006) Transparency in international commercial arbitration. *Univ Kans Law Rev* 54: 1301–1337
- Rojas Elgueta G (2011) Understanding discovery in international commercial arbitration through behavioral law and economics: a journey inside the minds of parties and arbitrators. *Harv Neg Law Rev* 16:165–191
- Strong SI (2009) Research and practice in international commercial arbitration: sources and strategies. Oxford University Press, Oxford
- Strong SI (2013) Mass procedures as a form of “regulatory arbitration” – *Abacat v. Argentine Republic* and the

International Arbitration

- Benson BL (2000) Arbitration. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics: the economics of crime and litigation*. Edward Elgar, Cheltenham, pp 159–193

international investment regime. *J Corp Law* 38:259–324

Trakman LE (2012) The ICSID under siege. *Cornell Int Law J* 45:603–665

United Kingdom Department for Business, Innovation & Skills, Analytical Framework for Assessing Costs and Benefits of Investment Protection Treaties (2013) Available at <https://www.gov.uk/government/publications/analytical-framework-for-assessment-costs-and-benefits-of-investment-protection-treaties>

Van Harten G (2012) Arbitrator behaviour in asymmetrical adjudication: an empirical study of investment treaty arbitration. *Osgoode Hall Law J* 50:211–268

Welsh NA, Schneider AK (2013) The thoughtful integration of mediation into bilateral investment treaty arbitration. *Harv Neg Law Rev* 18:71–144

International Litigation

Barker GR (2012) Third-party litigation funding in Australia and Europe. *J Law Econ Policy* 8:451–524

Bhattacharya U et al (2007) The home court advantage in international corporate litigation. *J Law Econ* 50:625–659

Buxbaum HL (2006) Transnational regulatory litigation. *Va J Int Law* 46:251–317

Cortés P (2013) A comparative review of offers to settle – would an emerging settlement culture pave the way for their adoption in continental Europe? *Civil Justice Quart* 23:42–67

Drahozal CR, O'Connor EO (2014) Unbundling procedure. *Fla L Rev* 66:389–430

Hoffman DA (2014) Whither bespoke procedure? *Ill Law Rev*

Hylton KN (2012) The economics of third-party financed litigation. *J Law Econ Policy* 8:701–741

Mehra SK (2004) More is less: a law-and-economics approach to the international scope of private antitrust enforcement. *Templ Law Rev* 77:47–70

Nicola FG (2008) Transatlanticisms: constitutional asymmetry and selective reception of U.S. law and economics in the formation of European private law. *Cardozo J Int Comp Law* 16:87–153

Parisi F (2002) Rent-seeking through litigation: adversarial and inquisitorial systems compared. *Int Rev Law Econ* 22:193–216

Rotem Y (2010) The problem of selective or sporadic recognition: a new economic rationale for the law of foreign country judgments. *Chic J Int Law* 10: 505–537

Steinitz M (2011) Whose claim is this anyway? Third-party litigation funding. *Minn Law Rev* 95:1268–1338

Strong SI (2014) Limits of procedural choice of law. *Brooklyn J Int Law* 39

Trachtman J (2001) Economic analysis of prescriptive jurisdiction. *Va J Int Law* 42:1–79

Visser L (2012) A law and economics view on harmonisation of procedural laws. In: Kramer XE, Rhee CH (eds) *Civil litigation in a globalising world*. Springer, New York, pp 65–92

Internet Governance

Philip C. Hanke

Department of Public Law, University of Bern, Bern, Switzerland

Abstract

The Internet is not regulated by a central authority, but instead by a plethora of institutions involving numerous stakeholders. Internet governance can be understood as a field that not only pertains to infrastructure regulation, but also to other legal questions such as copyright, hate speech, cyber-bullying, data protection, and others.

Definition

Internet governance, according to the UN-initiated World Summit on the Information Society, is “the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet” (“Report of the Working Group on Internet Governance” 2005).

Introduction

The Internet is a globally distributed network constituting a patchwork of interconnected autonomous networks. As such, there is no centralized governing body overseeing it. However, a governance structure is necessary for coordination and the assignment of certain property rights (e.g., domain names and Internet Protocol addresses). One early characterization of Internet governance identified three layers of regulation: the physical infrastructure layer contains the hardware, the logical layer contains the software, and the content layer contains the transmitted information. Each layer comes with its own questions, such as network access, copyright issues, or monetization (Benkler 2000).

The Internet, in a way, can be considered a public good (Spar 1999). Consumption is basically non-rivalrous. Although congestion can be an issue with increasing usage, its infrastructure is highly scalable and expandable almost without limits. The Internet also fulfills the second requirement of public goods, namely, that users cannot be excluded from accessing it. Certain services (e.g., access through specific Internet providers) are excludable, but the architecture as a whole is not.

Despite its characteristic as a public good, it has nevertheless successfully transferred into the private sector over the last 20 years. Although it first developed at universities and government agencies, the infrastructure, and arguable most of its content, is provided by private companies.

A Brief History of Internet Governance

The Internet as it is known today evolved from several precursor networks. One important network was ARPANET, developed and funded by the Advanced Research Projects Agency (ARPA) of the US Department of Defense (DoD) and established in 1969. In 1981, it was expanded by the National Science Foundation's Computer Science Network to connect its supercomputing facilities, research networks, and university campus networks. Since 1982, ARPANET has used the TCP/IP protocol suite, which is the cornerstone of modern Internet infrastructure. Throughout the 1980s, other countries were connected to ARPANET/NSFNET. In the 1990s, the network became the Internet: ARPANET was terminated in 1990, and the NSF started to first allow commercial usages of NSFNET in 1991, lifting the final restrictions in 1995. The NSF ended its sponsorship of the backbone services, and the private sector took over the provision of infrastructure. An originally state-owned network had thus become a mainly private network regulated by standards and commercial agreements (for a historical overview, see, e.g., Abbate 2000; Ryan 2013). Certain legacies from the Internet's early era, such as the role of the US government, survived until very recently.

Who Governs the Internet?

The operation of the Internet requires certain administrative tasks, such as assigning domain names (e.g., [google.com](https://www.google.com)), IP addresses (e.g., 216.58.214.110 is an IP address that the domain name [google.com](https://www.google.com) points to – however, a large website like Google operates many servers around the globe and thus uses many IP addresses for their different services), and other parameters. This assignment of names and numbers is overseen by the nonprofit Internet Corporation for Assigned Names and Numbers (ICANN) based in Los Angeles, California. ICANN, in turn, is managed by an International Board of Directors representing the constituencies of ICANN, its Supporting Organizations (the Generic Names Supporting Organization, the Country Code Names Supporting Organization, and the Address Supporting Organization), and its subgroups. Governments enter the structure through the Governmental Advisory Committee, which has an advisory role.

The Internet's domain name system (i.e., the list that assigns domain names to IP numbers) is administered by the Internet Assigned Numbers Authority (IANA), managed by ICANN. Until 2016, the US Department of Commerce (DoC) could overrule any resolution made by that body. This special role of the US government was a contentious issue for a long time, and handing over full control over IANA to ICANN was an important step in fully transition Internet governance toward a fully multi-lateral, international system.

An important forum to develop Internet governance is the Internet Governance Forum (IGF). It was established in 2006 by the secretary-general of the United Nations and holds annual meetings. Its main institutional bodies are the Multi-stakeholder Advisory Group (MAG) and the Secretariat based in Geneva, Switzerland. The MAG currently consists of 55 members representing governments, private industry, civil society, academia, and technical communities. Membership is not permanent but instead rotates regularly, and the UN secretary-general has the final say in the selection process initiated by the undersecretary-general for Economic and Social Affairs.

The standardization of the Internet's core protocols takes place under the auspices of the Internet Society (an NGO founded in 1992 and based in Reston, Virginia, and Geneva, Switzerland), more specifically mainly within the Internet Engineering Task Force, an open standards organization run by volunteers (whose work is usually funded by their employers). It also serves the host institution for the Internet Architecture Board (a committee of the IETF), the Internet Engineering Steering Group (a committee responsible for the technical management of IETF activities), and the Internet Research Task Force (focused on longer-term research issues related to the Internet).

Standards relating to the World Wide Web, that is, the part of the Internet that is accessed through the hypertext transfer protocol (HTTP) using a web browser, are set by World Wide Web Consortium (W3C). The Consortium has well over 400 members, including NGOs, private businesses, academic institutions, governmental entities, and some private individuals. Membership is subject to fulfilling certain requirements and receiving the approval of the existing members.

Thus, Internet governance is a patchwork, and proposals for reform to construct a coherent regime have been made since the mid-1990s. A global framework, for example, through a convention, was discussed in the academic debate but did not find momentum at the policy level (see, e.g., Mueller et al. 2007; Weber 2014).

Contentious Issues

The view that the Internet is a border-free network, free from territoriality, and thus controlled by governments is, however, contested. It has been argued that in many Internet-related policy fields – such as e-commerce, privacy, speech and pornography, intellectual property, and cybercrime – national governments do play a dominant role in regulation (Goldsmith and Wu 2006).

Occasionally, big decisions regarding the infrastructure arise. One large issue was a change in the length of Internet Protocol (IP) addresses, that is, the unique numbers required to identify machines on the Internet to be able to exchange information. The original system (IPv4) allowed for 4.3 billion

addresses, which is not enough for modern uses of the Internet. Although a new standard – IPv6, allowing for 340 undecillion addresses (36 zeros) – was selected already in the 1990s, the conversion process is still ongoing, which has to do with a nonobvious yet present political process taking place in the background (DeNardis 2009). The interconnectedness between infrastructure standard setting and broader, more political questions such as openness and transparency can thus make fairly straightforward decisions like increasing the number of IP addresses more difficult.

Another highly contested issue in terms of infrastructure is net neutrality, that is, the principle that Internet service providers (ISPs) and regulatory agency should treat all data on the Internet the same – the Internet equivalent of the common carrier principle in common law countries and the public carrier principle in civil law countries. In the European Union, Art. 3(3) of EU Regulation 2015/2120 requires ISPs to “treat all traffic equally, when providing Internet access services, without discrimination, restriction or interference, and irrespective of the sender and receiver, the content accessed or distributed, the applications or services used or provided, or the terminal equipment used.” Some EU member states also have national laws ensuring net neutrality. In the United States, net neutrality has been the object of many years of lobbying. In 2015, the Federal Communications Commission set forth rules protecting the equal treatment of Internet traffic, mainly by classifying broadband as a common carrier. However, it is possible that these rules change again under the Trump administration. The policy of mandated net neutrality is highly controversial: while it is defended on the grounds that it protects consumers by keeping access to the Internet equal to all and is important for innovation (Lee and Wu 2009), some economists criticize and see it as price regulation hindering competition (Hahn and Wallsten 2006; Becker et al. 2010; see in particular Maillé et al. 2012 for an overview of the debate).

Conclusion

A narrow definition of Internet governance would solely address questions of Internet infrastructure.

However, a broader view would encompass aspects of policing contents delivered through the Internet and thus include problems such as cyberbullying, copyright law, data protection rules, as well as safety-related matters.

Traditional approaches based on the principles of state sovereignty and clearly defined territoriality would fail to address the inherent problem of providing and maintaining a decentrally provided public good. Besides questions of infrastructure, activities such as hate speech or copyright infringement lead to strong interjurisdictional externalities. As a result, Internet governance is an archetypical example of highly decentralized, multilayered, transnational, public-private governance. Many important institutions for standard setting (such as the IETF) remain rather informal organizations and show a certain degree of ad hocism. Functionally, there is a lot of overlap between infrastructure design, public policy, and education in the activities of the involved institutions. With increasing digitalization and reliance on the Internet, contested questions regarding Internet governance gain relevance as the latter determines Internet freedom and as a result increasingly freedom of society in general (DeNardis 2014).

Cross-References

- Copyright
- Hate Groups and Hate Crime
- Privacy Regulation
- Telecommunications

References

- Abbate J (2000) *Inventing the Internet*, 58839th edn. The MIT Press, Cambridge, MA
- Becker GS, Carlton DW, Sider HS (2010) Net neutrality and consumer welfare. *J Competition Law Econ* 6(3):497–519. <https://doi.org/10.1093/joclec/nhq016>
- Benkler Y (2000) From consumers to users: shifting the deeper structures of regulation toward sustainable commons and user access. *Fed Commun Law J* 52(3): 563–580
- DeNardis L (2009) *Protocol politics. The globalization of internet governance*. MIT Press, Cambridge, MA
- DeNardis L (2014) *The global war for internet governance*. Yale University Press, New Haven
- Goldsmith J, Wu T (2006) *Who controls the internet? Illusions of a borderless world*. Oxford University Press, Oxford
- Hahn RW, Wallsten S (2006) The economics of net neutrality. *Econ Voice* 3(6). <https://doi.org/10.2202/1553-3832.1194>
- Lee RS, Wu T (2009) Subsidizing creativity through network design: zero-pricing and net neutrality. *J Econ Perspect* 23(3):61–76. <https://doi.org/10.1257/089533009789176780>
- Maillé P, Reichl P, Tuffin B (2012) Internet governance and economics of network neutrality. In: Hadjiantonis AM, Stiller B (eds) *Telecommunication economics, Lecture notes in computer science*, vol 7216. Springer, Berlin, pp 108–116. https://doi.org/10.1007/978-3-642-30382-1_15
- Mueller M, Mathiason J, Klein H (2007) The internet and global governance: principles and norms for a new regime. *Glob Gov Rev Multilateralism Int Organ* 13(2):237–254. <https://doi.org/10.5555/ggov.2007.13.2.237>
- Report of the Working Group on Internet Governance (2005) *Château de Bossey*. United Nations, Switzerland
- Ryan J (2013) *A history of the internet and the digital future*, Reprint edition. Reaktion Books, London
- Spar D (1999) The public face of cyberspace. In: Kaul I, Grunberg I, Stern M (eds) *Global public goods: international cooperation in the 21st century*. Oxford University Press, Oxford
- Weber RH (2014) *Realizing a new global cyberspace framework: normative foundations and guiding principles*. Springer, Berlin, Heidelberg

Intrinsic and Extrinsic Motivation

Rustam Romaniuc¹ and Cécile Bazart^{2,3}

¹Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier; International programme in institutions, economics and law (IEL), University of Turin, Montpellier, France

²CEE-M – Univ Montpellier – CNRS – INRA – SupAgro, University of Montpellier, Montpellier, France

³Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier 1, Montpellier, France

Abstract

The question of human motivation is central to understand the effects of legal rules on people's behavior. One of the main tenets of law and economics is that incentive systems need to be

designed in order to minimize the difference between private and social interests. Legal rules, taxes, subventions, and other external interventions are regarded as necessary to motivate the internalization of externalities. Empirical evidence however suggests that motivation to engage in pro-social behavior may preexist to external incentives. People often avoid cheating, polluting, or littering, and they act pro-socially without considering consequences of deviation to do so. The traditional idea of “laws as price incentives” has a difficult time explaining these phenomena. This is why in the last two decades, one of the most promising research agenda consisted in distinguishing between “extrinsic” and “intrinsic” motivations. The former are linked to actions driven by external incentives, in contrast to the latter driven by personal and internal forces.

Synonyms

External incentives; Internal motivations; Self-interest; Social preferences

Definition

Intrinsic motivation was originally defined as some inherent interest in a task, which cannot be derived from the agent’s economic environment. As such, intrinsically motivated behaviors aim at bringing about certain internal rewarding consequences independent of any extrinsic rewards. With the multiplication of experimental results displaying high levels of cooperation when monetary and social sanctions are inoperative, intrinsic motivation has come to be viewed as a natural response to the idea that socially desirable conduct can solely be motivated by economic incentives. The overarching idea is that because individuals have intrinsic motivations to behave pro-socially, extrinsic incentives such as promises of reward or threats of punishment were not always needed, and, under some conditions, it is desirable to rely on intrinsic motivation alone.

Incentives and Motivations

Selfish decisions, made by one person, can have very bad consequences for other people and society at large. Law and economics refers to this problem as “social costs” (Coase 1960) or “externalities.” Thus, it is important to temper individual selfishness to guarantee both: the functioning of the market, which rests on the respect of property rights, and the advancement of social welfare that rides on the provision of public goods. How can we discourage *socially* costly behaviors and encourage people to contribute time and money to the public interest? This question lies at the heart of the “new law and economics” (Kornhauser 1988). This strand of literature suggests that law can create material incentives, such as sanctions or rewards, for even the most selfish individuals to think twice before engaging in socially undesirable conduct. For example, tort law creates such external incentives through the determination of damages the injurer has to pay for the harm caused. This should deter potential injurers from harming others.

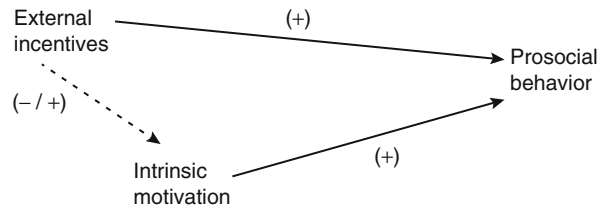
However, the limited and sometimes counter effects of external incentives on pro-social behavior are increasingly pointed out by psychologists (Deci and Flaste 1995) and economists (Frey 1992, 1997a, b). And the notion is gaining popularity within the law and economics community (Bohnet et al. 2001; Bohnet and Cooter 2003; Feldman 2009). The idea is that motivation is in fact dual, composed of an intrinsic element besides an extrinsic one.

Intrinsic motivation was originally defined as some inherent interest in a task (Deci 1971). However, it has long since been expanded to include: environmental morale (Frey and Stutzer 2008), civic duty (Frey et al. 1996), fairness and reciprocity (Fehr and Schmidt 1999), and potentially other pro-social unselfish motivations such as “conscience” (Stout 2011).

One of the major insights from this strand of literature is that trying to minimize the difference between individual and social interests by external incentives may actually undermine intrinsic motivation (Gneezy and Rustichini 2000; Bénabou and Tirole 2003). The two types of

Intrinsic and Extrinsic Motivation, Fig. 1

The interactions between external incentives and intrinsic motivation in increasing pro-social behavior



motivations are effectively interacting to produce the final behavioral response. This is called the inseparability thesis (Bowles 2011). Figure 1 illustrates the relationship among these variables, with the positive and negative signs indicating the direction of the effect.

At the core of the interrelationship between intrinsic and extrinsic motivations are the notions of *competence* and *autonomy* (Ryan and Deci 2000; Tirole 2009; Festré and Garrouste 2014). The prescriptive character of rewards and sanctions, such as imposed by the legal system, may indeed undermine people's confidence in their capacities to voluntarily restraint from undesirable conduct (Bénabou and Tirole 2003) and reduce their sense of autonomy by shifting the locus of control from inside to outside the person (Ryan and Deci 2000). Whether this substitution effect exists and to what extent it plays a part in people's reactions to external incentives is a debated empirical question to which we turn next.

Empirical Tests of Motivation Crowding (Out)

Research on the "crowding-out" effects of extrinsic incentives on people's intrinsic motivations can be divided into two groups. The first group focuses on the effects of rewards and sanctions on people's decision to engage in pro-social behavior. The second strand of literature takes other compliance instruments, which are generally labeled as "control" strategies, and attempts to shed light on their capacity to encourage desired conduct.

The Effects of Material Incentives

Contrary to economic theory's postulate that people supply more of a good or service when its

price goes up (holding all other prices constant), Titmuss's (1970) empirical investigation of blood donations in the United States of America and Great Britain indicates that paying people in return for their blood might decrease altruistic blood donations. The explanation being that under this monetary incentive, donors no longer consider their act as an altruistic blood donation but as an economic transaction. Following this puzzling idea, Frey et al. (1996) designed a field experiment to test the effects of monetary incentives on Swiss households' acceptance of a nuclear waste recycling plant in their neighborhood. They found that tangible rewards decrease acceptance due to the crowding-out of "public spirit." Along similar lines, Fehr and Rockenbach (2003) suggest that when people attribute their chosen behavior to external sanctions, they discount any altruistic motivations. Curiously, in their game, the amount of cooperation is higher in the absence of the sanction for noncooperative behavior. This is congruent with Bohnet et al. (2001) findings that provisions with respect to contract enforcement may crowd-out trustworthiness and negatively impact performance. In addition, an important aspect of this literature is that external incentives' detrimental effects on cooperation are nonmonotonic. In the case of contract enforcement, performance rates are high when the expected cost of breach is sufficiently large, decreasing for medium-sized expected sanctions and increasing again for sufficiently small ones. With respect to rewards for performance, there is a discontinuity at the zero payment in the effect of monetary incentives: "for all positive but small enough compensations, there is a reduction in performance as compared with the zero compensation, or, better, with the lack of any mention of compensation" (Gneezy and Rustichini 2000, p. 802). Another important

component of the crowding-out theory suggests that people do not really care about the outcome of a pro-social behavior per se, but instead care about how their behavior is perceived by others – they like to be perceived as fair (Andreoni and Bernheim 2009). Their behavior is thus driven by “image motivation” (Ariely et al. 2009). Within this perspective, extrinsic incentives have a detrimental effect on pro-social behavior due to the crowding-out of image motivation. Finally, Irlenbusch and Sliwka (2005) have shown that once a monetary incentive mechanism is introduced, its detrimental effect persists even after its removal and it spreads out to other areas by altering people’s cognition of the situation.

The Effects of Control

Another central tenet of economic theory, which is at the core of many law and economics models, is that monitoring improves agents’ performance. The legal system often specifies in advance what type of conduct is admissible and the conditions under which it may be carried out. However, the postulate that monitoring is always performance enhancing is jeopardized by experimental results. For instance, Falk and Kosfeld (2006) designed a principal-agent game in which the principal could specify in advance a minimum effort requirement for the agent to put in. Curiously, they found that “the decision to control significantly reduces the agents’ willingness to act in the principal’s interests” (p. 1611). Agents dislike control because it signals principal’s distrust in their capacities to perform a task. Agents “seem to believe that principals who control expect to receive less than those who don’t, and (...) agents’ beliefs correlate positively with their behavior” (p. 1612). This result is not confined to laboratory experiments. In an econometric study of 116 managers in medium-sized Dutch firms, Barkema (1995) showed that increases in personal control by superiors resulted in a significant reduction in the number of hours worked.

Finally, prescribing in detail a particular behavior through ex ante regulations has been found more detrimental than intervening ex post by sanctioning (rewarding) the undesired (desired) conduct. Motivation effects are indeed an

important aspect to be considered in the tradeoff between ex ante state control and ex post legal intervention. The legal system is less likely to reduce people’s sense of autonomy.

References

- Andreoni J, Bernheim BD (2009) Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. *Econometrica* 77:1607–1636
- Ariely D, Bracha A, Meier S (2009) Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am Econ Rev* 99:544–555
- Barkema HG (1995) Do executives work harder when they are monitored? *Kyklos* 48:19–42
- Bénabou R, Tirole J (2003) Intrinsic and extrinsic motivation. *Q J Econ* 70:489–520
- Bohnet I, Cooter R (2003) Expressive law: framing or equilibrium selection? UC Berkley Public Law Research paper no 138
- Bohnet I, Frey BS, Huck S (2001) More order with less law: on contract enforcement, trust, and crowding. *Am Polit Sci Rev* 95:131–144
- Bowles S (2011) Machiavelli’s mistake. Mimeo, Santa Fe Institute
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Deci EL (1971) Effects of externally mediated rewards on intrinsic motivation. *J Pers Soc Psychol* 18:105–115
- Deci EL, Flaste R (1995) Why we do what we do: Understanding self-motivation. Penguin, New York
- Falk A, Kosfeld M (2006) The hidden costs of control. *Am Econ Rev* 96:1611–1630
- Fehr E, Rockenbach B (2003) Detrimental effects on sanctions on human altruism. *Nature* 422:137–140
- Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. *Q J Econ* 114:817–868
- Feldman Y (2009) The expressive function of trade secret law: legality, cost, intrinsic motivation, and consensus. *J Empir Leg Stud* 6:177–212
- Festré A, Garrouste P (2014) Theory and evidence in psychology and economics about motivation crowding out: a possible convergence? *J Econ Surv*. <https://doi.org/10.1111/joes.12059>
- Frey BS (1992) Tertium datur: pricing, regulating and intrinsic motivation. *Kyklos* 45:161–184
- Frey BS (1997a) Not just for the money: an economic theory of personal motivation. Edward Elgar, Cheltenham
- Frey BS (1997b) A constitution for knaves crowds out civic virtues. *Econ J* 107:1043–1053
- Frey BS, Stutzer A (2008) Environmental morale and motivation. In: Lewis A (ed) *Psychology and economic behavior*. Cambridge University Press, Cambridge, pp 406–428
- Frey BS, Oberholzer-Gee F, Eichenberger R (1996) Te old lady visits your backyard: a tale of morals and markets. *J Polit Econ* 104:1297–1313

- Gneezy U, Rustichini A (2000) Pay enough or don't pay at all. *Q J Econ* 115:791–810
- Irlenbusch B, Sliwka D (2005) Incentives, decision frames, and motivation crowding-out: An experimental investigation. *IZA discussion papers* 1879
- Kornhauser LA (1988) The new economic analysis of law: legal rules as incentives. In: Mercurio N (ed) *Law and economics*. Kluwer, Boston, pp 27–55
- Ryan RM, Deci EL (2000) Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp Educ Psychol* 25:54–67
- Stout L (2011) *Cultivating conscience: How good laws make good people*. Princeton University Press, Princeton
- Tirole J (2009) Motivation intrinsèque, incitations et normes sociales. *Rev Écon* 60:577–589
- Titmuss RM (1970) *The gift relationship*. Allen and Unwin, London

Irregular Sector

► [Informal Sector](#)

Judgment

► Rationality

Judicial Behavior

► Empirical Analysis of Judicial Decisions

Judicial Decision-Making

Alessandro Melcarne
EconomiX, CNRS and Université Paris Nanterre,
Nanterre, France

Definition

Judges are not only the guardians of Human Rights and the Rule of Law. Justices equally play a key role in keeping legal certainty, thus providing a fertile ground for economic transactions. Nonetheless, scholars have only in the last two decades started to openly enquire judicial activity with the lenses of economic analysis. Claiming that judges might be sensitive to certain incentives does not necessarily undermine their independence. However, understanding the economics of judicial decisions is a key to properly

design an institutional framework that might incentivize their work efficiently.

Judicial Decisions and the Economy

A growing economy and sustainable development crucially depend on the functioning of judicial systems. Enforcing institutions have been proven to be fundamental for allowing the reduction of transaction costs and the protection of property rights (Acemoglu et al. 2001; Glaeser et al. 2004; La Porta et al. 1998). Judicial institutions and organizations thus play a fundamental role in the functioning of modern markets. Judges intervene in all litigated economic transactions assuming the role of the impartial “third party” among litigants. However, sometimes the impact of judges’ work might go beyond the mere sphere of interest of the litigants involved in a specific lawsuit. For example, in the case of “blockbuster” class-action cases, the consequences of judges’ decisions might have profound repercussions on society as a whole. Some other times, judicial actors are even called to intervene in the political arena and their behavior might have consequences even for the election of a country’s president (for example, the US Supreme Court in *Bush v. Gore*, 531 U.S. 98).

Despite the relevance of judicial decision-making, economic literature has relatively overlooked this topic. Until recently, few scholars

have tried to shift the rational choice model proper of the *homo-oeconomicus* to judges and their behavior. On the one side, Public Choice literature has focused its attention mainly on legislators and bureaucrats, failing to recognize the public relevance of judicial actors. On the other, legal scholarship has been very skeptical with respect to the idea of conceiving judges with a task that goes beyond the mere mechanical application of the law. According to the “legal formalist” tradition, judges are operating in a sort of *legal empyrean* while isolated from any form of mundane temptation (Melcarne 2017). However, even legal scholars who theoretically do not rule out the idea of interpreting judicial decision-making through the lens of a rational choice enquiry, discard this hypothesis in practice on institutional grounds (Epstein 1990; Epstein and Knight 1997; Heise 2002; Schauer 2000). In this perspective, the overall institutional framework, that we might informally define as judicial independence, has nowadays successfully isolated judges from any sort of “economic” incentive. According to this stream of literature, judicial behavior might be defined within a principal-agent relationship. Within this situation, judges play the role of agents, while the principal would be played, depending on the situation, by litigants, magistrates holding higher positions within the judiciary, politicians, or society as a whole. However, what really matters here is the fact that a relevant share of judges’ activity is characterized by asymmetric information. In this condition, there could be potential grounds for opportunistic behavior on the side of who has greater knowledge. In order to prevent similar circumstances, rules have been designed so to insulate judges from “outside economic incentives” and thus guarantee the credibility and impartiality of judicial enforcement. In this respect, judges are not allowed to accept any form of payment (i.e., bribes) from litigants and not allowed to be involved in extra-judiciary occupations (with some limits). All these institutional arrangements put together are a tool to preserve the genuineness of judges’ work with respect to those that appear before them: being an effective and impartial third-party when solving disputes. In this sense,

it becomes much harder for a judge to cheat and favor someone against what the law prescribes. As a consequence, judicial behavior is correctly determined by the law, as the orthodox view claims, thus leaving insignificant scope for a rational choice enquiry of judges’ actions.

Judges as Utility Maximizers

However, as strict as conduct rules and principal’s control might be, judicial decision-making does never collapses to a mere automatic application of a legal abstract rule to a practical case. A certain degree of freedom (thus discretion) is left to judges’ action; the scope of judges’ autonomy will, at least on a theoretical level, supply ground for their behavior to pursue self-interest and thus be the object of economic analysis (Posner 1993). What might be claimed is that, if judges maximize their utility just as anybody else, the institutional environment in which they have to operate them makes such optimization problem as a “constrained” one (Cass 1995).

Posner (1993) elegantly compares judicial behavior to the act of voting. People vote despite the fact that their individual vote will be marginally irrelevant. In the case of judges, their behavior is very relevant at least for the litigants that appear before them: sometimes, for a vast multitude in the case of courts of last resort. So just as voters extract utility from their votes, the same might be thought for judges. In the act of judging, many objectives might interact with decisions: leisure, reputation, promotion prospects. All these objectives are not strictly in contrast with the institutional design, so are left open to a rational choice individual to be pursued. The degree to which judges will be able to operate rationally, pursuing their personal interests, will depend on how strict the institutional setting is. This will ultimately vary to a great extent not only among different countries and their judiciaries, but also within the same national system, according to the jurisdiction and the hierarchy: the career concerns that might influence judicial decisions are different if considering a trial court judge or one sitting in a court of last resort.

Before Posner (1993), although with some exceptions (Cohen 1991; Cooter 1983; Higgins and Rubin 1980), the general sentiment declined the possibility of shifting economics tools to the investigation of judicial behavior. Scholars considered the overall institutional system granted by judicial independence sufficient to isolate judges from self-interest motivation in their work. Even works that explicitly deal with a formal model of judges' utility maximization (Cooter 1983), do so only with respect to "private" judges, that is, arbitrators. Similar works rule out *ex ante* the possibility that "public" judges (those belonging ordinarily to the judiciary) might behave according to their individual interest. It was only with the sentencing guidelines that studies were able to empirically address this issue (Cohen 1991; Taha 2004). The variance in the way this issue was disposed by judges across the USA allows inferring that it might be due more to judges' incentives, rather than exclusively to legal issues.

The isolation granted by the institutional setting, on the one side limits judges from seeking goals that will undermine the credibility of their impartiality. At the same time, however, such isolation endows judges with a position of advantage to exploit their discretion towards goals that, even if not illegal, might not be considered as strictly related to their duties (career, leisure). The degree of independence they enjoy will determine how binding are the institutional constraints, but the final objective will remain untouched: maximize their utility function. Ambitions are intrinsic in human action: if it is true that judges cannot directly increase their wage by means of their decisions (of course without considering illegal forms of bribing and corruption), there are other potential determinants of their utility functions that might equally bind and determine their decisional processes. In this sense, the research for promotion to higher ranks of the judiciary is even (indirectly) a way for judges to increase their income Melcarne and Ramello (2015). The utility maximization process embedded into judicial decision-making must not necessarily be related to morally doubtful or illegal behaviors. One might think, for example, to the imposition of a specific legal ideal through

judicial decisions, the egoistic use of judicial decision-making as a tool for reputation-building or, sometimes, other rather deplorable judicial behavior as corruption or more in general, the bias of the law from the general interest to a specific one. On the contrary, we might be thinking of much more mundane issues as shirking at the expenses of others' work. In the end, because of the independence enjoyed by members of the judiciary, they are not constrained in the achievement of specific performance goals. At the same time, judges are always free to calibrate their effort according to their career goals (when a career prospect is available) or to their leisure preferences.

Institutions, Incentives, and Judicial Behavior

Every judge is characterized by a peculiar utility function that might vary to different extents from that of an ordinary person. The determinants are probably the same (income, leisure, work satisfaction, reputation), although weighted differently. In general, one might assume that, given the general systems of incentives characterizing judicial systems, the average judge will be characterized by more risk-aversion than a fellow law graduate about to enter legal practice (Posner 2005). Accordingly, judges derive more utility from leisure and public recognition relatively to income. This is because judges' remuneration will not (strictly) depend on their marginal productivity, at least in the short run. Their salary cannot be lowered and they cannot be removed from their job. But indeed this does not exclude the fact that even these individuals might be characterized by a peculiar utility function, more or less constrained by the institutional environment, which they are willing to maximize for their personal well-being.

The main idea of the rational-choice model is that judges are intrinsically self-interested, but the concrete way through which such self-interest manifests itself will depend and vary in intensity according to the institutional framework. This is where differences in judicial systems across countries and even within the same judiciary might affect the study of judges' decisional processes.

The first macro-difference to be considered is the one between career-judiciaries (basically, most of civil law jurisdictions) and those following the Anglo-American model, i.e., recognition judiciaries (Garoupa and Ginsburg 2011). Even within the latter, differences emerge between elected judges and tenured judges. In the case of US elected judges, they will be subject to a vast number of incentives in their behavior, much like an elected politician. First of all, they have a time-limited tenure and so their behavior will change while in office. Second, they not only need to attract votes and thus be able to attract public opinion, but they also need to show good performance and focus their attention on those cases that will be more profitable for them in electoral terms. Furthermore, they need to raise money for their campaign from donors: usually lawyers that litigate in courts. The consequence is that elected judges are less independent, thus implying that their behavior is much more predictable than tenured judges. Agency costs are relatively smaller and thus their behavior is more suitable for economic analysis. Nonetheless, even when the institutional setting is so strict, so that judges are effectively insulated from the more canonical career incentives, this model remains relevant for a number of reasons. This will happen, for example, in the case of US federal district judges, since the probability of being chosen from those ranks to be promoted to courts of appeal is rather low but yet not equal to zero. Furthermore, judges are selected at a relatively older age, after they have already served as practicing lawyers. An environment characterized by such institutional constraints will not determine the total absence of judges' utility, but simply redefine its determinants. In a similar situation, judges are well aware that their decision-making will not affect concretely their future career, but still other factors might determine their decisional process, that are unlikely to be detected by efficient punishing mechanisms.

Since judges' work cannot be normally evaluated, it will be hard to assess if judges are exerting the expected effort. Accordingly, leisure will be a potential channel through which they might exploit their informational advantage. It is

predictable that the same person will be relatively less productive when working as a judge rather than as a lawyer, since the system of incentives is different. For this reason, moonlighting activity is generally prohibited, even if some exceptions are tolerated, as for example, teaching or writing books. However, leisure might be considered as a form of income. More leisure, decreases the amount of pecuniary income demanded by judges at a given quality. However, more leisure implies less work being done and thus more judges needed. In this sense, the entire debate on judicial performance might find a theoretical foundation on these grounds. Given the high standards that are adopted in judges' selection processes, the differences in their performance are hardly explainable only as a matter of different innate abilities or as the consequence of differences in the other productive factors of an ideal judicial production function (e.g., less clerks, or administrative staff, or computers). The strict selection is intended exactly to select only the best ones, but given the specific institutional settings that grant independence to judges, they are not any more bounded by the same set of incentives. Given the general ban of moonlighting jobs, if in a judge's utility function leisure has a much bigger space than careerism, this will induce shirking on her side. Since it is so hard, not only to be identified as lazy, and thus as a "cause" of judicial delay, but even to be "punished" when so, a big incentive will push such judges to slow down their activity, but yielding negative externalities for the judicial system and its users.

In this sense, the institutional isolation that judges enjoy might be a double-edged sword. On the one side, it prevents judges from being "bad," but on the other, it does not force them to be "good." A completely independent judiciary could contribute not to the establishment of the Rule of Law but rather to the "Rule of the Judges." Judiciaries must remain responsive towards society and individual judges must grant a certain level of professionalism and quality. One way to limit the possible negative consequences of such discretion is to combine independence institutions with accountability ones. Make judges accountable of their work is to conceive

a complementarity between two different institutional frameworks, trying to balance opposite incentives among judges (Voigt 2008).

Cross-References

- [Empirical Analysis of Judicial Decisions](#)
- [Independent Judiciary](#)
- [Judicial Delay](#)

References

- Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. *Am Econ Rev* 91(5):1369–1401
- Cass RA (1995) Judging: norms and incentives of retrospective decision-making. *Boston Univ Law Rev* 75:941–996
- Cohen M a (1991) Explaining judicial behavior or what’s “unconstitutional” about the sentencing commission? *J Law Econ Org* 7(1):183–199
- Cooter RD (1983) The objectives of private and public judges. *Public Choice* 41:107–132
- Epstein R (1990) Independence of judges: the uses and limitations of public choice theory. *Brigham Young Univ Law Rev* 3:827–850
- Epstein L, Knight J (1997) The new institutionalism, part II. *Law Courts* 7(2):4–9
- Garoupa N, Ginsburg T (2011) Hybrid judicial career structures: reputation versus legal tradition. *J Leg Anal* 3(2):411–448
- Glaeser E, La Porta R, Lopez-de Silanes F, Shleifer A (2004) Do institutions cause growth? *J Econ Growth* 9(3):271–303
- Heise M (2002) The past, present and future of empirical legal scholarship: judicial decision making and the new empiricism. *Univ Ill Law Rev* 3:819–850
- Higgins R, Rubin P (1980) Judicial Discretion. *J Leg Stud* 9(1):129–138
- La Porta R, Lopez-de Silanes F, Shleifer A, Vishny RW (1998) Law and finance. *J Polit Econ* 106(6):1113–1155
- Melcarne A (2017) Careerism and judicial behavior. *Eur J Law Econ* 44(2):241–264
- Melcarne A, Ramello GB (2015) Judicial independence, judges’ incentives and efficiency. *Rev Law Econ* 11(2):149–169
- Posner RA (1993) What do judges maximize? (The same thing everybody else does). *Supreme Court Econ Rev* 3:1–41
- Posner RA (2005) Judicial behavior and performance: an economic approach. *Fla State Univ Law Rev* 32:1259–1280
- Schauer F (2000) Incentives, reputation, and the inglorious determinants of judicial behavior. *Univ Cincinnati Law Rev* 68:615–636
- Taha AE (2004) Publish or Paris? Evidence of how judges allocate their time. *Am Law Econ Rev* 6(1):1–27
- Voigt S (2008) The economic effects of judicial accountability: cross-country evidence. *Eur J Law Econ* 25(2):95–123

Judicial Delay

Elena D’Agostino¹, Emiliano Sironi² and Giuseppe Sobbrìo¹

¹Department of Economics, University of Messina, Messina, Italy

²Department of Statistical Sciences, Catholic University of Milan, Milan, Italy

Definition

Justice delay is unanimously considered a negative externality both in terms of security and institutional trust (especially when referred to criminal cases) and in terms of economic development (especially when referred to civil cases). Despite several attempts to make systems more efficient have been taken in many countries, it is almost impossible to come to a unanimous solution able to fit different principles, structures, and procedures.

We review the economic literature on this topic, focusing on Gravelle’s (1990) theory and its applications.

Introduction

It is a matter of fact that asking for justice in front of a court of law requires a very long time to get to a verdict in most of European and non-European countries. Justice delay is unanimously considered a negative externality both in terms of security and institutional trust and in terms of economic development. E.g., Jacobi and Shar (2015) pointed out that companies may prefer to breach a loss contract and litigate just to postpone to pay its debts and/or an eventual insolvency. Despite several attempts to make systems more efficient have been taken in many countries, it is almost impossible to come

to a unanimous solution able to fit different principles, structures, and procedures.

The Market for Justice

Economically speaking, every justice system could be interpreted as a market where it is relatively easy to identify a demand (coming from litigants and their lawyers) and a supply (represented by judges, court administrative staff, etc.). This is probably an interesting starting point, well developed in the literature, in order to identify the effects of delay on the demand and the supply of justice. Looking at countries (like Italy) where judicial delay is particularly evident, what emerges is a gap between the demand and the supply sides.

On the demand side, the literature has emphasized the role of lawyers whose main interest is increasing their income. To do so, lawyers push their customers to start new disputes even when the probability of winning is relatively low. Evidence from Italy, a country characterized by strong delay in courts and a high number of lawyers, suggests that this induced demand effect is emphasized when competition among lawyers is higher, that is when the number of lawyers in respect to population is very high (see D'Agostino et al. 2012; Buonanno and Galizzi 2014). Moreover, there is an incentive to come to a dispute even for parties who are unlikely to win whenever legal interest rates are lower than market interest rates (see Marchesi 2007). Reducing the demand of justice, to be intended as the number of new cases filed in front of first instance courts, would have important effects on the justice system as a whole: evidence from Korea shows that lower caseloads would allow first instance judges to spend more effort on each case, improving the quality of their decision and contributing to lower the rate of appeal cases (see Kim and Min 2017).

On the supply side, improper incentive schemes for judges and administrative staff contribute to make more difficult to close disputes within a reasonable time (see Buscaglia and Dakolias 1996; Palumbo and Sette 2006). More in detail, transfers and insufficient numbers of judges combined with the time spent by them in

extrajudicial activities help increase the backlogs of judicial cases (Calvez 2007).

Creating right incentives is not necessarily difficult or expensive: e.g., Melcarne and Ramello (2015) find evidence from different countries that greater independence of judges from politics stimulates competition for professional upgrades, and positively affects court performances. Following a different perspective, Finocchiaro Castro and Guccio (2015) pointed out that inefficiencies and delay could be due not to judge laziness and/or lack of incentives. Focusing on the Italian experience, they find that the lack of managerial ability to manage the increasing demand of new disputes and the backlog damages the system and significantly contributes to determine court inefficiency in Italy (especially in the South of the country).

Furthermore, procedural rules imposing necessary waiting times, settlement attempts, and other actions make it difficult to shorten justice times (see Djankov et al. 2003; for Italy see also Di Vita 2010). An empirical analysis focusing on Belgium and conducted by Bielen et al. (2015) shows that the use of expert assessments has a negative impact on the duration of trials, as well as the number of pleadings. Focusing on Italy, Di Vita (2015) shown that decentralization of legislative production, such as in those sectors regulated by regional laws, has a negative effect on the duration of disputes in front of the regional administrative courts compared to cases where the applicable rules come from the European legislator. Laws could be (relatively) easily changed by reforms of national civil procedural codes, but the experience we have from several attempts made in different countries is that none of them has developed optimal rules and procedures yet. Evidence from Spain rather suggests the 2000 Procedural Law Reform, combined with some important socio-economic factors (like the economic regression), contributed to grow up the number of new civil cases between 1995 and 2010 (see Rosales and Jiménez-Rubio 2017).

If the main reason for judicial delay can be found in an excess of demand, traditional economic theory suggests that an effective policy could be increasing the price to get access to the market (see Vereeck and Muhl 2000). Such a

policy, however, should take into account the peculiarity of the justice system where efficiency must be accompanied by equity.

Gravelle's Theory

Against the conventional wisdom, however, Gravelle (1990) argues that a control based on price does not fit the justice system, but delay is itself a rationing mechanism able to control the demand for justice "until the number of trials demanded by litigants is equal to the capacity of courts."

The starting point of Gravelle's theory is that judicial procedures impose litigants to take decisions in sequence, such that trying to settle their dispute before going in front of a court of law. Gravelle argues that a rationing system based on higher prices does not ensure that litigants will take the right decision. E.g., suppose that litigants are not perfectly informed about the outcome of the dispute in front of a judge and both believe to be likely to win the case: they may prefer to pay a high price (in terms of taxes and other legal costs) to get access to the court instead of reducing their requests and coming to a settlement.

Rather, Gravelle points out that delay imposes a waiting list to litigants. It works as an additional cost able to push litigants to come to a settlement. As a result, the demand for justice should decrease to the sustainable level, that is, the level that the supply is able to cover.

D'Agostino et al. (2013) find empirical support to the Gravelle's theory testing the trend of appeals in Italy for ordinary, labor and social welfare disputes, separately taken, between 2000 and 2006. The authors find that for ordinary and labor disputes the number of new cases proposed in front of a court of appeal decreases as delay in first instance courts increases.

Conclusion

Judicial delay weakens rights, reduces trust among economic agents, and makes transactions less safe.

All these aspects are considered detrimental for the economic activity and are an obstacle for protecting the rights of citizens and of the economic actors. In this framework, it is difficult to suggest common solutions, because delay in courts is a generalized problem (Dakolias 1999), but with specificities that depend on the legal system of each country and of the type of proceeding (Calvez 2007). In addition to the theory exposed by Gravelle, alternative solutions (applied to general contexts) should be taken into consideration in literature to reduce the length of litigations: first, innovations in administration are possibly correlated with a decrease of the delay in courts. Secondly, an increase in the use of information technology is supposed to improve the efficiency of the justice system. All these possible solutions are effective only if combined with a generalized rationalization of the complexity of court rules.

Cross-References

- [Court of Justice of the European Union](#)
- [Litigation and Legal Expenses Insurance](#)

References

- Bielen S, Marleffe W, Vereeck L (2015) An empirical analysis of case disposition time in Belgium. *Rev Law Econ* 11(2):293–316
- Buonanno P, Galizzi MM (2014) *Advocatus, et non latro?: testing the excess of litigation in the Italian courts of justice*. *Rev Law Econ* 10(3):285–322
- Buscaglia E, Dakolias M (1996) *A quantitative analysis of the judicial sector: the cases of Argentina and Ecuador*. World Bank technical paper no. 353
- Calvez F (2007) *Length of court proceedings in the member states of the Council of Europe based on the case law of the European Court of Human Rights*. Council of Europe Publ, Strasbourg
- D'Agostino E, Sironi E, Sobbrío G (2012) *Lawyers and legal disputes. Evidence from Italy*. *Appl Econ Lett* 19(14):1349–1352
- D'Agostino E, Sironi E, Sobbrío G (2013) *The length of legal disputes and the decision to appeal in Italian courts*. *Riv Ital degli Econ* 18(1):47–63
- Dakolias M (1999) *Court performance around the world: a comparative perspective*, vol 23. World Bank Publications, Washington, DC

- Di Vita G (2010) Production of laws and delays in court decisions. *Int Rev Law Econ* 30(3):276–281
- Di Vita G (2015) Centralization versus decentralization of legislative production and the effect on litigation: a case study. *Rev Law Econ* 11(2):267–291
- Djankov S, La Porta R, Lopez-De-Silanes F, Shleifer A (2003) Courts. *Q J Econ* 118(2):453–517
- Finocchiaro Castro M, Guccio C (2015) Bottlenecks or inefficiency? An assessment of first instance Italian courts' performance. *Rev Law Econ* 11(2):317–354
- Gravelle H (1990) Rationing trials by waiting: welfare implications. *Int Rev Law Econ* 10:255–270
- Jacobi O, Sher N (2015) A commitment mechanism to eliminate willful contract litigation. *Rev Law Econ* 11(2):231–266
- Kim D, Min H (2017) Appeal rate and caseload: evidence from civil litigation in Korea. *Eur J Law Econ* 44(2):339–360
- Marchesi D (2007) The rule incentives that rule civil justice. ISAE working paper, no. 85
- Melcarne A, Ramello G (2015) Judicial independence, judges' incentives and efficiency. *Rev Law Econ* 11(2):149–169
- Palumbo G, Sette E (2006) Delayed decision: an application to the court, Working paper
- Rosales V, Jiménez-Rubio D (2017) Empirical analysis of civil litigation determinants: the case of Spain. *Eur J Law Econ* 44(2):321–338
- Vereeck L, Muhl M (2000) An economic theory of court delay. *Eur J Law Econ* 10(3):243–268

K

Keeping Promises and Contracts

Sergio Mittlaender

Max Planck Institute for Social Law and Social Policy, Munich, Germany

Abstract

There are reasons for individuals to keep contractual promises beyond legal remedies and reputational concerns. This entry reviews the leading explanations of promise-keeping behavior as well as recent evidence for the latter as collected in laboratory experiments. In the end, it discusses avenues for future research and recent applications of the results collected so far for the design of contracts and their legal enforcement.

Contracts and Promises

Contracts commit individuals to a future course of action. The contractual obligation is created, in common law systems, by the giving of a promise with consideration, and a contract is defined as “a promise or a set of promises for the breach of which the law gives a remedy” (Restatement, Second, of Contracts §1). Promises are therefore an integral part of contracts, and in imposing an obligation for the promisor to perform, or in inducing expectations in the promisee that she

will receive the promised performance, they are apt to motivate promisors to perform just as the legal remedy and reputational concerns do it.

Incentives created by the prospect of payment of damages for breach or by otherwise applicable legal remedies have been extensively studied in the literature (Shavell 1980, 1984). The promisor must anticipate that she will have to bear the cost imposed by the legal remedy if she does not perform, and legal enforcement thereby provides incentives for promisors to perform. Reputational concerns, involving, for instance, the prospect of losing contractual partners and the loss of future gains from trade associated with it, provide another reason for the promisor to make good on her word. They both create material incentives for promisors to perform and not to break their promises. They are, respectively, third- and second-party enforcement mechanisms, executed by the courts and by the promisee.

An individual that has an obligation to perform, and to do what she promised to the other party, might tend to keep her word for non-monetary reasons that operate in the promisor’s internal value system. The promisor might want to comply with the moral norm that promises ought to be kept – or that *pacta sunt servanda* – or she might want to avoid experiencing guilt from letting down the expectations of the promisee. These are the two main reasons why individuals might keep promises – (1) moral obligation and (2) guilt aversion – even in the absence of legal enforcement and reputational concerns.

There is by now substantial evidence that promises, even when they are only cheap talk (Crawford and Sobel 1982; Farrell and Rabin 1996) and their breach does not impose any material consequence on the promise breaker, are often kept at a cost for the promisor. Promises are subject to first-party enforcement, executed by the promisor herself, and are capable of inducing higher average levels of cooperation, trustworthiness, and sharing in games in which the strictly dominant strategy for the promisor is another one, such as in common-pool resources, holdup, trust, and dictator games (Ostrom et al. 1992; Ellingsen and Johannesson 2004; Charness and Dufwenberg 2006, 2011; Vanberg 2008). For sure, not all individuals keep their promises, but many do so, and the question that remains – and that has been subject to theoretical and empirical investigations recently – is *why* they do so.

The empirical evidence reviewed in this entry relies, most often, on the behavior of trustees in a modified version of the trust game originally developed by Berg et al. (1995). In this game, players have a certain endowment (usually \$10), are anonymously matched in pairs, and then are randomly assigned either to the role of a sender or a receiver. The sender (the first mover) can transfer any amount from her endowment to the recipient (the second mover) and keeps the remaining for herself. The amount transferred is multiplied by a factor larger than one (usually, it is double or tripled), which implies that it is socially optimal to transfer everything. The recipient can then freely decide on whether to reciprocate trust and transfer some amount back to the sender or rather keep the money for herself. The game is played only once, anonymously, and its backward-induction solution is, under traditional economic assumptions, unique: the sender must anticipate that the recipient will not repay any money and is therefore predicted not to send to the recipient any amount in the first place.

In reality, senders most often trust recipients and transfer positive amounts, and recipients often reciprocate by transferring back some amount to the sender. Still, around half of the possible gains from cooperation remain unrealized, as first movers transfer, on average,

only around 50% of their endowment (Johnson and Mislin 2011).

Communication between the parties has been found to significantly increase trust, just as it induces higher average levels of cooperation in other social dilemma games (Sally 1995; Balliet 2010). Subjects often take this opportunity to commit to take a certain future course of action that can lead to higher gains for both parties. The interest herein lies in the possibility for the recipient to send a message to the sender saying that she promises to be trustworthy if the sender trusts her and sends her money. The question is then *why* recipients that make such promises often indeed keep their word and reciprocate trust more often when they have promised than when they have not and in a situation in which they could earn more money by not keeping their word.

Theories of Promise Keeping

The first explanation for why individuals keep promises is provided by the theory of guilt aversion, introduced in economic theory by Charness and Dufwenberg (2006), formalized by Battigalli and Dufwenberg (2007, 2009), and expanded by Miettinen (2013) and Jensen and Kozlovskaya (2016). The theory assumes that individuals derive disutility from guilt when they believe they let the other's expectations down. If an act that could bring net gains to the agent disappoints another person's expectations, it is assumed to induce unpleasant and discomforting feelings of guilt. If the agent is sufficiently guilt averse, then she might refrain from taking that action in the first place, thereby forgoing monetary gains but avoiding the experience of guilt.

The theory of guilt aversion is apt to explain human behavior in various social contexts and is especially stark in the case of promises, for a promise induces the promisee to create expectations that she will receive what she was promised. Therefore, the promisor has a reason to believe that the promisee believes that she will receive the promised performance. The promisor that breaches disappoints expectations that she personally induced in the promisee by her own

autonomous act of making the promise and is predicted to suffer disutility from guilt if she breaks the promise. The promisor must weigh this loss of utility from guilt with the monetary gain that she would achieve by breaking the promise and will keep the promise if the former is larger than the latter.

This explanation for why individuals keep promises does not rely on the moral norm of keeping promises or that *pacta sunt servanda*. Guilt-averse individuals tend to keep promises in order to avoid letting the promisee's expectations down even if they do not care about promises and contracts per se. It may be objected that promises are apt to induce expectations in the promisee, and to alter the latter's beliefs only because such a moral norm exists, and that in its absence, a promise would never be taken seriously and would never cause people to attribute payoff expectations to others. While this argument is sound, it does not invalidate the theory but rather stresses that the existence of the moral norm is a prerequisite for such feelings of guilt and payoff expectations to arise. The motivating factor driving the decision to perform is still the desire to avoid the guilt from betraying the promisee's expectations.

The second explanation for why individuals keep promises relies on individual preferences for keeping promises and for respecting shared and accepted moral norms. The moral norm (Fried 1981) or societal convention (Rawls 1971) of promising establishes that "if one says the words 'I promise to do *X*' in the appropriate circumstances, one is to do *X*, unless certain excusing conditions obtain" (Rawls 1971, 344). Promises are therefore apt to affect the behavior of promisors directly, because individuals suffer disutility from violating the moral norm, and independent of the expectations of the promisee.

An individual that wants to keep the promise in order to comply with the moral obligation that she imposed upon herself by an autonomous act of the will will do so even if she believes that the promisee does not expect to receive the promised performance. This may be so because the promisee does not trust the promisor or perhaps because the promisee does not believe that the promise was serious. The promisor that wants to comply with

the moral norm but that knows that the other does not expect performance still has a reason to perform, and not to breach and thereby violate the moral norm that promises ought to be kept.

A third explanation for why individuals keep promises relies on moral identity and on the individual's beliefs about her own values (Bénabou and Tirole 2011). Although related to the previous explanation, it differs in its content. When deciding to keep or break a promise, the individual considers what type of person each of those actions would make her (a promise breaker or a promise keeper), the desirability of those, and how much the individual values each of these identities. Keeping promises is then similar to an investment in one's own identity, for it increases the individual's confidence that she is a moral person.

Breaking promises often involves gains for the individual, and these are challenges to her beliefs, identity, and self-esteem. Challenges to a weakly held belief ("I believe it is somehow good to keep promises") elicit conformity reactions ("I will therefore keep this promise and increase my own self-esteem from being a moral person"). Effective challenges, such as very high gains from breaking the promise, threaten the individual's belief, as the person might fall into temptation, break the promise, and abandon the identity attached to that belief. When such effective challenges are directed to a strongly held belief ("I believe keeping promises is certainly the right thing to do"), then they induce counteractions that aim at reinforcing the threatened belief and at increasing the payoff of the investment in one's own identity ("I will keep this promise even in this very hard situation and therefore I will tighten my belief that keeping promises is important and, as a consequence, that I am a very moral person").

Empirical Evidence

Charness and Dufwenberg (2006) provide initial evidence for the theory of guilt aversion in an experiment that implements a modified version of the trust game in which the sender could only decide to trust the recipient or not and the recipient

could only decide to repay the sender or not. The recipient had, in their experiment, the option to send a nonbinding, free-form message to the sender before the latter's decision to trust the recipient or not. The messages sent most often included a promise (defined by the authors as any "statement of intent") and were effective in inducing trust despite being, in game-theoretical terms, mere cheap talk in the implemented single interactions. When the trustee sent a message to the trustor that did not include a promise, then only 33% of trustees reciprocated trust. In case the message included a promise, 79% of trustees reciprocated trust and hence kept the promise.

The authors, moreover, elicited the players' beliefs after they played the game, and senders were asked to guess the likelihood of trustworthiness, and recipients were asked to guess the likelihood of trust. Trustors who received a promise, in effect, delivered higher average guesses about the likelihood of trustworthiness (64%) than those who received a message that did not include a promise (50%). Trustees who sent messages including a promise also reported higher guesses about trustors' guesses (63%) than those who sent a message that did not include a promise (55%). The authors concluded that this evidence supports the theory of guilt aversion because promises induced changes in second-order beliefs, as well as in behavior, increasing rates of repayment with respect to the situation in which the message did not include a promise.

Vanberg (2008) argued that Charness and Dufwenberg's experiment did not distinguish guilt aversion from moral obligation, as the promise made by the second mover is correlated with the observed changes in promisors' second-order beliefs. Changes in beliefs may follow behavior because players may believe that others can predict their behavior and the positive correlation between behavior and second-order beliefs induced by promises might be caused by the false consensus effect (Ross 1977). Vanberg implemented a modification of the game to achieve independent variation in promises and beliefs.

In his experiment, in some cases, the promisor was switched, and the promisee faced another

different player in the position of the promisor. The promisor knew whether the promisee had received a promise from another person or not, but promisees did not know that they faced another person. Promises significantly affected behavior and beliefs toward the person to whom the promise was given (in case the promisee was not switched), increasing trustworthiness from 52% to 73%, but promises made by others (in case the promisee was switched) induced only higher beliefs, but not higher trustworthiness. When promisors faced a promisee that had received a promise from another person, they did not reciprocate trust more often than in cases where there was no promise, and independent of the fact that they had higher beliefs about promisees' beliefs, and therefore not in accord with the guilt aversion explanation.

While Vanberg's innovative design provided substantial gains in the understanding of human behavior under promises, it does not necessarily imply that the guilt aversion explanation has no empirical adherence in the context of promises. If promisors tend to keep promises in order to avoid betraying the other's expectations, and experience guilt only if they break their own promises made to the original promisee, then the guilt aversion explanation remains stark. And, in effect, there are substantial moral reasons for why promises bind only those that made the promise, autonomously and freely, and for why one experiences guilt only when betraying the expectations of the person to whom the promisor had promised and not when disappointing the expectations of persons to whom they never promised anything.

Kawagoe and Narita (2014) assumed, in accordance with this argument, that promisors feel guilt when they betray the promisee's expectation only if that expectation was raised by their very own actions, typically by their promises, but find evidence against their theory of personal guilt aversion. Ederer and Stremitzer (2016) develop a lexicographic theory of guilt aversion in which a promisee's expectations matter only if there is a direct promissory link between the promisor and the promisee, and hence only if the promisee's expectations are supported by the promisor's own promise, and provide evidence that, in turn,

corroborates this lexicographic theory of guilt aversion. The evidence in favor of those two explanations is therefore mixed, and in certain experiments, both theories can explain many but not all aspects of the data (e.g., Charness and Dufwenberg 2011).

Conclusion, Implications, and Suggestions for Further Research

There is still a lively debate on the causes of promise-keeping behavior and on the reasons that lead individuals to make good on their word. Further research is necessary not only to resolve this ongoing debate, and to provide evidence for predictions from different theories (such as self-esteem, which remains largely unaddressed), but also to study under which conditions individuals are more likely to break their promises. They might experience guilt from breach of unilateral and bilateral promises differently and might tend to fulfill the moral obligation depending on whether the promise was the result of a bargain or not.

Moreover, the relationship between those reasons for promise-keeping behavior and legal enforcement remains largely unaddressed. While there is evidence on how threats, sanctions, and legal enforcement crowd out the intrinsic motivation to perform (Bohnet et al. 2001), less is known about how those different reasons for keeping promises interact with or perhaps even determine the crowding-out effect. This might be especially relevant when parties had agreed that, in case of breach, they can recur to courts, arbitration, or other means or when they had specified a liquidated damages clause.

Wilkinson-Ryan (2010) provides evidence that individuals would require a lower amount to accept breach of a hypothetical contract that included a liquidated damages clause than in cases where this clause was absent. Moreover, they thought that breaching a contract without that clause was marginally more immoral and more likely to upset the promisee. Hoepfner and coauthors (2017) provide evidence that liquidated damages, stipulated unilaterally by the principal,

induce lower levels of effort by the agent than exogenously set expectation damages and that while the first one creates gains in removing moral considerations that hamper socially efficient breaches, it also reduces levels of cooperation.

It would be interesting to investigate whether these results are due to moral concerns or to stronger feelings of guilt and whether a liquidated damages clause creates, in the promisor's view, an option for her to breach without violating a moral norm that would apply in the absence of such clause. Alternatively, liquidated damages might induce breaches because the promisor does not experience guilt in the same manner as she would experience in the absence of such clause, as the promisor might believe that the promisee does not invariably expect to receive performance in kind but instead expects to receive performance *or* damages in cases where parties bargained and agreed on such clause.

Some scholars have started to apply the theory and findings to contract design. Charness and coauthors (2013) argue that since breaking a promise creates a self-imposed cost to the promisor, contracts should assign the responsibility for installing the promise to the party who has the residual right to breach. They report that when workers propose the contract and make a promise specifying the effort level of the worker that is due, then actual effort choices and welfare are higher than when the firm proposes the contract or the target for effort.

Stone and Stremitz (2016) apply the theory of guilt aversion to reliance investments by promisees and hypothesize that these will overinvest in reliance in order to increase the guilt that promisors would experience when breaking the promise. Legal remedies for breach decrease the need for the promisee to rely on such "psychological lock-in" in order to increase the likelihood of performance and hence tend to decrease overinvestment with respect to the situation in which there is no legal enforcement. In effect, the authors find that overinvestment is higher in the absence of damages than in their presence and that those who overinvest do so to a higher extent in the absence of remedies than in their presence.

With respect to drafting contracts, Depoorter and Tontrup (2012) report how subjects were willing to sacrifice substantial gains to prevent breach when specific performance was the available remedy and how such standard made the moral norm of keeping promises more salient to promisees, hampering efficient renegotiation. Brooks and coauthors (2012) provide evidence that framing contracts in terms of losses, including deductions in case a low output is achieved, induces higher levels of effort than contracts framed in terms of gains, including bonuses in case of high outputs. Lastly, Wilkinson-Ryan (2012) finds that the transfer of contractual rights induces behavioral changes and erodes the likelihood and quality of performance by the promisor.

The inquiry into the causes of promise-keeping behavior and the collected evidence for the predictions from different theories that are apt to explain such behavior is fascinating not only in its own terms but also in its capacity to inform contractual design and contractual enforcement. It provides insights that are relevant for drafting contracts and for enforcing them, but further investigations, both theoretical and empirical, are needed to better understand how they *should* be done to provide for the welfare of the parties and of society.

Cross-References

- [Experimental Law and Economics](#)
- [Incomplete Contracts](#)
- [Intrinsic and Extrinsic Motivation](#)

References

- Balliet D (2010) Communication and cooperation in social dilemmas: a meta-analytic review. *J Confl Resolut* 54:39–57
- Battigalli P, Dufwenberg M (2007) Guilt in games. *Am Econ Rev* 97:170–176
- Battigalli P, Dufwenberg M (2009) Dynamic psychological games. *J Econ Theory* 144:1–35
- Bénabou R, Tirole J (2011) Identity, morals, and taboos: beliefs as assets. *Q J Econ* 126:805–855
- Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econ Behav* 10:122–142
- Bohnet I, Frey B, Huck S (2001) More order with less law: on contract enforcement, trust, and crowding. *Am Polit Sci Rev* 95:131–144
- Brooks R, Stremitz A, Tontrup S (2012) Framing contracts: why loss framing increases effort. *J Inst Theor Econ* 168:62–82
- Charness G, Dufwenberg M (2006) Promises and partnerships. *Econometrica* 74:1579–1601
- Charness G, Dufwenberg M (2011) Participation. *Am Econ Rev* 101:1213–1239
- Charness G, Du N, Yang C, Yao L (2013) Promises in contract design. *Eur Econ Rev* 64:194–208
- Crawford V, Sobel J (1982) Strategic information transmission. *Econometrica* 50(1431):1452
- Depoorter B, Tontrup S (2012) How law frames moral intuitions: the expressive effect of specific performance. *Ariz Law Rev* 54:673
- Ederer F, Stremitz A (2016) Promises and expectations. Cowles foundation discussion paper No. 1931
- Ellingsen T, Johannesson M (2004) Promises, threats and fairness. *Econ J* 114:397–420
- Farrell J, Rabin M (1996) Cheap talk. *J Econ Perspect* 10:103–118
- Fried C (1981) Contract as promise. Harvard University Press, Cambridge, MA
- Hoepfner S, Freund L, Depoorter B (2017) The moral-hazard effect of liquidated damages: an experiment on contract remedies. *J Institutional and Theoretical Econ* 173:84–105
- Jensen M, Kozlovskaya M (2016) A representation theorem for guilt aversion. *J Econ Behav Organ* 125:148–161
- Johnson N, Mislin A (2011) Trust games: a meta-analysis. *J Econ Psychol* 32:865–889
- Kawagoe T, Narita Y (2014) Guilt aversion revisited: an experimental test of a new model. *J Econ Behav Organ* 102:1–9
- Miettinen T (2013) Promises and conventions – an approach to pre-play agreements. *Games Econ Behav* 80:68–84
- Ostrom E, Walker J, Gardner R (1992) Covenants with and without a sword: self-governance is possible. *Am Polit Sci Rev* 86:404–417
- Rawls J (1971) A theory of justice. Harvard University Press, Cambridge, MA
- Ross L (1977) The “false consensus effect”: an egocentric bias in social perception and attribution processes. *J Exp Soc Psychol* 13:279–301
- Sally D (1995) Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Ration Soc* 7:58–92
- Shavell S (1980) Damage measures for breach of contract. *Bell J Econ* 11:466–490
- Shavell S (1984) The design of contracts and damages for breach. *Q J Econ* 99:121–148
- Stone R, Stremitz A (2016) Promises, reliance, and psychological lock-in. UCLA School of Law, Law & Economics Research Paper No. 15–17
- Vanberg C (2008) Why do people keep their promises? An experimental test of two explanations. *Econometrica* 76:1467–1480

- Wilkinson-Ryan T (2010) Do liquidated damages encourage breach? A psychological experiment. *Mich Law Rev* 108:633–671
- Wilkinson-Ryan T (2012) Transferring trust: reciprocity norms and assignment of contract. *J Empir Leg Stud* 9:511–535

Knowledge

Brishti Guha
Centre of International Trade and Development,
School of International Studies, Jawaharlal Nehru
University, New Delhi, India

Abstract

This essay focuses on two specific aspects of knowledge that make it interesting from the viewpoint of law and economics. First, we consider the nature of knowledge as a public good, discussing how non-excludability, non-rivalry, and positive externalities in knowledge production affect private incentives and generate a demand for state intervention. We discuss the role of patent and copyright laws, as well as other forms of state support, in this context. Secondly, considering knowledge as information, we examine the economic implications of asymmetries in information, and the extent to which laws and the legal mechanism can help solve the problems created by asymmetric information.

Definitions

(i) The sum of what is known. (ii) Facts, information, and skills acquired through education or experience

Introduction

Knowledge has several properties and facets that make it interesting from a law and economics viewpoint. I concentrate on two main facets of knowledge here: knowledge as a public good

and the implications of an even or uneven *distribution* of knowledge (symmetries or asymmetries in *information*).

Knowledge as a Public Good

Once a piece of knowledge is produced, it is “non-rival” and “non-excludable.” Non-excludability means that once something becomes known, everyone has access to this piece of knowledge. This will be true unless it is possible to restrict the spread of the knowledge to a limited number of people (who pay for this privilege, something we will discuss in more detail shortly). Non-rivalry means that when one person benefits from the new piece of knowledge, this does not decrease the amount of knowledge available to others. In economic terms, the marginal cost of providing the knowledge to one more person is zero (Arrow 1962).

Non-rivalry and non-excludability together make knowledge a “public good.” From the viewpoint of social efficiency, since it costs society nothing to share an already produced piece of knowledge with one more person, one could argue that, once a new piece of knowledge comes into being, the efficient solution would be free dissemination. That is, the non-excludability property should be preserved. However, this creates a problem, because it is at odds with private incentives to *produce* this knowledge in the first place.

Knowledge production requires considerable investment of resources and time. It is essentially a risky activity which may not yield returns. Therefore, an individual will only make such investments – and come up with knowledge-enhancing innovations – if he expects a high return from doing so. However, if the knowledge he produces immediately becomes freely available to everyone, his innovation can easily be replicated. No one has an incentive to pay him for his innovation, and he in turn has no incentive to undertake an expensive and risky activity.

Unless the state steps in, therefore, the amount of knowledge produced will be far below the socially efficient level. There is another, related, reason for this. Knowledge produces “positive externalities.” That is, when people acquire

knowledge, this benefits even those who do not acquire this knowledge in the first place. This is true not just of new, startlingly innovative knowledge; it is also true of education, which typically involves imparting knowledge which has been in existence for some time. For example, when children learn about basic healthcare at school, they can share this knowledge with their (perhaps less educated) parents. Another example is the effect on crime. When fewer people are uneducated, there is less crime, so that even people who have relatively little education or knowledge benefit from a safer society. Innovative knowledge generates positive externalities of its own. New techniques learnt by workers in one firm can easily spill over into other firms if some of these workers change their jobs after mastering these techniques. Positive externalities, however, do not enter private calculations; individuals who generate the knowledge (or provide the education) know that they will not be able to capture all of the benefits of this knowledge and therefore produce less knowledge or provide less education than is desirable from the viewpoint of society.

The “free-rider” problem generated by the public good character of knowledge – the problem that anyone can freely replicate an original idea or creation – has become particularly severe in the Digital Age since the Internet makes a great deal of knowledge – articles, music, and books – available. However, it has also been an important force in earlier times. For instance, in medieval times, countries which prospered were countries that were able to make major discoveries regarding new sea routes, acquiring information about ocean currents and winds, the geography of oceans, coastlines, islands, and indeed a whole new world (Guha and Guha 2014). As such information rapidly becomes common knowledge and also because investing in acquiring such information is immensely costly and risky, the countries which made these discoveries were also the ones that received major state support in the initial phases of exploration and discovery.

These issues directly impinge on one major area of law and economics, that dealing with patent or copyright infringement. The state may try to promote incentives to come up with new pieces of knowledge by assuring individuals of

their “intellectual property rights.” This involves granting the producer of the original piece of knowledge a monopoly (for a fixed period of time before the expiry of the patent or copyright) over his creation. Prohibition of free replication and sharing thus protects the innovator’s returns, at least for a certain period of time, and enhances incentives for knowledge creation. Patents sometimes create problems of their own; a strictly enforced patent regime, for example, may also give rise to patent trolls (firms which apply for patents not because they want to develop the patented product, but to prevent other competitors from doing so). It may also induce firms to constantly sue other firms claiming patent infringement (claims which may sometimes be frivolous). This imposes some social costs in terms of the legal resources spent on such lawsuits; also, potential new knowledge producers might be discouraged lest their contribution unintentionally violate an existing patent. At the same time, given digital piracy concerns, protecting the intellectual property rights of artists (particularly musicians) is becoming increasingly challenging.

Other ways in which the state could try to resolve the problem of underproduction of knowledge by the market include direct state subsidization of research. This was what Spain and Portugal did in the medieval era (the crown funded voyages of discovery). A vast fraction of US scientific and technological knowledge (computers, lasers, aerospace, etc.) is the direct or indirect product of state-supported research, particularly defense research. Since the provision of education by the market is also below the social optimum (because of knowledge externalities and because those who impart education wish to restrict access to those who can pay), the state can also intervene in education by providing a public education system.

Knowledge as Information: Informational Asymmetries and Law and Economics

If one thinks of knowledge as information, the distribution of this information among parties engaged in an economic transaction has major implications. Consider what happens when this

information is asymmetrically distributed (*informational asymmetries*). This provides scope for the informed party to cheat the uninformed party (Akerlof 1970). For example, consider a seller and a buyer; suppose the buyer cannot tell the quality of the good he is buying on inspection. However, the seller knows the quality, either because he has produced it himself or because he has been using it for some time (in the case of secondhand sales). This generates a temptation, a *moral hazard* on the seller's part to pass off an inferior product as a high-quality product and extract a high price from the buyer while disposing of a low-quality good. This temptation is particularly strong in relatively anonymous contexts, where the seller does not have a reputation to maintain. However, the fear that this might happen might keep buyers from entering the market in the first place. Therefore, informational asymmetries can generate market collapse. In contrast, if information – or ignorance – is symmetrically distributed, moral hazard or market collapse does not arise. If buyers are informed in addition to sellers, it is obvious that they can assess the true quality of the product, accordingly leaving sellers with no opportunity to cheat them; the two can then agree on a price based on the true quality of the product. More surprisingly, there is also no problem if both buyers and sellers are *uninformed*. If the sellers are ignorant of quality themselves, they cannot systematically pass off low-quality goods as high-quality ones; buyers realize this and pay a price based on average quality, and the market does not break down.

Again, asymmetry in the distribution of information is an important issue from a law and economics perspective. Product warranties, for instance, are one solution to the abovementioned problem; they allow markets to exist by giving customers some assurance of quality. Customer protection laws also help. Ordinarily, contracts guarantee both parties in an economic transaction some rights; however, a contract can only be upheld in court if its provisions are “verifiable.” Verifiability means that a third party – lawyers or judges – should be able to check whether the contract was, indeed, violated, and it must be possible to do this without incurring a prohibitively high cost. However, asymmetry of information often

implies that contracts are *incomplete*. It is not possible to establish in court whether all their provisions were honored; this knowledge is *private information* known only to the informed party in the transaction. In this event, these contracts are not legally enforceable and become informal understandings rather than formal legal contracts. Such understandings may be enforced by means which do not rely on the legal mechanism. Some examples of such means are reputation (the ability to spread information about cheats may induce them not to alienate potential new customers by cheating existing ones), peer monitoring (in cases of group activity, see Ghatak 1999), long-term (repeated) relationships (if a customer may become a repeat customer, it may not be worthwhile to alienate him for some quick profit), competition, and the ability to signal quality convincingly (Spence 1973).

References

- Akerlof G (1970) The market for “Lemons”: quality uncertainty and the market mechanism. *Q J Econ* 84:488–500
- Arrow K (1962) Economic welfare and the allocation of resources for invention. In: *The rate and direction of inventive activity: economic and social factors*. Princeton University Press, Princeton, pp 609–625
- Ghatak M (1999) Group lending, local information and peer selection. *J Dev Econ* 60:27–50
- Guha, AS, Guha, B (2014) Reversal of fortune revisited: the geography of transport and the changing balance of world economic power. *Rivista di Storia Economica* 30 (2):161–188.
- Spence M (1973) Job market signaling. *Q J Econ* 87:355–374

Knowledge-Specific Patents and the Additionality Constraint

Cristiano Antonelli
Dipartimento di Economia e Statistica “Cognetti de Martiis”, Università di Torino, Collegio Carlo Alberto, Torino, Italy

Definition

This entry elaborates the implications for a new knowledge policy of the full range of effects of

its limited appropriability and exhaustibility. The analysis of the appropriability trade-off identifies the dual effects of knowledge spillovers consisting not only in the reduction of incentives to the generation of knowledge but also in the reduction of R&D costs stemming from the access to the quasi-public stock of knowledge. The positive effects of knowledge spillover may compensate for the reduction of the price of innovated goods with its well-known negative effect in terms of reduced pay out of R&D activities and eventual underproduction of knowledge. The appreciation of the full spectrum of possibilities of the appropriability trade-off further refined by the analysis of the effects of the limited exhaustibility of knowledge and of the distinction between imitation and knowledge externalities enables to articulate a new framework of knowledge policy. These results, in fact, suggest the introduction of knowledge-specific patents based upon varying levels of exclusivity and call attention on the relevance of the actual additionality of public subsidies to research and development activities performed by firms.

Introduction

The recent advances of the economics of knowledge about the properties of knowledge as an economic good question the foundations of the basic framework upon which current knowledge policies are built. This entry advocates the introduction of knowledge-specific patents based upon varying levels of exclusivity and calls attention on the relevance of the actual additionality of public subsidies to research and development activities performed by firms.

The new analysis of the properties of knowledge as an economic good has shed new light on their economic effects with the identification and appreciation of the appropriability trade-off, the limited exhaustibility of knowledge, and its intrinsic heterogeneity.

The analysis of knowledge as an economic good enabled to appreciate its idiosyncratic characteristics. At first the analysis focused its limited

appropriability and, in a second stage, its limited exhaustibility. The owner of a standard economic good can fully appropriate the entire stream of benefits that stem from its use. This does not take place with knowledge because of the uncontrolled leakage: third parties can appropriate part of its economic benefits. The appropriability trade-off appreciates the dual effects of spillover where the reduction of R&D costs stemming from the access to the quasi-public stock of knowledge may compensate for the reduction of the price of innovated goods with its well-known negative effect in terms of reduced pay out of R&D activities and eventual underproduction of knowledge (Arrow 1962; Antonelli and David 2015).

The appreciation of the limited exhaustibility of knowledge is the result of a second stage of investigation of the economics of knowledge. Economics assumes that standard economic goods wear and tear. Their – repeated – use has negative consequences on their functionality. Food, once consumed, disappears and cannot be consumed again. Durable consumer and capital goods have an intrinsic limited time horizon that is further reduced by the frequency of use. The rate of usage of standard durable goods reduces their functionality. This does not take place with knowledge: knowledge can be used and used again without any negative consequence on its functionality. The limited exhaustibility of knowledge is not only physical but also economic. A standard durable economic good may become obsolete, even if it retains all its physical functionalities, because of the introduction of a superior good. The case of economic obsolescence, as distinct from a physical obsolescence, may take place with knowledge. Economic obsolescence applies to knowledge only to an extent, however. The “old” vintages of knowledge do retain an indispensable role as inputs in the recombinant generation of new knowledge. Next to its limited appropriability, knowledge is characterized, also, by a limited exhaustibility (Antonelli and Link 2015).

The identification of the limited exhaustibility of knowledge enables to qualify and specify the appropriability trade-off and to appreciate

its positive effects in terms of duration and cumulativeness that may compensate and occasionally overcompensate the negative effects of its limited appropriability (Antonelli 2017).

The identification of the new properties of knowledge as an economic good calls attention on the merits and strengths of knowledge as an economic good and its pivotal role as the engine of knowledge externalities and consequent dynamic increasing returns and questions the Arrowian assumption about the limits and weaknesses of knowledge because of its limited appropriability and the consequent market failure in terms of knowledge undersupply stemming from the lack of appropriate levels of incentives to its generation and use.

The new appreciation of the economic properties of knowledge enables also the identification of its intrinsic heterogeneity according to the varying levels of natural appropriability and exhaustibility. Knowledge should be regarded as a bundle of heterogeneous knowledge items with varying levels of exhaustibility and appropriability, rather than a homogeneous and undifferentiated stock.

This new approach that highlights the merits, as opposed to the limits, of knowledge as an economic good and its heterogeneity as opposed to its homogeneity, has strong implications in terms of knowledge policies.

Knowledge policies have been heavily influenced by the large consensus about the Arrowian knowledge market failure. The analysis of the negative economic consequences of the limits of knowledge as an economic good has been implemented for several decades and provided the base for an articulated set of economic policies aimed increasing the incentives to the generation of knowledge deemed to be insufficient.

The cornerstone of the knowledge policy that stems from the hypothesis of a generalized knowledge market failure due to the limited appropriability of knowledge is the implementation of the intellectual property rights regime designed to help reducing the negative effects of the limited appropriability and the provision of public subsidies designed to reducing the costs

of knowledge-generating activities so as to balance the negative effects of the persistent limited appropriability of proprietary knowledge even after the protection provided by homogeneous patents characterized by high levels of exclusivity.

The aim and the use of these tools have been criticized in terms of efficiency and effectiveness. As Heller and Eisenberg (1998) show, the strengthening of the intellectual property right regime that has characterized recent decades may actually deter innovation and make the case for anticommons. The current intellectual property right regime together with high transaction costs in the markets for knowledge and excess expectations of patentees regarding the value of their knowledge assets produce a fragmented knowledge landscape where owners of small complementary bits of knowledge are unable to participate in the collective effort that is needed to generate new knowledge as an output while using existing knowledge as input (David and Hall 2006). The consideration of the appropriability trade-off stemming from the dual role of knowledge spillovers as the cause of the reduction of the benefits that the “inventor” can appropriate and yet an essential facility indispensable for the recombinant generation of new knowledge (Antonelli 2007) and a cause of the knowledge externalities that help reducing the costs of new knowledge so as to compensate for the negative effects of the limited appropriability adds to the need of a new knowledge policy framework (Antonelli 2017).

The evidence shows on the other hand that automatic public subsidies and tax credits granted to R&D activities performed by private firms yield only a limited net increase of the R&D actually carried out. A large literature confirms not only that the elasticity of the actual levels of R&D activities performed by the firms that receive automatic subsidies to their R&D activities to public subsidies is well below one, but also that the additionality of public R&D subsidies – measured by the ratio of the public subsidy to the value of the additional R&D activities actually carried out – is below 100%. An important share of public subsidies

substitutes internal funds with a crowding out effect (David et al. 2000).

A New Knowledge Policy Framework

The results of the analysis carried on so far question the generalized application of the standard Arrovian framework and its strong implications in terms of knowledge policies. The limited exhaustibility of knowledge yields positive effects that may compensate and occasionally overcompensate the negative effects of its limited appropriability. The understanding of the appropriability trade-off suggests that the dissemination of knowledge helps strengthening the role of knowledge spillovers as an essential facility indispensable for the recombinant generation of new knowledge (Antonelli 2007) and the knowledge externalities that help reducing the costs of new knowledge so as to compensate for the negative effects of the limited appropriability (Antonelli 2017).

The actual increase of the amount of R&D expenditures carried out by firms matters as soon as the effects of the limited exhaustibility of knowledge in terms of the long-term accumulation of a stock of knowledge with beneficial effects on knowledge costs. This dynamics is all the more relevant when the positive effects of the limited appropriability of knowledge are considered. The knowledge externalities stemming from nonexhaustible knowledge spillovers benefit all the system. Hence, the larger are the flows of R&D activities at each point in time and the faster the accumulation of the quasi-public stock of knowledge, the larger the amount of knowledge externalities that become available.

The new evidence calls for a new knowledge policy framework able to take into account the joint effects of both the appropriability trade-off and exhaustibility of knowledge.

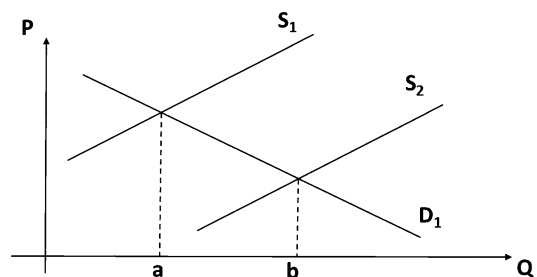
The provision of public subsidies to R&D activities and the intellectual property right regime should be differentiated so as to take into account the full array of outcomes of the varying levels of appropriability and exhaustibility of the bundle of heterogeneous knowledge items. Let us explore the matter in turn.

R&D Subsidies and Additionality Requirements

The rationale of the provision of public subsidies including tax credits to R&D activities carried out by firms needs to be reconsidered. The standard assumption upon which R&D subsidies and tax credits are expected to support innovation is synthesized by Fig. 1.

Automatic public subsidies and tax credits should reduce the costs of R&D activities: the supply curve of knowledge shifts from S_1 to S_2 . For a given derived demand for knowledge, the equilibrium solution changes, and on the horizontal axis, the equilibrium amount of knowledge actually generated increases from a to b . Public intervention is able to restore the equilibrium conditions that apply to standard goods and remedy the intrinsic problems of knowledge as an economic good. The system, now able to produce the right amount b of knowledge, will reach the Pareto equilibrium condition increasing the general amount of goods produced. This argument, however, rests on strong implicit assumptions about the actual shape of the derived demand for knowledge. The effects of R&D subsidies and tax credits are strongly influenced by: (i) the slope of the derived demand for knowledge and (ii) the amount of the shift. Let us consider these effects in turn.

The larger is the slope of the derived demand, the larger the output elasticity of knowledge in the technology production function. When the slope of the derived demand is large, the reduction of R&D costs engendered by public subsidies yields only a limited increase of the actual amount of



Knowledge-Specific Patents and the Additionality Constraint, Fig. 1 Expected effects of R&D subsidies

R&D activities performed by the firms that receive the subsidies.

The effects of the subsidies and tax credits on the actual levels of R&D activities depend also and primarily upon the original position of the R&D cost curve and the amount of subsidies. The downward shift from large cost levels is likely to have larger effects than additional subsidies that lower further the position of the R&D cost curve. The large empirical evidence about the limited elasticity of automatic public subsidies including tax credits to the actual levels of R&D activities and the low levels of actual additionality of public subsidies and tax credits can be interpreted as the effect of a kink in the derived demand for knowledge. For behavioral reasons, firms are reluctant to expand their R&D activities beyond a given level (Clarysse et al. 2009) (see Fig. 2).

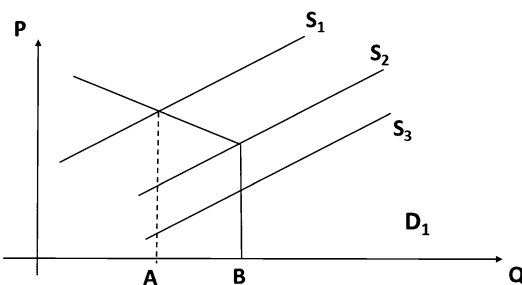
The derived demand for knowledge before the kink is more price elastic than after it. The slope of the derived demand reflects the willingness of firms to increase their R&D activities. Before the kink, additional R&D activities and the related introduction of innovations are compatible with the current organization and strategy of firms. When the amount of R&D expenditures reaches a point, however, firms become reluctant to increase further their commitment to research activities. Additional R&D activities would imply radical changes in the organization of the firm and in its strategies. Beyond that point – the kink – the price elasticity shrinks.

The effects of public subsidies that reduce the costs of knowledge are represented by the shifts of

the supply curve from S_1 to S_2 and S_3 . Because of the kink, the subsidies that made possible the shift from S_1 to S_2 have positive effects on the equilibrium levels of knowledge generated and used in the economic system. The subsidies that made possible the shift of the supply curve of knowledge from S_2 to S_3 , instead, have negligible effects on the equilibrium levels of R&D activities. The shift brought the supply curve S_3 beyond the kink. Here crowding out effects take place. Public subsidies and tax credits have reduced the cost of knowledge but have not induced any increase in the amount of R&D activities actually performed. The amount of technological change that was deemed necessary by firms, according to the conditions of product and factor markets into which they are embedded and to whose changes they try and react by means of the introduction of innovations and hence the amount of knowledge that was required, has been already put in place. Additional funds targeted to reduce the cost of knowledge are welcomed as they do reduce the costs of the given amount of knowledge that was deemed necessary by the firm's strategy, but by no means can be used to introduce further innovations. The actual effect of the subsidies that helped to shift the knowledge supply curve from S_2 to S_3 is the use of the financial resources made available for other purposes. R&D subsidies in this case have low levels of additionality.

The lower the additionality is, the larger is the amount of the public subsidies and the larger the output elasticity of knowledge in the technology production function. The issue of the low levels of actual additionality of public subsidies to R&D becomes crucial as soon as the positive effects of R&D activities in terms of synchronic knowledge externalities are taken into account. R&D subsidies, in fact, benefit not only the recipient firm that is compensated of the negative effects of the limited appropriability of knowledge, but the whole system provided that the subsidies yield positive effects in terms of additional volumes of R&D activities and hence additional flows of spillover and consequent knowledge externalities and rates of increase of total factor productivity.

The additionality of public subsidies is easily defined by the ratio of additional R&D



Knowledge-Specific Patents and the Additionality Constraint, Fig. 2 The “true” derived demand for knowledge and the risks of crowding out

expenditures of the recipient firm i ($R\&D_i$) to the amount of public subsidies (SUB_i) received by the very same firm.

The lower is the exhaustibility, the larger the additionality should be ($\delta R\&D_i/SUB_i > 1$). The lower the exhaustibility, the larger the positive welfare effects that stem from the dynamic increasing returns associated with the increasing size of the stock of quasi-public knowledge.

Public subsidies that yield only the reduction of R&D costs and do not lead to net additional R&D activities may be justified only in a perfect Arrovian market failure when appropriability is very low and exhaustibility very high.

According to the actual levels of appropriability – and related knowledge spillover with the consequent reduction of revenue for inventors as well as increase of knowledge externalities – and exhaustibility, a range of possible additionality requirements can be considered:

- (i) Automatic R&D subsidies and tax credits with high levels of additionality requirements should be granted to knowledge items with high levels of appropriability and low levels of exhaustibility. The markets for knowledge are able to provide quasi-optimal incentives and public subsidies are mainly targeted to increase the size of the stock of quasi-public knowledge and to empower the dissemination of additional spillovers.
- (ii) Automatic R&D subsidies and tax credits with low levels of additionality requirements can be granted to knowledge items with low levels of appropriability and high levels of exhaustibility. The classical Arrovian market failure applies and R&D subsidies do compensate for the missing incentives.

Not only the types of additionality requirements but also the mix of public interventions should reflect the actual levels of appropriability and exhaustibility. Crowding out, in fact, is less likely to take place when public subsidies and tax credits target specific research programs. In this case, the State can use the provision of selective public subsidies to effectively influence the levels

and the direction of the research activities according to the public relevance of the research projects that have been submitted by firms. Targeted research subsidies and tax credits moreover help directing the research activities of a variety of firms so as to increase their compatibility and complementarity helping the dissemination of knowledge spillovers and related knowledge externalities at the system level (Antonelli and Crespi 2013).

The direct supply of knowledge by means of a public infrastructure consisting of dedicated research centers has indeed played a major role in increasing the supply of knowledge and its dissemination in the economic system especially when and if it has been coupled with dedicated and targeted public interventions especially on the demand side activating the competent demand pull mechanisms, the provision of targeted research incentives and most importantly the active role of state owned enterprises (Antonelli et al. 2014).

Table 1 summarizes the arguments elaborated so far.

The differentiation of public interventions according to the levels of knowledge appropriability and exhaustibility should concern also the structure of intellectual property right regimes.

Knowledge-Specific Patents and the Additionality Constraint, Table 1 Types of knowledge and knowledge policies

	Low exhaustibility	High exhaustibility
High appropriability	Automatic R&D subsidies and tax credits with strong additionality requirements Targeted R&D programs Direct supply of knowledge	
Low appropriability		Automatic R&D subsidies and tax credits with weak additionality requirements

Knowledge-Specific Patents: Exclusivity and Length

The analysis of knowledge properties and their effects suggests to explore the possibility to introduce knowledge-specific, rather than industry-specific, patents and to modulate not only their length (and breadth) but also their levels of exclusivity according to their actual levels of natural appropriability and exhaustibility. The case for a differentiated intellectual property rights regime based upon two types of knowledge-specific patents rests on three distinct arguments.

First, the classical argument in favor of the introduction of differentiated patents relies exclusively upon the analysis of the types of market forms of the different industries (Gilbert and Shapiro 1990). The literature that explores the optimal length and breadth of patents concentrates the analysis upon the types of competition and the market structure of industries and does not explore the differences of the knowledge items. According to this line of analysis, the length of patents should be industry-specific: the more intense is competition; the higher the productivity of R&D activity; the more intricate the reverse engineering and the shorter should be the duration of the patent (Mosel 2011). Yet it is clear that the very same knowledge may be relevant in a broad array of industries and many different types of knowledge are relevant within the same industry. The cases of general purpose technologies, such as ICT (Information and Communication Technologies) and biotechnologies, show how large can be the scope of some knowledge items that apply to a variety of well distinct industries with well-differentiated types of competition and market structure (Helpman 1998).

The analysis of general purpose technologies unveils a second major limit of the current literature of optimal breadth and length of patents: it misses the crucial distinction between imitation externalities and knowledge externalities. Imitation externalities apply when knowledge spills within the same industry and rivals can use it in their technology production function as an input at costs below equilibrium. Intraindustrial spillovers engender imitation externalities that

in turn engender the fall of the price-cost-margin of inventors with the well-known negative consequences in terms of missing incentives and knowledge undersupply. This is not the case of knowledge externalities that take place when knowledge spills across industries and is not ready-to-be-used again, but can be used only as an input in the generation of new knowledge. Knowledge externalities have no effects on the price of goods produced with knowledge in the original product market, but strong positive effects on the costs of knowledge that is used as an input for the introduction of innovations in other product markets. In this case, the outcome of the appropriability trade-off does not confirm the Arrovian postulate (Antonelli 2017).

Third, the literature has paid much attention to the possible differentiation of patents in terms of length and breadth but has paid little attention to the possibility to differentiate intellectual property rights according to their level of exclusivity. The introduction of knowledge-specific patents with both compulsory licensing and royalties enables to explore the possibility of varying levels of exclusivity as a second layer of differentiation of patents.

The design of a differentiated intellectual property right regime, based upon two types of knowledge-specific patents, should take into account the actual levels of exhaustibility and natural appropriability of the knowledge items, and of their scope of application whether intra- or interindustrial, and concern not only the length and breadth of patents but also their exclusivity. It is necessary, in fact, to pay attention not only to the negative effects of intraindustrial spillovers and imitation externalities in terms of reduced incentives but also to the positive effects of knowledge externalities and of the limited exhaustibility of knowledge. The generalized exclusivity of intellectual property rights has strong negative effects in terms of reduction of knowledge externalities that stem from the delays to the access to knowledge.

The introduction of compulsory licensing to the use of proprietary knowledge as an input in the generation of knowledge relevant in other industries different from that of inventors might

help reducing the negative effects of the standard exclusivity of the current intellectual property right regime. The definition of a fair level of royalties is clearly crucial. In this context, the identification of an optimum level of royalties paves the way to implementing an intellectual property right regime based upon the liability rule. Users have the right to access proprietary knowledge provided they pay a fair royalty. Patent holders cannot impede the use of their proprietary knowledge. The selective introduction compulsory licensing can play a major role to limit the anticommons effects of the current intellectual property right regime. In the recombinant knowledge generation process, formalized by the knowledge generation function, knowledge is both an output and an input. The impediment to use of knowledge as an input that shares the characteristics of an essential facility can block the rate of inventive activity and is in any case the cause of major costs in terms of research duplication and inventing around (Antonelli 2007).

Following Griliches (1979) and Weitzman (1996), it seems possible to assume that the inputs of the knowledge generation function are: (i) the current R&D expenditures, (ii) the internal stock of knowledge, and (iii) the external knowledge extracted from the quasi-public stock of external knowledge. The price of knowledge affects directly the revenue. The cost of knowledge as an input affects the costs. In a Cobb-Douglas specification of the knowledge generation function, standard substitution – albeit limited by substantial complementarity – between the three basic inputs implies that the changes in their cost affects total and average costs with nonlinear effects. All changes in the price of knowledge have, instead, linear effects on the revenue. It is consequently possible to identify an optimum price of knowledge and use it to implement the working of compulsory licensing-cum-royalties (Antonelli 2013).

The identification of a fair level of royalties stems from two distinct layers of analysis: (i) the analysis of the effects of the actual levels of appropriability and exhaustibility on the knowledge generation function at the firm level and (ii) the analysis of the effects of the actual levels

of appropriability and exhaustibility at the system level. Let us analyze them in turn.

At the firm level, the actual levels of appropriability and exclusivity of specific knowledge items have direct effects on the fair level of the price of knowledge and hence on the levels of royalties. The price of knowledge items characterized by low levels of natural appropriability should be larger than the price – and royalty – levels of knowledge items characterized by high levels of natural appropriability. The cost of knowledge items characterized by high levels of exhaustibility is larger than the cost of knowledge items characterized by low levels of exhaustibility: they can be used again and again with low levels of wear and tear. Hence the fair level of the royalties should be lower for knowledge items characterized by low levels of exhaustibility and larger for knowledge items characterized by high levels of exhaustibility.

Taking into account the two layers of analysis and the distinction between imitation and knowledge externalities, it seems possible to modulate exclusivity according to the levels of royalties associated to compulsory licensing. In intellectual property regimes, knowledge-specific patents are characterized by the inclusion of patents with compulsory-licensing-cum-royalties. In this case, intellectual property rights are no longer exclusive: licensing is compulsory, but not free. The levels of the royalties define the levels of *de facto* exclusivity. High-level royalties imply strong *de facto* exclusivity. Low-level royalties imply weak *de facto* exclusivity.

Intellectual property right regimes characterized by high exclusivity should not be granted to knowledge items characterized by low levels of exhaustibility for two strong reasons: (i) low exhaustibility compensates already the negative effects of low appropriability, (ii) exclusive property rights for knowledge items with low levels of exhaustibility have large opportunity costs at the system level as they limit and delay the knowledge spillovers and hence the access to knowledge externalities and the recombinant generation of new knowledge (Antonelli 2017).

The design of differentiated intellectual property rights in terms of both terms and exclusivity

Knowledge-Specific Patents and the Additionality Constraint, Table 2 Types of knowledge and externalities and types of knowledge-specific patents

	Low exhaustibility	High exhaustibility
High appropriability	Short-term patents with weak exclusivity especially for interindustrial uses	
Low appropriability		Long-term patents with strong exclusivity especially for intraindustrial uses

levels should be implemented according to the intrinsic heterogeneity of knowledge in terms of both natural appropriability and exhaustibility with the introduction of knowledge-specific patents:

- (i) Short-term patents with weak exclusivity should be granted to knowledge items characterized by low exhaustibility and high natural appropriability and apply to interindustrial users.
- (ii) Long-term patents with strong exclusivity can be confirmed only to knowledge items with high levels of exhaustibility and very low levels of natural appropriability and apply to intraindustrial users (Table 2).

Conclusions

The new understanding of the appropriability trade-off and of the limited exhaustibility of knowledge questions the basic frame of current knowledge policy. The appreciation of the positive effects of the limited appropriability of knowledge in terms of reduced knowledge costs and of the limited exhaustibility of knowledge on its intertemporal derived demand and on the long-term duration of the positive effects of the spill-over contrasts the foundations of the Arrovian postulate. The current frame of knowledge policy builds up the Arrovian postulate according to

which the limited appropriability of knowledge is itself a cause of market failure consisting in the undersupply of knowledge because of the lack of appropriate incentives to generate it. Consistently, the current frame of the knowledge policy aims at increasing the incentives to the generation of knowledge so as to remedy the negative effects of its intrinsic limits. The appreciation of the appropriability trade-off and of the limited exhaustibility of knowledge, instead, calls attention on the opportunities for growth that stem from the properties of knowledge in terms of dynamic increasing returns rather than on its limits as a cause of market failure.

The appreciation of the new properties of knowledge and the distinction between imitation and knowledge externalities provide basic support to an alternative knowledge policy framework that should be aimed at strengthening the opportunities for growth that stem from the increase of the amount of knowledge available in a system at each point in time as a source of knowledge externalities.

Public subsidies to R&D activities performed by firms should be finalized to increase the actual amount of knowledge generated in the system, rather than to compensate firms for the missing revenue caused by the limited appropriability of knowledge by lowering knowledge costs. The actual additionality of public subsidies to R&D activities should become the basic constraint and the basic target. Public subsidies should be granted to firms that are actually able to use the subsidies not only to reduce their R&D costs but also to increase the amount of knowledge generated.

Intellectual property rights characterized by intrinsic exclusivity have been justified and implemented to remedy and increase the low levels of natural appropriability of knowledge. The exclusivity of patents has major negative consequences in terms of reduced dissemination of knowledge that limits the access and use of proprietary knowledge as an input into the recombinant generation of new knowledge. Exclusive intellectual property rights do increase appropriability and hence the incentives to generate knowledge but are the cause of major

opportunity costs in terms of missing opportunities to reduce the actual costs of new knowledge. Their use should be limited to intraindustrial uses.

The design of intellectual property rights with limited exclusivity based upon compulsory licensing cum fair royalties seems better able to combine the necessary support to knowledge appropriation and hence to the incentives to its generation together with a reduction of the severe limits caused by exclusivity to the dissemination of knowledge and hence to the levels of knowledge externalities that are crucial to support growth. Such patents should apply to all uses of knowledge outside the industry of inventors.

Finally, it seems necessary a radical departure from the standard assumption of the homogeneity of knowledge that underlies the current frame of knowledge policy. It is by now necessary to regard knowledge as a bundle of highly differentiated items. Knowledge items differ substantially in terms of levels of natural appropriability and exhaustibility. Knowledge policies should be differentiated accordingly so as to modulate both subsidies and the design of intellectual property rights according to their levels of natural appropriability, exhaustibility, and scope of usage.

Knowledge-specific patents as well as differentiated public subsidies with varying levels of additionality requirements should be designed and implemented to take into account the actual properties of the knowledge item under consideration.

References

- Antonelli C (2007) Knowledge as an essential facility. *J Evol Econ* 17:451–471
- Antonelli C (2013) Compulsory licensing: the foundations of an institutional innovation. *WIPO J* 4:157–174
- Antonelli C (2017) Endogenous innovation. The economics of an emergent system property. Elgar, Cheltenham
- Antonelli C, Crespi F (2013) The “Matthew effect” in R&D public subsidies: the case of Italy. *Technol Forecast Soc Chang* 80:1523–1534
- Antonelli C, David PA (eds) (2015) The economics of knowledge and the knowledge driven economy. Routledge, London
- Antonelli C, Link A (eds) (2015) Handbook of the economics of knowledge. Routledge, London
- Antonelli C, Barbiellini Amidei F, Fassio C (2014) The mechanisms of knowledge governance: state owned corporations and Italian economic growth, 1950–1994. *Struct Chang Econ Dyn* 31:43–63
- Arrow KJ (1962) Economic welfare and the allocation of resources for invention. In: Nelson RR (ed) The rate and direction of inventive activity: economic and social factors. Princeton University Press for NBER, Princeton, pp 609–625
- Clarysse B, Wright M, Mustar P (2009) Behavioral additionality of R&D subsidies: a learning perspective. *Res Policy* 38:1517–1533
- David PA, Hall BH (2006) Property and the pursuit of knowledge: IPR issues affecting scientific research. *Res Policy* 35:776–771
- David PA, Hall HB, Toole AA (2000) Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Res Policy* 29:497–529
- Gilbert R, Shapiro C (1990) Optimal patent length and breadth. *Rand J Econ* 21:106–112
- Griliches Z (1979) Issues in assessing the contribution of research and development to productivity growth. *Bell J Econ* 10(1):92–116
- Heller MA, Eisenberg RS (1998) Can patents deter innovation? The anticommons in biomedical research. *Science* 280:698–701
- Helpman E (ed) (1998) General purpose technologies and economic growth. MIT Press, Cambridge
- Mosel M (2011) Competition imitation, and R&D productivity in a growth model with industry-specific patent protection. *Rev Law Econ* 7:601–652
- Weitzman ML (1996) Hybridizing growth theory. *Am Econ Rev* 86:207–212

L

Labeling

John M. Crespi
Iowa State University, Ames, IA, USA

Synonyms

[Certification Labeling](#); [Marking](#)

Definition

Labeling is the affixing of a mark and symbol or identifying word or phrase to a product or process for the purpose of providing information about an undetectable attribute. The attribute could be a product ingredient, a manufacturing process, or an aspect of the production process anywhere along the supply chain.

Labeling

The economic study of *labeling* generally refers to the research regarding the welfare impacts of descriptions, marks, or symbols affixed to a product for the purpose of providing consumer information, typically concerning a credence attribute. A credence attribute (Darby and Karni 1973) exists when consumers cannot detect the attribute even after consuming the product and must rely

upon provided labeling information. Examples of labeled attributes are nutritional labels, organic labels, eco-labels, or safety information. Such attribute labels concern either the final-stage manufacturing of the product or some aspect of the production process itself. Further examples include wine appellations, production methods such as “sustainably grown” or “bird friendly,” the inclusion or exclusion of genetically modified ingredients, “fair trade,” “country of origin,” or even more well-known labels such as single malt or Champagne, which refer to processes or regions that imbue the product with a demand-shifting attribute. Along with being credible, labels, to be successful, must overcome coordination and free-riding problems, and as such, many fall under governmental regulations. In the United States, marketing orders and checkoff programs allow growers to jointly market a product using a producer association label (e.g., “Certified Angus Beef”), while other nations use similar regional and product appellations. Some labels are quite broad, encompassing large classes of products such as France’s “Label Rouge” program to indicate quality. Labels may be for attributes desirable by all consumers as in the case of food safety or nutrition (e.g., vertical differentiation) or desirable to some consumers, while other consumers show relative indifference as in the case of organics and genetically modified ingredients (e.g., horizontal differentiation). The label can be either positive, indicating the presence of the

attribute, or negative, indicating the attribute's absence. Most economic research on labeling has examined descriptions on food products.

In its broadest sense, a label could encompass branding and other private-party signaling although more often research on labeling focuses on governmental or producer association labels not directly affiliated with any one manufacturer where truth in labeling statutes might be difficult to apply (Giannakas 2002). For example, while consumers might trust a winery's claim that the grapes used are 100% cabernet, the consumer may be less inclined to trust the winery on whether its workers were paid a fair wage or whether the grapes were organically produced. Conversely, such a winery may wish to build a reputation for incorporating such attributes in its manufacturing, and seeking out a third-party certifier could be a smart business decision.

The economic research often takes as a datum that consumers would not trust the manufacturer to correctly self-report a credence attribute. There are practical reasons for this. It may be impossible to detect through chemical testing, for example, whether a firm's ingredients are organic, and it would be arduous for consumers to detect whether a firm's suppliers follow environmental safeguards. Certification agents (public or private) who specialize in such detection are seen as necessary in cases where the labels signal the production methods, regional sourcing, environmental impacts, and safety or quality of a good. The absence of the label for a desirable attribute creates a so-called lemon problem (Akerlof 1970) where consumers who have a higher willingness to pay for a good with some attribute (e.g., free of antibiotics or genetically modified ingredients) or produced using a preferred production method (e.g., kosher or dolphin safe) have no way of discerning the quality in the absence of the label. If other firms could make the claim without adhering to the labels' standards, consumers would learn to distrust any labeled good resulting in a crowding out of the desired attribute despite the presence of consumers who would purchase it.

Numerous studies have examined consumers' willingness to pay for or avoid paying for particular credence attributes (see Caswell and Mojduszka 1996 and the extensive discussions

of studies in Krarup and Russell 2005). Research on the upstream environmental benefits of labeling consumer products is also an area of much research (Krarup and Russell 2005). Along with the welfare implications in general, research on policy effectiveness have incorporated models of product differentiation and market structure to examine the impacts of labeling on vertical coordination, market power, and international trade (Bonroy and Constantatos 2015; Lapan and Moschini 2007; Roe and Sheldon 2007), reducing costly searches (Teisl and Roe 1998), and whether eco-labeling in particular might be counterproductive to environmentalists' desires (Matoo and Singh 1994). Research also delves into label proliferation, which arises when products contain multiple, sometimes competing labeled, attributes (Kiesel and Villas-Boas 2013; Marette 2014); behavioral economics through framing (how the wording of a label impacts welfare; Levin and Gaeth 1988; Crespi and Marette. 2003a, b); comparing label policies with other policies (Bonroy and Constantatos 2015; Marette and Roosen 2011); voluntary versus mandatory labeling (Roe et al. 2014); and the impact of labels on brain functions (Bruce et al. 2014).

Cross-References

- [Food Safety](#)
- [Horizontal Product Differentiation](#)

References

- Akerlof G (1970) The market for "Lemons": quality uncertainty and the market mechanism. *Q J Econ* 84:488–500
- Bonroy O, Constantatos C (2015) On the economics of labels: how their introduction affects the functioning of markets and the welfare of all participants. *Am J Agric Econ* 97:239–259
- Bruce AS, Lusk JL, Crespi JM, Cherry JBC, McFadden BR, Savage CR, Bruce JM, Brooks WM, Martin LE (2014) Consumer brain responses to controversial food technologies and price. *J Neurosci Psychol Econ* 7:164–173
- Caswell J, Mojduszka M (1996) Using information labeling to influence the market for quality in food products. *Am J Agric Econ* 78:1248–1253
- Crespi JM, Marette S (2003a) Some economic implications of public labelling. *J Food Distrib Res* 34:83–94

- Crespi JM, Marette S (2003b) “Does Contain” vs. “Does Not Contain”: how should GMO labelling be promoted? *Eur J Law Econ* 16:327–344
- Darby M, Karni E (1973) Free competition and the optimal amount of fraud. *J Law Econ* 16:67–88
- Giannakas K (2002) Information asymmetries and consumption decisions in organic food product markets. *Can J Agric Econ* 50:35–50
- Kiesel K, Villas-Boas SB (2013) Can information costs affect consumer choice? Nutritional labels in a supermarket experiment. *Int J Ind Organiz* 31:153–163
- Krarup S, Russell CS (eds) (2005) *Environment, information and consumer behaviour. New horizons in environmental economics*. Edward Elgar, Northampton
- Lapan H, Moschini G (2007) Grading, minimum quality standards, and the labelling of genetically modified ingredients. *Am J Agric Econ* 89:769–783
- Levin IP, Gaeth GJ (1988) How consumers are affected by the framing of attribute information before and after consuming the product. *J Consum Res* 15:374–378
- Marette S (2014) Economic benefits coming from the absence of labels proliferation. *J Agric Food Ind Organ* 12:1–9
- Marette S, Roosen J (2011) Bans and labels with controversial food technologies. In: Lusk JL, Roosen J, Shogren JE (eds) *The Oxford handbook of the economics of food consumption and policy*. Oxford University Press, Oxford, pp 499–519
- Matoo A, Singh HV (1994) Eco-labelling: policy considerations. *Kyklos* 47:53–65
- Roe BE, Sheldon I (2007) Credence good labelling. The efficiency and distributional implications of several policy approaches. *Am J Agric Econ* 89:1020–1033
- Roe BE, Teisl MF, Deans CR (2014) The economics of voluntary versus mandatory labels. *Ann Rev Resour Econ* 6:1–21
- Teisl MF, Roe BE (1998) The economics of labeling: an overview of issues for health and environmental disclosure. *Agric Resour Econ Rev* 27:140–150

Laffer Curve

► Laffer Effect

Laffer Effect

Francesco Forte
Department of Economics and Law, Sapienza - University of Rome, Rome, Italy

Abstract

The Laffer effect, that takes its name from Arthur Laffer, the economist who presented

it in discussions to support tax cuts by US President Ford (1974–1977), consists of the increase of the tax revenue caused by reductions of tax burdens. This principle had already presented in the past by Suetonius (119/122), Pufendorf (1672), Hume (1742), Montesquieu (1748), in various contexts, either for the maximization of tax revenues or for that of national wealth and welfare. In contemporary economics, the tax cuts to increase tax revenue and GDP have been theorized in a supply side and public choice approach by James Buchanan and others, either as mere tax policies or in broader supply side-policy frame, as that of deregulation. The Laffer effect may be misunderstood through fiscal illusions. It has often been muddled with Keynesian demand-side approaches.

Definition

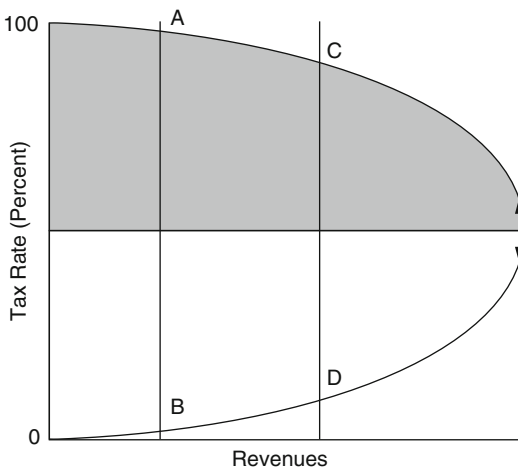
Effect of tax rate reduction which does not reduce the tax revenue but may even increase it.

Synonyms

Laffer curve

Laffer curve takes its name from Arthur Laffer. Wanniski (1978) writes that this economist and professor of Business Economics at the University of Southern California – and adviser of the president of the USA Gerard Ford in 1974–1977 – presented it in a discussion, to support a tax cut, drawing the curve of Fig. 1 and telling that “There are always two tax rates that yield the same revenues.”

The tax rate is on the vertical axis, while the revenues are on the horizontal axis. The revenue depends from the size of the taxable basis (which does not appear in the graph) and from the level of the rate. The tax rate affects the revenues through its relation with the public expenditure and through its effect on the behavior of the taxpayers. Initially an increased revenue devoted to public expenditure gives benefits greater than the cost of



Laffer Effect, Fig. 1 The Laffer curve

the taxes in terms of loss of wealth and incentives. Therefore, the revenue increases both because the rate increases and because the taxable basis increases. Subsequently the benefits of the cost of the tax for the taxpayers overcome the benefits of public expenditure, and the negative effects of the tax increase reduce the taxable basis at an increasingly rate. Therefore, the revenue increases at a reduced rate because the reduction of the taxable basis reduces the revenue effect of the increase of the tax rate. After a point, the taxable basis diminishes in a proportion greater than the increase due to the rate increase and the revenue diminishes. Thus, any amounts of revenue except the point of maximum revenue may be achieved with two alternative tax rates as those indicated with $A = B$ and, respectively, $C = D$. Wanniski, however, appears to give a *wrong supply-side* explanation of the Laffer curve because for him the maximum rate seems the best not only from the point of view of revenue maximization but also from the point of view of production maximization. This is not true because the maximum revenue does not imply a maximization of the economic activity taxed as shown clearly by Monissen (1985, 1999). The situation, indeed, is similar to that of a monopolist, as pointed out by Brennan and Buchanan (1980), dealing with a “Leviathan state.” Wanniski, as Laffer’s predecessor, quotes Montesquieu and Hume. But while in Montesquieu the principle of the optimal tax rate

seems to be in relation to the maximization of the budget revenue (Montesquieu (1748), Book 13, Chapter VII, writes that the free government maximizes its revenue when the revenues of taxpayers are maximized and, in Chapter XV quoted by Wanniski, adds that the free government that increases too much its taxes shall lose the power in favor of a despotic state with lower taxes.), in Hume (According to Hume (1742), VIII § 8, increases of taxes may induce taxpayers to increase their efforts, but, after a point, the opposite taxes place.), it appears to be in relation to the maximization of the production taxed, which may imply the maximization of the national wealth (The same is true for Adam Smith (1776) fourth maxim of taxation, as for taxes that obstruct the industry of the people or discourage them.). With this ambiguity, the Lafferian principle goes back to Suetonius (Suetonius (119/122), Chapter III, The life of Tiberius, § 32) – who refers that the emperor Tiberius “to governors who recommended burdensome taxes for his provinces, wrote in answer that it was the part of a good shepherd to clear his flock, not skin it.”

Thus, the Laffer curve presents two ambiguities: it may be conceived for a partial equilibrium, with GDP as given, or for a general equilibrium. While in the short run the GDP may be given, in the longer term it may change, through the effect of taxation and of exogenous factors, and this may imply that the tax revenue Laffer effect differs from the tax revenue/GDP effect. Therefore, many fiscal illusions may arise (Fedeli and Forte 2014a, b). Pufendorf (1672, Book VIII, Ch. I §5) mentioned by Gerloff (1948) (see Gerloff (1948) pp. 2010–2014 on the Gesteiz der Verringerung der Steuerfälle in Blankart (1991)) presents a partial equilibrium case of the revenue of an excise tax in a scenario of international or interregional tax competition, where a government may get a gain of revenue, by reducing its tax to the level of that of competing governments.

In Forte, Bondonio, and Jona (1980), p. 48, I presented the (partial equilibrium) curve of Fig. 2 similar to that of Laffer, with inverted axes, where the tax rate and the taxable basis are considered *under a given GDP*. The taxable basis is negatively influenced by the tax

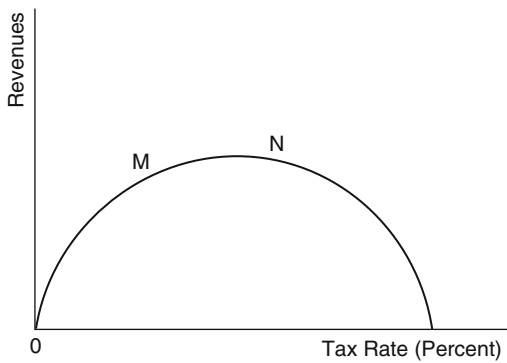
avoidance – done by evasion, elusion, and outflow of the taxable matter – caused by rate increases. A reduction of the rate may increase the tax revenue, at any level of GDP, but then also GDP may change and the elasticity of the taxable basis to GDP may change.

A similar Lafferian curve is given by Gutman (1981) and by Frey and Weck (1983) as presented in Blankart (1991) Chapter 11, § 5), considering the effects of the level of the tax rate on shadow economy. In these cases the tax revenue/GDP ratio in the subsequent years is influenced both by the effect of tax burden on the shadow economy and by the way in which it is assessed in GDP. Buchanan and Lee (1982a, b) argue that reduction of taxes with balanced budget, in

medium term, may give increased GDP and additional revenues, thus allowing an increased public spending, with a balanced budget and a smaller government size as in Fig. 3 presented in Frey and Weck (1985), p. 103, which is valid also for the tax revenues/GDP ratio.

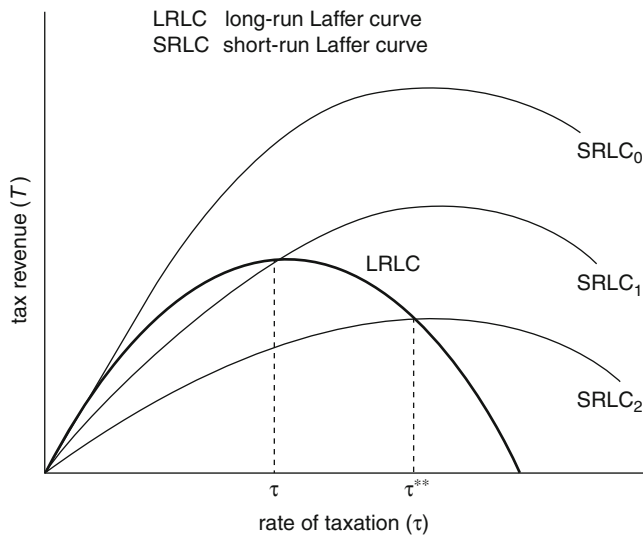
Later on J. Buchanan connected these Laffer effects to increasing returns and work ethics (Forte 2008).

A dangerous ambiguity of the Laffer curve, in the short and longer term, derives from its extension to the Keynesian tax reduction from a demand-side point of view, approved by Laffer (2004) (The ambiguity of Laffer’s own view is increased by the fact that he considered Kennedy’s tax cut, inspired by the Keynesian doctrine, as an application of his principle. Canto, Joines, and Laffer (1982), however, put the Laffer curve only in a supply-side context.), quoting Keynes (1933) who argues that a reduction of taxes *in deficit* may increase the national income by increasing the demand, thus bringing back a balanced budget at a higher level of income (Keynes (1933), p. 338, “taxation may be so high as to defeat its object, and that, given sufficient time to gather the fruits, a reduction of taxation will run a better chance than an increase of balancing the budget”). The deficit-tax cut theorized by Keynes and many neo-Keynesians introduces a fiscal policy “Trojan horse” that may lead



Laffer Effect, Fig. 2 Forte (1982)’s Laffer curve

Laffer Effect, Fig. 3 The long- and the short-run Laffer curves



to a “democracy in deficit” Buchanan and Wagner (1977). Hypothetical future Laffer effects may be invoked to justify a popular budget unbalance and redistributions of taxation from consumptions to savings and from poor to rich. But while in a Keynesian world this can increase GDP by increasing consumption, the opposite may happen untrue in a supply-side approach.

Tax illusions are very relevant as for the Laffer effect. Forte (1987) shows that a fiscal bureaucracy pursuing the gross revenue maximization may not maximize the net revenue because the marginal costs of collection may exceed the marginal increase of revenue.

Laffer effects are also conditioned by the institutional scenario. Tax cuts joint with a reduction of regulations have caused Laffer effects as shown by Fedeli and Forte (2008) and Fedeli, Forte, and Zangari (2008) for reductions of social security contribution accompanied by deregulation of the labor market. Trabandt and Uhlig (2009), within a neoclassical model of growth, assessed how much the USA and EU may still increase taxes without a negative Laffer effect. Fedeli and Forte (2014a) demonstrate by empirical research on OECD countries that high deficits and high taxes may reduce GDP with implications for long-run Laffer effect. Tanzi (2014) defines the Laffer curve as a muddle. Nevertheless, in this muddle, one carefully looking may find gold nuggets.

Cross-References

► Fiscal System

References

- Blankart CB (1991) Öffentliche Finanzen in Demokratie. Verlag Vahlen, München
- Brennan G, Buchanan JM (1980) The power to tax: analytical foundations of a fiscal constitution. Cambridge University Press, Cambridge
- Buchanan JM, Lee DR (1982a) Tax rates and tax revenues in political equilibrium. *Econ Inq* 20(3):344–354
- Buchanan JM, Lee DR (1982b) Politics, time and the Laffer curve. *J Polit Econ* 90:816–819
- Buchanan JM, Wagner R (1977) Democracy in deficit. The political legacy of Lord Keynes. Academic, New York
- Canto Victor A, Douglas H. Joines, Robert I. Webb (1982) The revenue effects of the Kennedy tax cuts. In: Victor A. Canto, Douglas H. Joines, and Arthur B. Laffer (eds) Foundations of supply-side economics – theory and evidence. Academic Press, New York
- Fedeli S, Forte F (2008) The Laffer effects of a program of deregulation cum detaxation: the Italian reform of labor contracts in the period 1997–2011. *Eur J Law Econ* 27:211–232
- Fedeli S, Forte F (2014a) Political and fiscal illusion and the Laffer curve reflections on an Italian case. In: Backhaus J (ed) Fiscal sociology. Peter Lang, Frankfurt am Main, pp 95–114
- Fedeli S, Forte F (2014b) Deficits, tax burden and unemployment. In: Forte F, Mudambi R, Navarra PM (eds) A handbook of alternative public economics. Elgar, Cheltenham, pp 116–139
- Fedeli S, Forte F, Zangari E (2008) An econometric analysis of the employment and revenue effects of the Treu reform in the period 1997–2001. *Riv Polit Econ* XC VII:215–248
- Forte F (1987) The Laffer curve and the theory of the fiscal bureaucracy. *Public Choice* 52(2):101–124
- Forte F (2008) On the ethics of the Laffer curve. In: Forte F (ed) Money, markets and morals. Acedo, München, pp 75–84
- Forte F, Bondonio PV, Jona L (1980) Il sistema tributario. Boringhieri, Torino
- Frey BS, Weck H (1988) Bureaucracy and shadow economy: a macro-approach. In: Hanusch H (ed) Anatomy of government deficiencies. Berlin, Springer, pp 89–105
- Gerloff W (1948) Die öffentliche Finanzwissenschaft, Vol. I, Frankfurt, Klostermann, II ed
- Gutman PM (1981) Implications of the subterranean economy. In: Bove RX, Klingenstein RD (eds) Wertheim’Grund economy conference, Wertheimpp, pp 31–58
- Hume D (1742) Essays moral, political and literary. Liberty Fund, Indianapolis
- Keynes JM (1933) The means to prosperity. In: Essays on persuasion, vol IX (1972) of collected writings, Macmillan, Cambridge University Press, pp 335–366
- Laffer A (2004) The Laffer curve. Past, present and future. Heritage Foundation
- Monissen HG (1985) Optimale Staatsgröße in einem Einkommensteuerregime. In: Milde H, Monissen HG (eds) Rationale Wirtschaftspolitik in komplexen Gesellschaften. W. Kohlhammer, Stuttgart
- Monissen HG (1999) Explorations of the Laffer curve, Würzburg economic papers 99-09 reedited in Forte F (2008), Money, markets and morals. Acedo, München, pp 61–83
- Montesquieu C (1748) De l’esprit des Lois. Garnier-Flammarion, Paris
- Pufendorf De Jure Naturæ et Gentium Libri Octo (1672) Liber VIII, I, 5, English trans. by G. Carew (1729), as of the Law of Nature and Nations: Eight Books, Buffalo, New York, William S. Hein reprint, London, Aris 1729 available on Google

- Smith A 1937 (1776) An inquiry into the nature and causes of the wealth of nations. Modern Library, New York
- Suetonius, (119/122), The lives of the twelve Caesars (English trans: Rolfe JC, Teubner, 1907)
- Tanzi V (2014) The Laffer curve muddle. In: Forte F, Mudambi R, Navarra PM (eds) A handbook of alternative public economics. Elgar, Cheltenham, pp 104–115
- Trabandt M, Uhlig H (2009) How far are we from the slippery slope? The Laffer curve revised. NBER working paper no 15343
- Wanniski J (1978) Taxes, revenues, and the Laffer curve. Public Int 50:3–16

Language of Economics

Magdalena Bielenia-Grajewska
Intercultural Communication and
Neurolinguistics Laboratory, Department of
Translation Studies, Faculty of Languages,
University of Gdansk, Gdansk, Poland

Abstract

The aim of this contribution is to discuss the characteristic features of economic discourse. Moreover, the language of economics is studied through the prism of domains, approaches, and perspectives used to investigate the complexity of economic communication. In addition, different methods of researching the language of economics are presented and discussed, including interdisciplinary methodologies.

Introduction

Modern economics, as other domains in life, can be characterized by different features of the twenty-first century. One of the key factors shaping modern economic reality is globalization, represented in, among other things, relatively easier access to different resources, of material and nonmaterial character, that have been previously limited because of, among other things, geographical or technological barriers. The power of technology (in terms of transport and communication possibilities), represented in its influence on the global market, has led to the growing importance

of linguistic skills, indispensable to meet the needs of stakeholders coming from different parts of the world. Since the modern economic emporium has shrunk as far as geographical or technological distance is concerned, companies may operate simultaneously in various countries by communicating via telephone, emails, or social networking tools with the representatives of markets that could not have been reached in the past. Communication itself also varies in comparison with the type of interactions popular in the past century. In the area of Tofflerian prosumers, customers are not only passive receivers of goods and services but they are also active creators of merchandise and organizational culture. They do not only design products themselves but they also construct the dialogic sphere in which they participate as proper members, together with producers, users, and the nonhuman entities, such as goods and the broadly understood technological environment of economic reality. The interactional approach is also highlighted by Klammer (2007, p. 15), who states that *economics is a conversation, or better, a bunch of conversations, and economists are economists because they are in conversation with other economists*. The development of economic discourse takes place in various domains, leading to the appearance of new subdomains and methodological coexistence with other academic disciplines. The burgeoning research encompassing the representatives of different disciplines makes the language of economics a complex and multidimensional phenomenon. As far as scientific studies are concerned, the article by William Warrand Carlile (1909) was one of the first works on the language of economics. Nowadays, the role of economic communication is studied by both researchers and businesspeople in order to make this type of language for special purposes even closer to speakers representing different levels of economic knowledge. At the individual level, the popularity of the language of economics is also connected with the growing role of goods in the life of individuals and the importance of material possession for some people. Taking the organizational dimension into account, effective communication exercised both internally and externally is one of

the key determinants of company effective performance on competitive markets.

The Language of Economics: A Definition

The language of economics can be understood in at least two ways. The first method is to treat language in the broad sense, analyzing both verbal and nonverbal (e.g., pictorial or olfactory) communication. The language of economics investigated in this way can be observed from a more holistic perspective, drawing one's attention not only to the linguistic layer of economic discourse but also to other senses that shape the rhetoric of economics. Taking marketing as an example, verbal communication stimulates purchasing behaviors of customers, together with other types of experience that determine the way reactions and behaviors are shaped. For example, smells and sounds used in shops may intensify the perception of words aimed at making buyers interested in products. Thus, the approach to language is synesthetic, showing how verbal, pictorial, olfactory, and audio dimensions create the way a given phenomenon is encoded and decoded. Another perspective is to concentrate exclusively on the linguistic level of economic communication, focusing only on words and their role in this type of language for special purposes.

The next field of analysis is the dichotomy of inner and outer dimensions of economic discourse. At the internal (mainly organizational) level, the language of economics plays the following functions. First, it facilitates internal communication between specialists representing different areas of economics, offering them the common ground for specialized interactions. Secondly, organizational discourse offers the possibility of "closed and exclusive communication," relying on selected terms understood only by the members of a given organizational community. This results in a creation of a corporate code and, consequently, a strengthening of organizational identity. Analyzing the language of economics from a more external perspective, it stimulates effective communication between economists

and laymen, often coming from different linguistic and professional backgrounds. For example, the dominance of English terms in economic discourse makes it easier for specialists speaking different languages to communicate by relying on international concepts understood by different users, regardless of their mother tongues.

The Language of Economics: General Characteristics

The language of economics does not exist in a vacuum; it shapes and, at the same time, is shaped by other domains of life, such as politics, culture, geography, technology, and economics. The multidimensionality characterizing the language of economics can also be observed at the level of stakeholders. The participants in economic dialogue can be categorized by taking into account their level of knowledge on economic matters and their professional involvement in organizational matters (employees vs. nonemployees). The language of economics should also take into account users with special needs, visible in, e.g., adjusting product information for the blind. Thus, the studies on economics discourse should also encompass communication types and tools tailored to the needs and expectations of a diversified audience. Apart from the mentioned external determinants, the language of economics is a complex phenomenon within itself. One of its key characteristics is the high number of borrowings and loans in its lexicon. In many languages of economics, loanwords constitute a large part of economic dictionaries. Taking the example of English, many economic concepts come originally from French, making it the most powerful language donor, contributing about 40–60% of all terms, depending on subdomains, such as law, accounting, or general business. An example can be coupon used in finance, entrepreneur used in management, or force majeure used in business law. Another key donor language is Latin, with terms such as *moneta*, *pondus*, and *centus* representing the monetary dimension of economics. Another important foreign language that influenced the current state of English economic lexicon is Japanese. One of

the areas rich in Japanese terms is management philosophy, with such terms as *genba shugi*, *keiretsu*, and *zaibatsu*. Another field that relies on Japanese words is technical analysis, with such names describing candles as *Harami*, *Marubozu*, or *Doji*. The next linguistic donor is Greek, with such terms as kappa, rho, or phi used to describe options. Other foreign concepts in the English language of economics come from Italian (e.g., *agio* and *mezzanine*), Spanish (e.g., *cargo* and *gambit*), as well as the North Germanic languages and German (e.g., *blitzkrieg* *tender offer*) as discussed by Bielenia-Grajewska (2009a). The linguistic dimension of economic discourse is also connected with literal and figurative economic communication. *Literal economic communication* can be investigated through the way certain linguistic tools are selected and used, such as verbs, nouns, and adjectives, and the way their direct meaning influences communication. *Figurative economic communication* encompasses such notions as metaphor, metonymy, idiom, personification, and simile. Since not only verbal communication constitutes the language of economics, rituals, rites of passage, and the arrangement of furniture or office space constitute the identity of modern companies. Focusing on the verbal dimension of economic discourse, McCloskey (1995, p. 218) states that *economists are poets but they do not know about it*. When one observes the language of economics, it is noticed that there are many metaphors used to denote economic reality. They serve different functions. On the individual level, they are used by economists to construct their own idiolect that can be viewed as an important element of their identity. In addition, metaphors make economists or businessmen outstanding and remembered by others. The same can be observed in the case of economic journalists, who rely on figurative metaphors to make their content more interesting and intriguing. This feature is important in the case of covers and headlines that are to attract potential readers to the articles themselves. On the social level, they stimulate effective communication between people of different knowledge levels on economic topics. A metaphor, relying on a well-known domain, is more easily perceived and understood

than a complicated economic term. The application of metaphors to describe economic environments can be observed at different levels of economic discourse. Taking the microsphere into account, metaphors are used to denote products or strategies. The examples include such terms as *porcupine defense* or *black knight*, used to describe the world of mergers and acquisitions (Bielenia-Grajewska 2009b). The meso level encompasses organizations; in that case, metaphors are used to create the image of a company in the eyes of stakeholders. The following examples come from the food industry: organization as a teacher, organization as a network, organization as a protector, organization as a traditionalist, organization as a travel guide, and organization as a family (Bielenia-Grajewska 2014). Company linguistic identity can be studied through the prism of the 3Ps model of company linguistic identity and its metaphorical dimension. The application of three angles, such as personnel, purchasers, and products, stresses how metaphors determine corporate communication and how they shape the selection of products by attracting stakeholders to companies and their offers (Bielenia-Grajewska 2015). The examples to support the discussed phenomenon can come from different domains. One of them is the animal world; animals, having distinctive features and being widely known, constitute an important metaphorical donor. Another one is weather, being a varied phenomenon, offering both positive and negative notions. In addition, the unpredictability of weather conditions is used to denote economic phenomena that cannot be foreseen and controlled. Another domain is literature, with metaphorical names originating from tragedies (e.g., *Lady Macbeth Strategy*) or myths (*Sisyphean struggle*). The next domain is medicine, with such metaphors as *pills* or *heal* used to describe how economic conditions can be improved.

The Language of Economics: Approach Perspectives

Another way to study the growing interest in the language of economics is to investigate it through

the prism of approaches and theories determining modern reality. Different assumptions originating from social studies and the humanities can be used to research the language of economics as a scientific phenomenon since they often stress the changeability and complexity of economic discourse.

The Language of Economics as a Re-enchanted and Fluid Reality

The interest in the language of economics can be viewed from the perspective of re-enchanted reality. As Reed (2002, p. 35) discusses, *re-enchantment refers to a symbolic or discursive, rather than material or structural, reworking of the ways in which organizations discipline and control their members. It shifts the focus of attention away from material technologies and organizational structures to the cultural and linguistic forms through which members represent and communicate their organizational identities*. Thus, modern business can be studied as an example of re-enchanted entities, with language showing how people create the economic reality, exercise power, and become competitive. An example is the symbolic layer of economic communication, with metaphors, fairytales, and myths used to denote organizational culture.

Applying a broader theory, such as postmodernism, allows researchers to study the language of economics as a dynamic phenomenon that escapes easy categorization. For example, the Baumanian concept of liquid modernity offers the opportunity to investigate language without posing any rigid boundaries between language, economics, politics, and private life, but by treating language as an element of fluid reality that can merge with other entities and influence them.

The Language of Economics as a Systemic Entity

A similar perspective is connected with looking at the language of economics as an open system. In this approach, attention is drawn to the language of economics as an element of a more complex system and, at the same time, to economic communication as an entity constituted of smaller

elements. This line of investigation stresses the multilayered and interdependent character of economic discourse. The mentioned systemic perspective exemplifies different subsystems within the language of economics itself (such as the language of accounting, the language of banking, and the language of management) as well as highlighting that the language of economics is an element of a more compound phenomenon, such as the domain of languages for special purposes. At the text level, the traces of systemic approaches are visible in the notion of a paratext by Gerard Genette (1997). Paratextual materials accompany the main text in different ways. In the case of economic books, such notions as reviews, cover layout, or interviews with authors can be examples of paratexts. Another textual approach is intertextuality as discussed by Julia Kristeva (1980), stressing the role of other texts or their elements in creating subsequent works. The systemic character of economic communication is also represented by network approaches. For example, Actor-Network Theory (ANT) stresses that both living and nonliving elements shape the performance of a given entity. Taking the language of economics into consideration, not only human beings determine the way language is created and used. For example, technological tools, such as the Internet and telephone, and standard modes of information creation and distribution, including books or magazines, shape and maintain economic discourse. Another network theory is social network analysis (SNA) that offers discussion on grids, lattices, and relations between those creating and using organizational networks. The next concept that stresses the interrelation of linguistic phenomena is heteroglossia. This term, originating from the Greek words hetero (different) and glossa (language), was introduced by the Russian linguist Mikhail Bakhtin (1986) to discuss, e.g., different parts of speech in literary works. Nowadays the term heteroglossia is implied to stress the plurality and complexity of modern communication, visible in the concept called *heteroglossic linguistic identity of modern companies* (Bielenia-Grąjewska 2013). Applying this notion to the discussion on the language of economics facilitates the study on the

multivocality of economic discourse, observed at micro, meso, and macro levels. The micro perspective is dominated by hybridity at a word level, represented in terms originating from different domains of life. The meso level, on the other hand, can be viewed through the prism of textual representation, studying different types of texts used in economic discourse. The macro approach, on the other hand, is connected with different voices shaping the language of economics, represented in the way different experts and laymen communicate. The application of hybrid approaches to the studies on the language of economics can be observed in the concept of *hybrid linguistic identity* (Bielenia-Grajewska 2010); the exposure to different linguistic codes leads to the creation of a new linguistic representation that possesses unique features. It is not only the fusion of different languages or dialects but it also results in an exclusive communication style, being a novel linguistic strategy that offers new possibilities of usage and research. The hybridism of economic communication is represented at different levels, starting from the word level and finishing with the language of economics as a complex phenomenon.

The Language of Economics: Technological Perspectives

The language of economics is a dynamic phenomenon. The mentioned dynamism has different reasons, with technology being an important determinant shaping economic discourse. Technological determinism is visible not only in the way the language of economics is created but also in the way it is stored and offered. As far as the creation of economic discourse is concerned, technological advancements have resulted in a language of economics that is efficient and economical. Technology has also created new places where the language of economics can be stored. In addition, new economic terms and expressions are coined and used to denote novel technological and economic reality. The language of economics can also be studied from the perspective of written and spoken forms of interaction. As far as the written types are concerned, they involve both standard and novel modes of communication, such as

letters, emails, offers, websites, social media communication tools, and leaflets. The mentioned channels of economic communication can also be subcategorized through the perspective of synchronicity and possibilities of alternation. For example, websites can be updated very quickly, whereas new catalogues and leaflets have to be prepared, printed, and distributed. The spoken side of economic discourse encompasses business talks, negotiations, speeches, as well as telephone or online conversations. Both written and spoken channels of economic communication undergo technological changes. The growing popularity of online communication tools, such as electronic billboards, social media, and communicators, has enriched the spectrum of communication possibilities.

The Language of Economics: Discipline Perspectives

The Language of Economics: Linguistic Perspectives

In scientific literature on economic discourse, one can also come across such terms as *economese* or *econospeak* and *dialogical economics*. Although they vary as far as the field of interest is concerned, they draw attention to the linguistic side of economics. Linguistic perspectives focus, as their name suggests, on the language-related aspects of economic communication. Detailed characteristics of this approach are presented in the section devoted to the main features of economic discourse.

The Language of Economics: Educational Perspectives

The discussion on the language of economics from educational perspectives is connected with investigating how the language of economics can be taught and learnt. It focuses on the studies devoted to languages for special/specific purposes and their relation to general language. Using English as an example, the language of economics is often researched and taught as English for Business. This line of teaching provides a type of instruction that meets the needs of economists.

Apart from the general focus on the language of business as such, teachers also offer more tailored courses, such as English for Marketing, English for Accounting, and English for Management, to meet the expectations of accountants, marketers, and managers. Educational perspectives also encompass necessary books and other sources to study the language of economics. Thus, the educational dimension also includes printed and online dictionaries, coursebooks, CDs, and supplementary materials.

The Language of Economics: Economic Perspectives

The relations between language and economics can be studied by investigating how certain linguistic policies and behaviors increase or decrease economic performance. One aspect is the connection between knowledge flows and economic efficiency. For example, providing proper translation increases the chances of companies to be successful on foreign markets. Apart from translation, the notion of localization becomes more and more popular since online content does not only have to be translated but also localized, taking into account the needs and expectations of the target market. For example, George Akerlof (1970) discusses the notion of *adverse selection*. This concept denotes situations when the asymmetry of information leads to less favorable positions of products or services on the market and lower selection rates among customers, in comparison with products that benefit from efficient informational coverage. Another notion that can be studied is trust. It is also language dependent since effective communication between interlocutors determines their cooperation. The economic perspective of language can be exemplified at different levels. One of them is the supranational perspective, examining the role of a lingua franca in creating global economy. The national perspective may concern the place of a given language in creating a national economy and its role in international business. The national perspective may also concern how national linguistic policies determine the usage of national language in organizational settings. This dimension encompasses the attitude to national languages used in

international companies and the place of minority languages and dialects within national economies. An individual perspective can also be analyzed by looking at the predicted possible gains related to learning a given foreign language.

Apart from the language of economics, there are other concepts that sound similar but carry a slightly different meaning. One of them is the *economics of language* that focuses on how language determines the economic side of individual or organizational performance, showing, e.g., the relation between language and employment income, the economic dimension of second-language acquisition, language and immigration, or language and rational choice theory (e.g., Grin 1994). The topics include the following spheres of investigation: the economics of the multilingual workplace (Grin et al. 2010) or the relation between languages and economic advantage (Ginsburgh and Weber 2011).

The Language of Economics: Political, Social, and Historical Perspectives

The link between politics and economic discourse is visible in the creation of terms related to a given economic system in a country. Politics is represented in different regulations on national languages as well as the political system and its influence on the economy. Language also offers information on socioeconomic conditions. There are studies in literature stating how one's selection of semantic and syntactical choices is connected with one's upbringing, education, and profession. In the language of economics, the selection of words may mirror one's knowledge of a given economic domain. The historical perspective can be studied by looking at how loanwords entered the economic lexicon of a given language or how foreign syntax was adopted in economic communication.

The Language of Economics: Biological Perspectives

The dynamic perspective of economic communication can also be discussed from the perspective of memetics. Originating from genetics and studied by such scholars as Richard Dawkins or Susan Blackmore, memetics focuses on replication in

culture; it can be used to discuss how new economic terms “replicate” in a new economic reality. For example, new economic terms that appeared in Poland accompanying the change of the system in the 1990s first spread among economists and journalists and later became known also among laymen. A concept called *viral meme* that transmits because of its emotional connotations is also applicable to economic memes. Such economic phenomena as market crashes, mergers and acquisitions, or failures of financial institutions, together with terms denoting them, can be treated through the perspective of viral memes. Examples constitute metaphorical names describing M&A, such as *white knight* or *black knight*, with the “color of armor” showing the intentions of acquiring companies. Moreover, the success or failure of “infecting” the public with a new term depends on how the translation of a given term is accepted by the target audience. For example, some economists prefer the English term *futures* instead of long and descriptive versions in their mother tongues. The analogy to genes is also visible in the case of viral economic terms that gain popularity very quickly (Bielenia-Grajewska 2008). The memetic approach to economics can be noticed in the contribution by Měřo (2009), and the concept of *Mom* that can be defined as the piece of information that describes a company and together with other *Moms* creates economic entities.

The Language of Economics: Neuroscientific Perspectives

As has already been mentioned in the introductory part of this chapter, modern economics does not exist in a vacuum but it is shaped by other disciplines that were not directly linked with economic phenomena in the past. Neuroscience is an example of the domain that is more and more popular in different types of economic research. It has led to the creation of the discipline called neuroeconomics and other economic-related neuroscientific fields of investigation. For example, such subdisciplines as international neurobusiness, international neurostrategy, neuromarketing, neuroentrepreneurship, and neuroethics belong to international neuromanagement. *International*

neurobusiness can be perceived as the neuroscientific dimension of activities, hierarchies, and decisions related to multinational enterprises. Similarly, *international neurostrategy* can be defined as the neuroscientific level of organizational focus on reaching a competitive advantage on international markets. *Neuromarketing*, on the other hand, refers to the neuro side of marketing, represented in using neuroscientific developments and tools in advertising products and services. The more research-oriented perspective of studying the link between marketing and neuroscience is *consumer neuroscience*. *Neuroentrepreneurship*, on the other hand, helps understand the types of entrepreneurs and entrepreneurship and the way biological and cognitive factors determine leadership styles as well as individual and social entrepreneurial activities. *Neuroethics* involves moral approaches to studies in neuroeconomics (Bielenia-Grajewska 2013). Taking the linguistic aspect of neuromanagement into account, the mentioned popularity of neuroscience has led to the emergence of new terms within the economic lexicon that denote the newly created scientific field of neuroeconomics. Thus, such terms as fMRI or galvanic skin response, previously associated exclusively with neuroscience, have been incorporated into the economic discourse of the twenty-first century. Analyzing the rise of interest in the way the brain and the nervous system are influenced by economic stimuli, it can be expected that the economic lexicon will be further enriched with the appearance of even more sophisticated neuroscientific vocabulary.

The Language of Economics and Its Functions

Among its different functions, economic discourse offers successful tools for gaining economic advantage. This phenomenon can be understood in various ways. One of them is to focus on marketing and the way marketing communication shapes customers' behaviors and organizational identity. The linguistic sphere of marketing is visible in, among other things, the creation of brand names and marketing slogans. Moreover, language

is a tool fostering innovation and entrepreneurship, stimulating information flows between disciplines as well as knowledge creation and knowledge communication. In addition, language exhibits the level of knowledge in disciplines (Bielenia-Grajewska et al. 2013). The language of economics serves the communicative function on the inner and outer level of organizations, showing employees how tasks should be done and stimulating effective communication with the broadly understood stakeholders. The language of economics also mirrors the legal dimension of investigated entities. Linguistic policies can be studied by taking into account the micro level, that is, the organization as such, as well as more macro dimensions, such as national linguistic policies or the supranational ones, such as the EU regulations.

The Language of Economics and Methods of Investigation

The language of economics can be investigated by taking into account the methods and tool characteristic of linguistics. Thus, such domains as sociolinguistics, psycholinguistics, neurolinguistics, pragmatics, semantics, and syntax or corpus studies can be used in the discussion on economic communication. The selection of methods applicable to study the language of economics depends on research perspectives. Applying the micro (word) scope, cognitive linguistics may be used to observe the way words are created and understood. Moreover, corpus linguistics facilitates the studies on concordance or collocations. A more macro perspective on the language of economics is offered by discursive approaches, such as critical discourse analysis (CDA). CDA approaches language as a social practice, investigating how language creates social relations and how these relations are perceived by using a selected communicative repertoire. Moreover, critical discourse analysis does not only concentrate on the purely linguistic elements but it also studies nonlinguistic notions to show how both verbal and nonverbal tools determine the interest in the perception, comprehension, and influence of a given economic text. Since CDA is an approach that is used to investigate such

notions as social problems and issues (e.g., dominance or inequality), this method of analysis is used to discuss texts related to economic imbalance, crisis, and corporate changes. Taking into account the growing role of neuroscience in different fields of life, it can be predicted that neuroscientific tools will become even more popular for checking how people create, use, and understand words. Thus, the noninvasive techniques popular in neuroscience, such as fMRI and galvanic skin response (GSR), provide information how an economic term is understood and how the mentioned comprehension facilitates decision-making processes. The language of economics can be investigated by using both qualitative and quantitative studies. Qualitative approaches offer the focus on tendencies and decision-making processes underlying modern economic discourse, such as the use of borrowings or the influence of technology on the way words are created and used. Quantitative studies provide, among other things, statistical data on economic discourse. Researchers investigating the language of economics may rely on methods and tools widely used in the humanities and social studies, such as interviews, questionnaires, expert panels, and case studies.

Conclusion

The multidimensionality of economic communication results in different linguistic and nonlinguistic layers of the phenomenon itself as well as the plethora of approaches that can be used to investigate it. Taking into account the coexistence of economics with other domains of life, it can be predicted that new disciplines will appear that will also focus on the language of economics, at least to some extent. For example, neuroeconomics and neurolinguistics separately and together provide new and complex ways of analyzing how economists speak.

Cross-References

- [Entrepreneurship](#)
- [Financial Education](#)
- [Knowledge](#)

References

- Akerlof G (1970) The market for “lemons”: quality uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Bakhtin M (1986) *Speech genres and other late essays*. University of Texas Press, Austin
- Bielenia-Grajewska M (2008) Mem jako jednostka translacyjna w przekładzie terminów ekonomicznych. In: Pstyga A (ed) *Słowo z perspektywy językoznawcy i tłumacza*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, pp 225–230
- Bielenia-Grajewska M (2009a) Linguistic borrowing in the English language of economics. *Lexis – E-J Engl Lexicol* 3:107–135
- Bielenia-Grajewska M (2009b) The role of metaphors in the language of investment banking. *Iber* 17:139–156
- Bielenia-Grajewska M (2010) The linguistic dimension of expatriatism – hybrid environment, hybrid linguistic identity. *Eur J Cross-Cult Competence Manag* 1(2/3):212–231
- Bielenia-Grajewska M (2013) International neuro-management: deconstructing international management education with neuroscience. In: Tsang D, Kazeroony HH, Ellis G (eds) *The Routledge companion to international management education*. Routledge, Abingdon, pp 358–373
- Bielenia-Grajewska M (2014) CSR online communication: the metaphorical dimension of CSR discourse in the food industry. In: Tench R, Jones B, Sun W (eds) *Communicating corporate social responsibility: lessons from theory and practice*. Emerald, Bingley, pp 311–333
- Bielenia-Grajewska M (2015) Metaphors and risk cognition in the discourse on foodborne diseases. In: Mercantini JM, Faucher C (eds) *Risk cognition*. Springer, Berlin/Heidelberg, pp 89–105
- Bielenia-Grajewska M (2015) The role of figurative language in knowledge management. In: Khosrow-Pour M (ed) *Encyclopedia of information science and technology*. IGI Publishing, Hershey, pp 4728–4735
- Bielenia-Grajewska M, Carayannis E, Campbell D (2013) Linguistic dimension of creativity, invention, innovation, and entrepreneurship. In: Carayannis E (ed) *Encyclopedia of creativity, invention, innovation and entrepreneurship*. Springer, New York, pp 1206–1215
- Genette G (1997) *Paratexts: thresholds of interpretation*. Cambridge University Press, New York
- Ginsburgh V, Weber S (2011) *How many languages do we need? The economics of linguistic diversity*. Princeton University Press, Princeton
- Grin F (1994) The economics of language: match or mismatch? *Int Polit Sci Rev* 15(1):25–42
- Grin F, Sfreddo C, Vaillancourt F (2010) *The economics of the multilingual workplace*. Routledge, New York
- Klamer A (2007) *Speaking of economics: how to get in the conversation*. Routledge, Abingdon
- Kristeva J (1980) *Desire in language: a semiotic approach to literature and art*. Columbia University Press, New York
- McCloskey D (1995) Metaphors economists live by. *Soc Res* 62(2):215–237
- Mérő L (2009) *Die Biologie des Geldes: Darwin und der Ursprung der Ökonomie*. Rowohlt Taschenbuch Verlag, Reinbek
- Reed M (2002) From the ‘cage’ to the ‘gaze’? The dynamics of organizational control in late modernity. In: Morgan G, Engwall L (eds) *Regulation and organisations: international perspectives*. Routledge, London, pp 17–49
- Warrand Carlile W (1909) *The language of economics*. *J Polit Econ* 17(7):434–447

Further Reading

- Bielenia-Grajewska M (2011) Teaching business communication skills in a corporate environment. In: Komorowska H (ed) *Issues in promoting multilingualism. Foundation for the Development of the Education System*, Warsaw, pp 39–57
- Klamer A, McCloskey D, Solow RM (1988) *The consequences of economic rhetoric*. Cambridge University Press, Cambridge
- McCloskey D (1986) *The rhetoric of economics*. The University of Wisconsin Press, Madison

Lapse of the Contract Basis

► Impracticability

Law and Economics

Pablo Salvador Coderch¹ and Antoni Terra Ibáñez²

¹Universitat Pompeu Fabra, Barcelona, Spain

²Stanford Law School, Stanford, CA, USA

Synonyms

Economic Analysis of Law

Translated by Leah Daniels Simon, Bachelor’s Degree in Law (2014), Universitat Pompeu Fabra (Barcelona).

Definition

Law and Economics is economics applied to the analysis of statutory law systems or subsystems or to legal policy proposals.

Historical Remarks

To talk about Law and Economics or, equally, of the Economic Analysis of Law is to refer to the application of economic science tools to national and international systems of statutory law, as well as to legal policy proposals.

Positive analysis is to be distinguished from normative analysis. Within the framework of positive analysis, economic science is employed within a system of statutory law to explain and predict the consequences of the application of one or more sets of judicial system rules in questions concerning human conduct. It consists, therefore, in elucidating the manner in which people react to the enactment, enforcement, abolition, or reformation of such rules of law. In this sense, the Economic Analysis of Law is a system of theories regarding human behavior. The key question regarding positive analysis is to understand the law as a system of incentives.

Within the framework of normative analysis, on the other hand, economics is applied to the evaluation of rules of law in accordance with the degree to which they contribute to bettering the economic efficiency of the analyzed conduct: given two alternative regulations, the economic normative analysis would give preference to the one whose application, according to predictions based on its own analysis, would generate the most efficient results.

The application of economic tools to the analysis of law is as old as economics itself (Smith 1776). The first systematic applications were focused, understandably, on criminal law (Bentham 1789) – one of the earliest fields in which national legal systems received scientific analysis being that of crime investigation (see the entry ► [forensic science](#)). However, the emergence of the positive-normative movement called Law and Economics took place much later, during

the second half of the twentieth century in the United States of America, driven fundamentally by the influence of academics from the Chicago School. Thus, economists Ronald Coase (1960) and Gary Becker (1968) or jurist Richard Posner (1973) started to massively apply neoclassical microeconomics to law in both analytical and normative terms. In a synthesis reminiscent to European academics of Georg Wilhelm Friedrich Hegel's philosophy, the first Economic Analysis of Law came close to claiming that reality is rational and that rationality must be turned into reality. In Posner's well-known formulation, North American common law is essentially efficient; thus, in the measure in which it has not yet achieved that status, it will end up being so.

Nonetheless, liberally oriented lawyers and economists, such as Guido Calabresi (1961) in the United States or Pietro Trimarchi (1961) and Hans-Bernd Schäfer and Claus Ott (1986) in Europe, were also pillars of the birth of the Law and Economics model.

In the 1970s and 1980s, the Law and Economics movement, as it was defined, was advocated by the best US law schools; however, although the program was also accepted in continental European culture, its penetration in European law faculties was less intense than that which took place on the other side of the Atlantic. See, for instance, Boudewijn Bouckaert and Gerrit De Geest (2000), Peter Cane and Herbert Kritzer (2010), or Régis Lanneau (2014).

The degree of influence of the Economic Analysis of Law program has stabilized since the last decade of the twentieth century; the evolution of Law and Economics has continued to be characterized by its increasing specialization in such a way that currently each legal area or subject possesses top-level economic applications with very rich empirical and econometrical contents. Nowadays, it is inconceivable that a top-level economic analyst specializing in Family Law would also be a specialist, in similar academic conditions, in US Law and the Market of Telecommunications, in the Law of Publicly Traded Companies, in Biopharmaceutical Law, in International Taxation, in Product Liability of Health Agencies, and in International Sales Contracts.

For each and every one of these specializations, readers should consult the corresponding entries of this encyclopedia.

Currently, the level of specialization of economics applied to law and the intensity of its econometric sophistication (see the papers presented at the 24th Annual Meeting of the American Law and Economics Association [ALEA], University of Chicago, 2014, <http://www.amlecon.org/2014-Program.final.revised.pdf>) allow us to state that Law and Economics is a victim of its own success. As it is true of the relationships between Criminal Law, Criminology, and Legal Medicine, or Patent Law and Engineering, along with generalists that hold a good background in economics and a primarily legal education, the different fields within the Economic Analysis of Law would be more ideally developed by multidisciplinary teams in which economists and lawyers collaborate. In the end, as we shall see in this same entry, the specific relationship between economics and law represents a specific instance between the different sciences and the law itself (Lawless et al. 2010).

In any case, and in the current academic environment, top-level North American or European law schools take into account the Economic Analysis of Law and legal institutions although in none of them is the economic and normative Analysis of Law the dominant paradigm.

Legal Assumptions of Economic Analysis

In general terms, economics studies the efficient assignment of scarce resources in specific markets and in the economic system as a whole. Furthermore, political and public sector economics analyze, in a manner independent of market functioning, the feasibility of political and legal proposals regarding wealth redistribution and the establishment of infrastructures or direct public service provisions by governments (Stiglitz 2000; Barr 2012).

In an advanced economy, the objective of economic analysis, i.e., the market, is exogenous to the analytical instruments themselves as well as to their applications, as every market requires some implicit economic assumptions in order to exist,

consisting, at least, of the following two: a system of property rights which formally recognizes who is the owner and what are his or her faculties of use, disposition, and exclusion, and a system of law of contracts that permits owners to exchange goods and services in order to achieve more efficient allocations. In addition to the above, and given a minimally developed economic system, two additional legal systems will be required: Company Law, in particular for limited liability companies and, within these, for publicly traded ones, and, finally, a capital market subjected to one or more public regulations that coordinate, evaluate, and, when applicable, sanction the conducts of those who operate in it (Hadfield and Weingast 2013). These issues do not constitute specific study objectives in any standard economic or microeconomic textbook, but perhaps for this same reason, it is convenient to align economic analysis with the legal assumptions of its own traditional aim (Mankiw 2014; Krugman et al. 2013; Varian 2014; Mas-Colell et al. 1995).

The previous four legal subsystems shape the national or, when appropriate, international systems of private law and are the objective of economic analysis. But, in turn, they also presuppose a public system of legal, governmental, and judicial remedies given the violation of property rights, the breach of contracts, the dysfunction of limited liability companies, or the defective behavior of capital markets. Although it is a fact that the four subsystems have always been partially controlled by social norms, it is also true that, in general terms, private law subsystems depend on public law systems which, operate through coercively adopted central decisions independent of the economic markets.

Economic Assumptions of Legal Analysis

In turn, and in a symmetrical fashion, historical experience substantiates the thesis that a legal system has never been able to survive independently of capital markets or, in other words, the necessity to acknowledge the systematic assignment of economic resources in accordance with the economic analysis itself (von Mises 1920).

From a historical economic point of view, the final destiny of national legal systems depends on their economic successes. The main analytical question continues to reside in the causes of wealth and poverty of nations (Acemoglu and Robinson 2012; Cooter and Schäfer 2012).

Traditionally, as summarized by Polinsky and Shavell (2008), the Economic Analysis of Law has been the target of criticism by those who have pointed out its alleged shortcomings, especially regarding the following three considerations: First, it is at times insufficient and, at other times, inexact to presuppose, as the Economic Analysis of Law would have us do, that individuals and organizations are rational maximizers of well-being. Second, it implies limiting oneself to a positive and normative legal system analysis which takes exclusively into account economic efficiency considerations without regard to the legal consequences resulting from the initial assignment of property rights over production and consumption goods or to the distribution of income and wealth. Third, and lastly, such an analysis would be incomplete if, together with the objective of economic efficiency, it did not take into account criteria of fairness or of basic notions concerning corrective or distributive justice.

The limitations of the neoclassical analysis derived from the assumption of rational human behavior have been corrected, at least in part, by the incorporation of behavioral economics (Akerlof and Kranton 2010; Kahneman 2011; Sunstein 2000). It consists, basically, of enriching the legal analysis with a multidisciplinary treatment of the rules and principles that define it or, in other words, an analysis which is not only economic but also psychological, sociological, and political, as we shall soon see.

As for the criticisms derived from the justice theory or from the lack of attention to particular fairness aspects, it is worth pointing out that the former raise basic economic analysis and jurisprudence issues – jurisprudence being understood as a philosophy of ideas concerning the law without regard to an empirical analysis – while the second criticism often refers to the lack of specific considerations concerning the empirical

consequences that any given regulation could exert on the real distribution of income and wealth in a given society at a specific moment.

The Economic Analysis of Law as an Instance of the Application of the Current State of Scientific and Technological Knowledge to Legal Systems

Economics recognizably forms a branch of scientific activities belonging specifically to that of the social sciences. Within this field, it is part of the formal, natural, and life sciences. Given, then, that law can be analyzed from the perspective of any of the formal, natural, life, or social sciences, as well as from the perspective of technology or the applied sciences to which it is related and not only from an economic standpoint, it is undeniable that to explain and predict the human conduct subject to legal rules, the Economic Analysis of Law is at the same time both necessary and insufficient. For this reason, it can be affirmed that presently the reduction of the analysis of human conduct to economics is no longer endorsed, in the same way that now no one holds the reduction of the analysis of nature to physics (see Schurz 2013).

Therefore, a basic question regarding the relationship between economics and a particular national or supranational legal system is the relevance of a specific social science, economics, in the legal system concerned. Regarding this topic, it is convenient to point out that different national legal systems take into consideration the current state of scientific and technological knowledge in distinct ways and with varying intensity.

To quote an example, the United States' federal law system has, since 1993, established a normative doctrine regarding the consideration an allegedly expert witnesses' testimony should merit to a judge or federal court. The expert is presented before the court itself as a scientist or a technologist, whose aim is to illustrate facts and natural or social issues that might be important to the court for a resolution in which the state of scientific or technological knowledge could be relevant. In

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993), a product liability lawsuit, the issue at stake was whether a particular drug had caused or not the damages suffered by the plaintiffs. Given that this matter had a fundamentally scientific basis, the debate was later centered on another issue regarding the reliability, scientific in this case, of the expert witnesses presented by the parties. In this regard, the US Supreme Court established that judges, when faced with scientifically relevant issues and, thus, not subject to the interpretation of the law itself – in other words, out of the realm of their own expertise – should carry out the function of gatekeepers and distinguish between good and junk science. That is, they should exclude a priori witnesses which are purely partial and elaborated with disregard to accepted scientific methods. To that effect, the judge should carry out a preliminary evaluation to determine if the reasoning or the underlying methodologies emanating from the testimony were valid and supported by facts. Firstly, the hypothesis, formulated and presented, should be empirically contrastable – falsifiable, as determined by a then Popperian court. Secondly, the theories or techniques presented to the court should have been subjected to peer review and should preferably have already been published and hence should be accessible to the scientific community. Additionally, the known or potential margin of error should be taken into consideration. And finally, the degree of acceptance of the formulated theories or hypothesis should also be taken into account in distinguishing between good and junk science. The standard established by *Daubert* was later incorporated by the Rule 702 of the Federal Rules of Evidence and has been followed by subsequent cases.

However, not all national legal systems include rules that establish control filters or demarcation operational standards concerning the testimonies of expert witnesses and their scientific or technological reliability. Nor do all national legal systems foresee that before the enactment of a certain statute or regulation, a scientific or technological – and, when applicable, also economical – assessment should necessarily take

place regarding the consequences that the rule would have if it were to be enacted and enforced. Ultimately, the relevance of the Economic Analysis of Law depends on the law itself. To sum up, not everything that is efficient is necessarily fair, but if everything is inefficient, nothing can be fair.

Cross-References

- ▶ [Behavioral Law and Economics](#)
- ▶ [Coase, Ronald](#)
- ▶ [Coase Theorem](#)
- ▶ [Economic Growth](#)
- ▶ [Efficiency](#)
- ▶ [Empirical Analysis](#)
- ▶ [Forensic Science](#)
- ▶ [Incomplete Contracts](#)
- ▶ [Legal Disputes](#)
- ▶ [Market Definition](#)
- ▶ [Rationality](#)
- ▶ [Smith, Adam](#)

References

- Acemoglu D, Robinson J (2012) Why nations fail: the origins of power, prosperity, and poverty. Crown Business, New York
- Akerlof G, Kranton R (2010) Identity economics: how our identities shape our work, wages, and well-being. Princeton University Press, Princeton
- Barr N (2012) Economics of the welfare state, 5th edn. Oxford University Press, Oxford
- Becker G (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Bentham J (1789) An introduction to the principles of morals and legislation. Payne, London
- Bouckaert B, De Geest G (eds) (2000) Encyclopedia of law and economics. Edward Elgar, Cheltenham
- Calabresi G (1961) Some thoughts on risk distribution and the law of torts. *Yale Law J* 70(4):499–553
- Cane P, Kritzer H (eds) (2010) The oxford handbook of empirical legal research. Oxford University Press, Oxford
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cooter R, Schäfer H-B (2012) Solomon's knot: how law can end the poverty of nations. Princeton University Press, Princeton
- Hadfield G, Weingast B (2013) Microfoundations of the rule of law. Stanford law and economics olin working paper no 453

- Kahneman D (2011) *Thinking, fast and slow*. Farrar, Straus and Giroux, New York
- Krugman P, Wells R, Graddy K (2013) *Essentials of economics*, 3rd edn. Worth Publishers, New York
- Lanneau R (2014) To what extent is the opposition between civil law and common law relevant for law and economics? In: Mathis K (ed) *Law and economics in Europe. Foundations and applications*, vol 1, *Economic analysis of law in European legal scholarship*. Springer, Dordrecht, pp 23–46
- Lawless R, Robbennolt J, Ullen T (2010) *Empirical methods in law*. Wolters Kluwer, New York
- Mankiw G (2014) *Principles of economics*, 7th edn. Cengage Learning, Boston
- Mas-Colell A, Whinston M, Green J (1995) *Microeconomic theory*. Oxford University Press, Oxford
- Polinsky M, Shavell S (2008) *Economic analysis of law*. In: Durlauf S, Blume L (eds) *The new palgrave dictionary of economics*, 2nd edn. Palgrave Macmillan, Basingstoke
- Posner R (1973) *Economic analysis of law*, 1st edn. Little, Brown and Company, Boston/Toronto. Last edition: (2011) 8th edn. Aspen, New York
- Schurz G (2013) *Philosophy of science: a unified approach*. Routledge, London
- Smith A (1776) *An inquiry into the nature and causes of the wealth of nations*. Cannan, London
- Stiglitz J (2000) *Economics of the public sector*, 3rd edn. W. W Norton, New York
- Sunstein C (ed) (2000) *Behavioral law and economics*. Cambridge University Press, Cambridge
- Trimarchi P (1961) *Rischio e responsabilità oggettiva* [Risk and strict liability]. Giuffrè, Milano
- Varian H (2014) *Intermediate microeconomics: a modern approach*, 9th edn. W. W Norton, New York
- von Mises L (1920) *Die Wirtschaftsrechnung im sozialistischen Gemeinwesen*. Archiv für Sozialwissenschaften, 47. *Collectivist economic planning* (trans: Hayek FA, 1935). George Routledge & Sons, London
- Parisi F, Fon V (2009) *The economics of lawmaking*. Oxford University Press, Oxford
- Polinsky M (2011) *An introduction to law and economics*, 4th edn. Aspen, New York
- Posner R (2014) *Economic analysis of law*, 9th edn. Aspen, New York
- Schäfer H-B, Ott C (1986) *Lehrbuch der ökonomischen Analyse des Zivilrechts*. Springer, Berlin/New York. English edition: Schäfer H-B, Ott C (2005) *The economic analysis of civil law*. Edward Elgar, Cheltenham
- Shavell S (2004) *Foundations of economic analysis of law*. Harvard University Press, Cambridge, MA
- Smelser N, Baltes P (eds) (2001) *The international encyclopedia of the social & behavioral sciences*. Elsevier, Amsterdam

Law and Economics, History of

Martin Gelter¹ and Kristoffel Grechenig²

¹Fordham University School of Law, New York, NY, USA

²Max Planck Institute for Research on Collective Goods, Bonn, Germany

Abstract

The roots of law & economics lie in late 19th century Continental Europe. However, this early movement did not persist and was essentially cut short in the 1930s. After World War II, modern law & economics was (re-)invented in the United States and subsequently grew into a major field of research at U.S. law schools. In Continental Europe, law & economics was re-imported as a discipline within economics, driven by economists interested in legal issues rather than by legal scholars. Hence, the European discourse was more strongly influenced by formal analysis, using mathematical models. Today, research in the U.S., Europe, and in other countries around the world, including Latin America and Asia, uses formal, empirical, and intuitive methods. New subfields, such as behavioral law & economics and experimental law & economics, have grown in the U.S. and in Europe during the past two decades.

Further Reading

- Calabresi G (1970) *The costs of accidents: a legal and economic analysis*. Yale University Press, New Haven
- Cooter R, Ulen T (2011) *Law and economics*, 6th edn. Addison Wesley Longman, New York
- Harrington J, Vernon J, Viscusi W (2005) *Economics of regulation and antitrust*, 4th edn. MIT Press, Cambridge, MA
- Jolls C, Sunstein C, Thaler R (1998) A behavioral approach to law and economics. *Stanford Law Rev* 50:1471–1550
- Kaplow L (1986) An economic analysis of legal transitions. *Harv Law Rev* 99:509–617
- Kornhauser L (2011) *The economic analysis of law*. In: Zalta E (Ed) (2014) *The stanford encyclopedia of philosophy*. <http://plato.stanford.edu/archives/spr2014/entries/legal-econanalysis>
- Miceli T (2008) *The economic approach to law*, 2nd edn. Stanford University Press, Stanford

Definition

A survey of the development of law and economics from the late nineteenth century until today.

Historical Antecedents

Precursors to modern law and economics can be identified as early as the nineteenth century, particularly in German-speaking Europe (Gelter and Grechenig 2007; Grechenig and Gelter 2008 referring to work by Kleinwächter 1883, Mataja 1889, Menger 1890, and Steinitzer 1908). Perhaps the defining piece for this period was the monograph on tort law by Victor Mataja (1888) titled “*Das Recht des Schadensersatzes vom Standpunkte der Nationalökonomie*” (*The law of civil liability from the point of view of political economy*), in which Mataja anticipated central ideas of the American law and economics movement, developed almost a century later. Prefiguring modern law and economics, Mataja focused on the incentive effects of tort law. While the book did not go unnoticed among legal scholars and influenced policy debates, including the drafting of the German Civil Code (Mataja 1889), it had no lasting influence on legal analysis (Englard 1990; Winkler 2004). While the University of Vienna, where Mataja held a position, integrated law and economics into one faculty, and economists in academia were typically trained in law, no law and economics movement developed (Grechenig and Gelter 2008; Litschka and Grechenig 2010).

Generally, one explanation for why the early law and economics movement left no lasting impression is the increasing specialization of the social sciences (Pearson 1997, pp. 43, 131). In the German-speaking countries in particular, legal scholarship remained strongly under the influence of a tradition that had grown out of the nineteenth-century “Historical School.” While conceptual jurisprudence gave way to the more functional jurisprudence of interests, legal scholarship continued to be seen as a hermeneutic discipline focused on a coherent interpretation of the law based on an internal consistency of the system in

terms of language and value judgments (e.g., Grimm 1982, p. 489; Wieacker 1967, p. 443). Policy arguments remained outside of the purview of legal scholarship. Moreover, nascent alternative views that may have led to more openness toward interdisciplinary work, such as the sociological jurisprudence pioneered by Eugen Ehrlich and Hermann Kantorowicz’s “Free Law School,” petered out in the 1930s and were finally cut short by the Nazi regime and World War II. Postwar jurists had no interest in portraying the law as an objective system in order to maintain the legitimacy of the legal profession (Grechenig and Gelter 2008; see also Curran 2001).

By contrast, when the modern economic analysis of law developed in the USA in the second half of the twentieth century, legal theory was far more conducive to integrating interdisciplinary and specifically economic approaches. The Langdellian orthodoxy of the late nineteenth century and the conceptual jurisprudence of the Lochner period up to the 1930s were thoroughly discredited by the legal realist movement. With “the great dissenter” Oliver Wendell Holmes on the Supreme Court as their role model, the legal realists criticized the formalism of the majoritarian jurisprudence, arguing that the law itself was to a large extent indeterminate. As for the rejection of its analogues in Germany, there is also a strong political component to the historical development in the USA: The Lochnerian judges on the Supreme Court defended the previous legal order against interventionist “New Deal” policies of the Roosevelt administration which they declared were unconstitutional. Only when Roosevelt threatened to “pack the court” with more compliant justices this jurisprudence changed, and the realists, who were generally New Deal progressives, won. Even if there was no agreement on the extent of indeterminacy, the insight that judges enjoyed great discretion in interpreting and shaping the law with policy took a strong foothold. The void left by the abandonment of formalism was filled with innovative jurisprudential movements in the second half of the twentieth century, including the legal process school, critical legal studies, and not least law and economics. After 1980, with the older generation of scholars having

left the scene, legal scholars could finally say that they were “all realists” now (Singer 1988, p. 467; Reimann 2014, p. 15). Another important issue was the prevalent role of utilitarian philosophy, on which welfare economics is based, in the US legal discourse compared to other countries (Grechenig and Gelter 2008, pp. 319–325). Finally, the maturation of economics as a discipline meant that it was better equipped for addressing legal issues than it was in the late nineteenth century (Litschka and Grechenig 2010).

The Development of Modern Law and Economics

The contemporary law and economics school is typically traced back to around 1960, specifically to the work of Ronald Coase and Guido Calabresi (e.g., Schanze 1993, pp. 2–3). However, from an institutional perspective, the basis was laid at the University of Chicago in the 1940s and 1950s, when economists first taught at the law school. Most prominently, Aaron Director began to teach at Chicago in 1946 and initiated interdisciplinary discussions both inside and outside of the classroom, most significantly in antitrust law (Duxbury 1995, pp. 342–345). He was the first editor of the *Journal of Law and Economics*.

The development of a “law and economics movement” can probably be credited to the publication of Ronald Coase’s “Problem of Social Cost” in that particular journal in 1960 (Coase 1960). The article’s core insight about the reciprocity of the relationship between the tortfeasor and the victim, and hence the significance of transaction cost, helped the economic method to expand into fields where the application of economic principles did not seem immediately obvious, such as contract or tort law. In this intellectual climate, other economists developed economic theories pertinent immediately to legal questions. Gary Becker of the Chicago economics department can be credited for applying economic principles to crime (Becker 1968), racial discrimination (Becker 1957), and family life (Becker 1981).

The prominence of the economic analysis of law in the USA today, however, probably must be

attributed to its adoption by *legal* scholars, for whom the law is – other than for most economists – the primary field of research. Among these, Guido Calabresi, Henry Manne, and Richard Posner stand out as some of the pioneering law and economics scholars.

Yale professor (later federal judge) Calabresi, in 1960, apparently independently from Coase, began a research program that led him to publish a series of articles on tort law, in which he explained its structure on the basis of simple economic principles (Calabresi 1961, 1965, 1968, 1970, 1975a, b; Calabresi and Melamad 1972; Calabresi and Hirschhoff 1972).

Henry Manne, who worked in corporate and securities law, critiqued the wisdom prevailing in these areas from the 1960s onwards (Manne 1962, 1965, 1967), attracting particular attention for his view that insider trading should be legal (Manne 1966a, b). He succeeded in transforming the George Mason University’s fledgling law school into a law and economics powerhouse (see Manne 2005). He established intensive courses on microeconomics for judges and for law professors. The fact that about 40% of federal judges had taken such a course by 1990 helped the acceptance of law and economics in the courts (Butler 1999).

Richard Posner, as a young professor at the University of Chicago Law School, not only established the *Journal of Legal Studies* in 1972 but became a trailblazer within legal academia by publishing the first edition of his monograph *Economic Analysis of Law* (Posner 1973). This standard text was the first to subject almost the entire legal system in its full breadth to a systematic analysis from an economic perspective. Posner’s publication output, both before and after joining the federal bench, is unparalleled, as he continued to influence the development of both the law and legal scholarship with his articles, books, and opinions. One of the most noted ideas was the theory, originally proposed in his textbook, that efficiency (largely defined as wealth maximization) could explain the structure of the common law across the legal system. Given that an inefficient precedent was likely to be questioned and subsequently overruled, in this view the common law tends to develop efficient solutions in the long run (Posner

1979, 1980). Obviously, the thesis has remained controversial, both as a descriptive account of the common law and as to whether wealth maximization should, normatively, be the objective of policy analysis in law (see Parisi (2005, pp. 44–48) for a summary of the debate).

Continental Europe has been described as lagging behind the USA by at least 15 years in terms of the development of law and economics (Mattei and Pardolesi 1991). While law and economics has been growing steadily in Continental Europe since the publication of Schäfer and Ott's textbook in 1986 (Schäfer and Ott 1986), important differences remain in both the methodological mainstream approach and the quantity of law and economics scholarship published in domestic law reviews. Scholars have attempted to explain the divergence by institutional factors at university level, such as publication incentives, hiring policies, and the legal curriculum (Gazal-Ayal 2007; Garoupa and Ulen 2008; Parisi 2009), on the one hand, and the institutional environment on a state level, including the separation of the legislature and the judiciary and legal positivism (Kirchner 1991; Weigel 1991; Dau-Schmidt and Brun 2006), on the other. Others have extended these approaches and focused on the legal discourse that was connected to the institutional setup. Since the Continental European concept of a separation of powers implies that a judge may only "interpret" the law, policy arguments such as those provided by law and economics were outside the scope of the legal discipline. To the extent that the legal discourse allowed for policy arguments, European legal doctrine was more strongly based on deontological philosophy than scholarship at US law schools (Grechenig and Gelter 2008).

Maturation into an Established Discipline

In the USA, economic analysis became one of the main methods of legal scholarship – both descriptive and normative – in legal academia. Other than in the early twentieth century, and in sharp contrast to Continental Europe, law is not recognized

as an autonomous discipline but a field to be studied from various social science perspectives (Posner 1987). Leading law schools often hired economists to teach and research in the area of economic analysis of law, and in the course of the 1980s and 1990s, the number of faculty with an interdisciplinary background, e.g., with a J.D. and a Ph.D. in Economics, increased (e.g., Ellickson 1989). While some legal academics publish in economics journals or in specialized law and economics journals where formal modeling or econometric analysis is typically required, a lot of economic analysis of law takes the intuitive, non-formal/non-mathematical form that is acceptable in law reviews, and one does not necessarily need to be a trained economist to be able to follow the law and economics approach.

In contrast, European law and economics scholarship was more formal (both theoretical and empirical) and was primarily driven more by economists – a characterization that remains true today (Depoorter and Demot 2011), even though non-formal law and economics scholarship at law faculties continues to grow. Several universities (primarily in the Netherlands, Germany, and Italy) have established research centers for law and economics; there are also a number of public research institutes with a strong focus on law and economics, such as the Max Planck Institute for Research on Collective Goods in Bonn, which emphasizes experimental research, and private initiatives, e.g., the Center for European Law and Economics (CELEC). Several professorships for law and economics have been established, for example, at the Universities of Amsterdam, Bonn, Frankfurt, Hamburg, Lausanne, and St. Gallen as well as at the EBS. International programs for the study of law and economics (EMLE, EDLE), both at undergraduate and at graduate level, allow students to specialize at the intersection of the two fields. An increasing number of national programs complement the options students have today. From a historical perspective, the European Association of Law and Economics (EALE), founded in 1984, seven years before the formation of the American Association (ALEA), has played an important role in the emergence of law and economics. A significant number of national and

international associations for law and economics have been established since, including in Latin America (Latin American and Iberian Law and Economics Association – ALACDE), Asia (Asian Law and Economics Association – AsLEA), Israel (Israeli Association for Law and Economics – ILEA), and other non-European countries, as well as in several European countries. These associations typically hold annual conferences, facilitating the cooperation between law and economics scholars. Today's conferences and meetings are typically more formal in terms of research methods than they used to be, although some meetings, for example, in Latin America, focus on non-formal methodology (a verbal, law-review style). An increasing number of law and economics journals, textbooks, and treatises have become available as publication outlets. Since the beginning of law and economics scholarship, research has risen to a countless number of associations, conferences, articles, etc. As a consequence, some scholars have claimed that the divergence between European and American law and economics has become much smaller than often perceived (Depoorter and Demot 2011). However, it is certainly fair to say that the influence of law and economics on the mainstream legal discourse is still much larger in the USA than in Europe. Scholars continuously argue that courts outside the USA rarely take economic consequentialist arguments and empirical evidence into account (Grechenig and Gelter 2008; Petersen 2010).

New Developments in Law and Economics

Current law and economics incorporates much of the critique that has been brought forth throughout the past decades, for example, regarding the notion of efficiency. While the traditional rational-choice approach still plays an important role, there is a growing literature on behavioral and experimental law and economics (Jolls et al. 1998; Sunstein 2000; Gigerenzer and Engel 2006; Engel 2010, 2013; Towfigh and Petersen 2014). One of the most notable movements

includes empirical legal studies. With the formation of a Society for Empirical Legal Studies (SELS), the launch of the *Journal of Empirical Legal Studies* (JELS), and an annual Conference on Empirical Legal Studies (CELS) (all between 2004 and 2006), this field became an important discipline closely connected to law and economics. Behavioral economics and empirical legal studies have helped the economic analysis of law broaden its scope by including studies with field data and experimental data, for example, experiments with judges (e.g., Guthrie et al. 2001) and laboratory experiments that demonstrate how humans behave under legally relevant circumstances (McAdams 2000; Arlen and Talley 2008; Engel 2010). Recently, a professorship for experimental law and economics, held by Christoph Engel, was established at Erasmus University in Rotterdam. The coming decades will show the way to a closer connection between the two disciplines and an enhanced use of economics in legal research, making use of whole spectrum of economic methods and possibly extending to new fields such as law and neuroeconomics.

Cross-References

- ▶ [Austrian School of Economics](#)
- ▶ [Becker, Gary S.](#)
- ▶ [Coase, Ronald](#)
- ▶ [Consequentialism](#)
- ▶ [Empirical Analysis](#)
- ▶ [Experimental Law and Economics](#)
- ▶ [Posner, Richard](#)
- ▶ [Rationality](#)

Acknowledgments We thank Emanuel Towfigh and Michael Kurschilgen for helpful comments, Brian Cooper for proofreading, and Jessica Beyer for helpful research assistance.

References

- Arlen J, Talley E (2008) *Experimental law and economics*. Elgar, Cheltenham
- Becker GS (1957) *The economics of discrimination*. University of Chicago Press, Chicago

- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Becker GS (1981) *A treatise on the family*. Harvard University Press, Cambridge, MA
- Butler HN (1999) The Manne programs in economics for federal judges. *Case Western Reserve Law Rev* 50(2):351–420
- Calabresi G (1961) Some thoughts on risk distribution and the Law of torts. *Yale Law J* 70(4):499–553
- Calabresi G (1965) The decision for accidents: an approach to non-fault allocation of costs. *Harv Law Rev* 78(4):713–745
- Calabresi G (1968) Transaction costs, resource allocation and liability rules: a comment. *J Law Econ* 11(1):67–73
- Calabresi G (1970) *The cost of accidents*. Yale University Press, New Haven
- Calabresi G (1975a) Optimal deterrence and accidents. *Yale Law J* 84(4):656–671
- Calabresi G (1975b) Concerning cause and the law of torts: an essay for Harry Kalven, Jr. *Univ Chicago Law Rev* 43(1):69–108
- Calabresi G, Hirschoff JT (1972) Toward a test for strict liability in torts. *Yale Law J* 81(6):1055–1085
- Calabresi G, Melamad AD (1972) Property rules, liability rules and inalienability: one view of the cathedral. *Harv Law Rev* 85(6):1089–1128
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Curran VG (2001) Fear of formalism: indications from the fascist period in France and Germany of judicial methodology's impact on substantive law. *Cornell Int Law J* 35(1):101–188
- Dau-Schmidt KG, Brun CL (2006) Lost in translation: the economic analysis of law in the United States and Europe. *Columbia J Transnat Law* 44(2):602–621
- Depoorter B, Demot J (2011) The cross-atlantic law and economics divide: a dissent. *Univ Illinois Law Rev* 2011:1593–1606
- Duxbury N (1995) *Patterns of American jurisprudence*. Oxford University Press, Oxford
- Ellickson RC (1989) Bringing culture and human frailty to rational actors: a critique of classical law and economics. *Chicago-Kent Law Rev* 65(1):23–56
- Engel C (2010) The multiple uses of experimental evidence in legal scholarship. *J Inst Theor Econ* 166:199–202
- Engel C (2013) *Legal experiments: mission impossible?* Eleven International Publishing, The Haag
- England I (1990) Victor Mataja's liability for damages from an economic viewpoint: a centennial to an ignored economic analysis of tort. *Int Rev Law Econ* 10(2):173–191
- Garoupa N, Ulen TS (2008) The market for legal innovation: law and economics in Europe and the United States. *Alabama Law Rev* 59(5):1555–1633
- Gazal-Ayal O (2007) Economic analysis of law and economics. *Capital Univ Law Rev* 35(3):787–809
- Gelter M, Grechenig K (2007) Juristischer Diskurs und Rechtsökonomie. *Journal für Rechtspolitik* 15(1):30–41
- Gigerenzer G, Engel C (2006) *Heuristics and the law*. MIT Press, Cambridge, MA
- Grechenig K, Gelter M (2008) The transatlantic divergence in legal thought: American law and economics vs. German doctrinalism. *Hastings Int Compar Law Rev* 31(1):295–360
- Grimm D (1982) Methode als Machtfaktor. In: Horn N (ed) *Europäisches Rechtsdenken in Geschichte und Gegenwart – Festschrift für Helmut Coing*. Beck, München, pp 469–492
- Guthrie C, Rachlinski JJ, Wistrich AJ (2001) Inside the judicial mind. *Cornell Law Rev* 86(4):777–830
- Jolls C, Sunstein CS, Thaler R (1998) A behavioral approach to law and economics. *Stanford Law Rev* 50(5):1471–1550
- Kirchner C (1991) The difficult reception of law and economics in Germany. *Int Rev Law Econ* 11(3):277–292
- von Kleinwächter F (1883) *Die Kartelle – Ein Betrag zur Frage der Organisation der Volkswirtschaft*. Wagner, Innsbruck
- Litschka M, Grechenig K (2010) Law by human intent or evolution? some remarks on the Austrian school of economics' role in the development of law and economics. *Eur J Law Econ* 29(1):57–79
- Manne HG (1962) The higher criticism of the modern corporation. *Columbia Law Rev* 62(3):399–432
- Manne HG (1965) Mergers and the market for corporate control. *J Polit Econ* 73(2):110–120
- Manne HG (1966a) In defense of insider trading. *Harv Bus Rev* 44(6):113–122
- Manne HG (1966b) *Insider trading and the stock market*. Free Press, New York
- Manne HG (1967) Our two corporate systems: law and economics. *Virginia Law Rev* 53(2):259–284
- Manne HG (2005) How law and economics was marketed in a hostile world: a very personal history. In: Parisi F, Rowley C (eds) *The origins of law and economics*. Elgar, Cheltenham, pp 309–327
- Mataja V (1888) *Das Recht des Schadenersatzes vom Standpunkt der Nationalökonomie*. Duncker & Humblot, Leipzig
- Mataja V (1889) *Das Schadensersatzrecht im Entwurf eines Bürgerlichen Gesetzbuches für das Deutsche Reich*. *Archiv für Bürgerliches Recht* 1:267–282
- Mattei U, Pardolesi R (1991) Law and economics in civil law countries: a comparative approach. *Int Rev Law Econ* 11(3):265–275
- McAdams R (2000) Experimental law and economics. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics*, vol 1. Elgar, Cheltenham, pp 539–561
- Menger A (1890) *Das bürgerliche Recht und die besitzlosen Volksklassen : eine Kritik des Entwurfes eines bürgerlichen Gesetzbuchs für das Deutsche Reich*. Laupp, Tübingen
- Parisi F (2005) Methodological debates in law and economics: the changing contours of a discipline. In: Parisi F, Rowley C (eds) *The origins of law and economics*. Elgar, Cheltenham, pp 33–52
- Parisi F (2009) Multidisciplinary perspectives in legal education. *University Of St Thomas Law Journal* 6(2):347–357

- Pearson H (1997) *Origins of law and economics: the economists' new science of law, 1830–1930*. Cambridge University Press, Cambridge, MA
- Petersen N (2010) Braucht die Rechtswissenschaft eine empirische Wende? *Der Staat* 49(3):435–455
- Posner RA (1973) *Economic analysis of law*, 1st edn. Little Brown, Boston
- Posner RA (1979) Utilitarianism, economics, and legal theory. *J Leg Stud* 8(1):103–140
- Posner RA (1980) The ethical and political basis of the efficiency norm in common law adjudication. *Hofstra Law Rev* 8(3):487–507
- Posner RA (1987) The decline of law as an autonomous discipline: 1962–1987. *Harv Law Rev* 100(4):765–780
- Reimann M (2014) The American Advantage in Global Lawyering. *Rabels Zeitschrift für ausländisches und internationales Privatrecht* 78(1):1–36
- Schäfer HB, Ott C (1986) *Lehrbuch der ökonomischen Analyse des Zivilrechts*, 5th edn. Springer, Berlin
- Schanze E (1993) Ökonomische Analyse des Rechts in den USA: Verbindungslinien zur realistischen Tradition. In: Assmann H-D et al (eds) *Ökonomische Analyse des Rechts*. Mohr, Tübingen, pp 1–16
- Singer JW (1988) Legal realism now. *Calif Law Rev* 76(2):465–544
- Steinitzer E (1908) *Ökonomische Theorie der Aktiengesellschaft*. Duncker & Humblot, Leipzig
- Sunstein CR (2000) *Behavioral law & economics*. Cambridge University Press, Cambridge, MA
- Towfigh E, Petersen N (2014) *Economic methods for lawyers* (forthcoming)
- Weigel W (1991) Prospects for law and economics in civil law countries: Austria. *Int Rev Law Econ* 11(3):325–329
- Wieacker F (1967) *Privatrechtsgeschichte der Neuzeit unter besonderer Berücksichtigung der deutschen Entwicklung*, 2nd edn. Vandenhoeck & Ruprecht, Göttingen
- Winkler V (2004) Ökonomische Analyse des Rechts im 19. Jahrhundert: Victor Matajas “Recht des Schadensersatzes” revisited. *Zeitschrift für neuere Rechtsgeschichte* 26:262–281

Law and Finance

Afef Boughanmi¹ and Nirjhar Nigam²

¹University of Lorraine (IUP Finance),
BETA-CNRS Laboratory, Nancy, France

²ICN Business School, CEREFIGE and LARGE
Laboratory, Metz, France

Abstract

“Law and finance” is a rather new and evolving research topic, initiated by La Porta, Lopez-de

Silanes, Shleifer, and Vishny (henceforth LLSV). They investigated the differences between legal origins and their impact on economic performance. Two seminal papers published by LLSV in 1997 and 1998 addressed an important question: Does law matter? Their main idea was to evaluate the impact of legal protection of investors (shareholders and creditors) on three key areas: corporate governance, structure of ownership, and control and orientation of financial system. Their work was extensively referred by other researchers and is now considered as a new finance theory. So as the reader is able to appreciate the nuances of the theory which was postulated by LLSV, we propose a two-step approach: first, a detailed description of the theory which would be followed by a summarized discussion of related works on law and finance.

Law and Finance Theory: Contribution by LLSV

Based on a sample of 49 countries, LLSV had investigated the differences between the four legal origins: common law, French civil law, German civil law, and the Scandinavian civil law.

The common law which originated in England provides great flexibility to its judges. They can administer a case according to their discretion, and the outcomes can vary on a case to case basis, as long as the judgments are fair and in conformity with the law of precedent. The civil law was pioneered in France and is replete with statutes and written codes. Herein, judges cannot act with full liberty to exercise their discretionary power. (French civil code is a derivative of the Napoleon Code of 1804. Its main objective was to standardize laws. However, the German civil law of 1896 differs from the civil code due to its different origins. Similarly, Scandinavian civil law differs from both the French civil law and common law. Most of the countries, according to La Porta et al. (1998), were categorized into one of the four aforementioned and dominant laws.)

Law and Finance, Table 1 “Rule of law” within common and civil law countries (The data was collected between 1994 and 1995)

Legal origin	Number Of countries	Shareholders' rights	Creditors' rights	Rule of law
Common law	18	4	3,11	7,11
French civil law	21	2,33	1,58	3,91
German civil law	6	2,33	2,33	4,66
Scandinavian civil law	4	3	2	5
International mean	49	3	2,3	5,30

The empirical comparative studies of LLSV (these papers were published in 1997, 1998, 1999, and 2000), used an indicator called “rule of law” (Table 1) which measures the legal protection of the minority investors and is an aggregated indicator of ten dummy variables: six of which describe the rights of shareholders and four describe the rights of the creditors. Each dummy variable has a value of 1 if the right is valid within a country and 0 if not. These rights represent many aspects of the law such as security exchange, bankruptcy, corporate law, and commercial code.

Their main results postulate the idea that common law countries generally provide better investor protection than civil law (French, German, and Scandinavian) countries whereas French civil law countries provide the least level of investor protection.

In order to evaluate the impact of law on financial and economic systems, LLSV used the legal investor protection variable for explaining the differences between all 49 countries of their sample with regard to (1) corporate governance, (2) structure of ownership of firms, and (3) structure and development of financial systems.

Law and Corporate Governance Models

LLSV proposed a new approach to corporate governance. If we consider corporate governance as a set of internal and external mechanisms which promote the efficiency of a corporation, then the legal system approach is a constituent of the external mechanisms. It is defined as a set of rules conceived to protect outside investors, in minority, against managers, the board of directors, and stakeholders. Hence, the need to integrate a sound legal structure and an effective form of corporate

Law and Finance, Table 2 Shareholders rights and structure of ownership and control

Shareholders rights index	Widely held firms	Families owned firms	State-controlled firms
High level	47,92%	24,58%	13,75%
The USA (5)	0,80	0,20	0,00
The UK (5)	1,00	0,00	0,00
Low level	27,33%	34,33%	22%
France (3)	0,60	0,20	0,15
Germany (1)	0,50	0,10	0,25
Finland (3)	0,35	0,10	0,35
Mean of the sample	36,48%	30%	18,33%

governance becomes imperative, even more amidst the wake of the failures of multinational firms such as Enron, Vivendi, Parmalat, etc. Similarly, bankruptcy laws are crucial in the prevention and resolution of financial distress of corporations.

Law and Structure of Ownership and Control

From a list of 27 richest countries (LLSV chose 27 richest countries based on the income per head in 1993), LLSV (1998 and) used a sample of 20 large private and nonfinancial firms for their analysis (Table 2).

LLSV divided their sample of 27 countries into two groups: one group with high level of legal protection of shareholders and the second with low degree of protection.

They concluded that a high level of dispersion of ownership and control is observed within countries offering good shareholder protection (as the USA and UK). It was observed that the percentage of large companies which were characterized by a dispersed shareholding was 47.92% in the first group and 27.33% in the second group.

Legal Structures and Development of Financial Systems

LLSV established an empirical correlation between the share of external financing (bank debt and equity) in the total financing of firms and the structure of the legal system. They observed that the common law countries have higher levels of financing by market capitalization as compared to civil law countries, particularly the French civil law countries. They demonstrated that countries that offer better protection to the shareholders benefit from more developed financial markets and a greater number of stocks per capita. Similarly, an important legal protection of creditors involves a developed banking system. In general, the common law countries have market-based financial systems, and civil law countries have bank-based systems. At the same time, LLSV remark that French civil countries are financially less developed than common law countries (Table 3).

The seminal papers of LLSV have been the subject of many criticisms and extensions. We will first present the principal critics and then the proposition for an extension of this legal approach. We can safely concur that these works are encapsulated within the framework of law and finance literature and have in their own rights the capacity to be termed as a new financial theory.

Law and Finance, Table 3 Investor protection and structure and development financial system

Countries	Market capitalization/ GNP	Number of listed companies per capita	Bank ^a
Common law	0,60	35,45	0,83
French civil law	0,21	10,00	0,67
German civil law	0,46	16,79	0,92
Scandinavian civil law	0,30	27,26	0,90
Mean of the sample	0,46	21,59	0,83

^aBank = the ratio of bank loans compared to the sum of assets related to domestic credit of the central bank and bank lending (Levine 1998)

Scope of the Literature of “Law and Finance”: A Critical Analysis

The work of LLSV, however, has also been the focus of much criticism. In particular, two groups of studies are worth a mention in our analysis and postulation. On the one hand, there is a group of several studies which presents the critical analysis of their methodologies and its deficiencies while offering alternative methodologies. These studies constitute the main core of “law and finance” literature. On the other hand, there is a second group of studies which offers critical insights into the theoretical aspects of the work of LLSV and tries to elaborate on their work. This kind of work is contributing to the growth and development of a “new political economy” theory.

A Critical Approach of LLSV’s Methodology

These types of articles proposed that LLSV’s seminal works have significant legal deficiencies of theoretical and empirical statures that are highlighted in several recent studies designed to challenge their main conclusions. We can segregate such papers into four categories:

1. *The legal indicators built by LLSV are composed of only ten rights of investors:* Many articles point out the fact that the postulated ten rights are not sufficient to effectively describe the national legal protection of investors and a legal system in entirety. Then several papers propose new legal indicators based on largely more than ten rights (e.g., Pistor et al. (2000)). Also, Holderness (2006) highlights that LLSV’s use of aggregated data, in particular, the indicator of the degree of concentration of ownership. According to LLSV, the latter is explained only by the legal system (national factor) and is not influenced by other factors (firm size, age, etc.).
2. *LLSV’s studies are cross-sectional studies:* This is also the case of many studies using legal indicators of LLSV, such as the “doing business” annual reports published by the World Bank. (WORLD BANK, DOING BUSINESS REPORTS (since 2004), available at <http://www.doingbusiness.org>.) Indeed, in

order to attribute scores of an annual growing sample of countries, the World Bank measures the “rule of law” with respect to a large number of countries all over the world. However, this method neglects the dynamic aspect of legal factors. Many studies have proposed more sophisticated legal indicators as compared to LLSV. Some of them focused on the time series data with respect to one particular country, while the others took multiple countries into account. For example, the research conducted by Hyytinen et al. (2003) about Finland shows the evolution of legislation over the period 1980–2000, particularly after the financial scandals. The authors find that the improvement of shareholder rights leads to restructuring of the financial market. Changes in investor protection are, therefore, a by-product of financial market development. And, similar conclusions are obtained by studies conducted on other countries, especially the ones with French civil law.

3. One of the criticisms of LLSV’s works is that their idea was to prove the legal and financial superiority of common law countries as compared to civil law countries. In order to challenge this conclusion, Blazy, Boughanmi, Deffains, and Guigou conducted a study in 2011 and found that the legal protection of investors in France is not as bad as presented by the works of LLSV, especially when we consider the dynamics of the evolving law and financial structures and the ability of the French system to quickly adapt progressively to the changing reforms. The highlights of this study demonstrate that study of LLSV is a relatively descriptive work used in order to prove some normative hypothesis.
4. *The classification of countries according their legal origins*: Common law versus civil law is deeply criticized. Modigliani and Perotti (2002), contrary to LLSV, base their analysis not on the legal origin but on the quality of legal rules. They thus prove the superiority of civil law countries as compared to Scandinavian Anglo-Saxon countries with regard to their impact on the quality of law enforcement on financial development. From a general point

of view, an international comparison of investor rights similar to those carried out by LLSV but on more recent data may not reach the same conclusions. This assumption is justified by the evolution of legislation in several countries, particularly following the recent financial scandals.

Ultimately, the LLSV vision is based on a “shareholder value” whose objective is to maximize shareholder value of the company. In these conditions, the superiority of common law countries with those of codified law is justified only by the higher level of legal protection of shareholders in the former than in the latter. Their results can be mitigated if we consider the “stakeholder value.” In this model, the objective is to maximize the total value of the company. The interests of all stakeholders are taken into account. Shareholders are set to the same status as other stakeholders (employees, customers, suppliers, etc.). An empirical study conducted by the OECD (1999) in fact shows the existence of two groups of countries. On the one hand are countries of continental Europe and Japan that are characterized by high protection of employees and low investor protection, while on the other hand are the USA and Great Britain where the reverse situation is observed (Pagano and Volpin 2001).

Theoretical Critic Research: The Political Theory

Some authors believe that political theory is more compelling to explain financial and economic development than the law and finance theory. Also, we can decide to classify these works as taking part of the “law and finance” theory.

Thus, the main idea of these critics’ articles is that financial development cannot be exclusively explained by legal origin theory, because the financial systems are evolving rapidly, while the legal origin remains stable and evolves languidly. In addition, the structure of the financial system has evolved over the last century. This change has been effected more by political lobbying, than by legal reforms (Rajan and Zingales 2001). Thus, the political theory raises the following questions: Why and how can the political process influence

financial systems? Is the effect of the political process on financial development more remarkable than the legal system? LLSV neglects the political aspect in their explanation of the superiority of common law as compared to civil law. They do not consider that the judges of common law countries may want to serve political interests. The political aspect is an important factor in the analysis of financial development and economic performance. Several studies are based on the proposition that political factors influence corporate governance, not only through the law but also through other channels of transmission. LLSV approve the idea but argue that the law remains the principal transmission channel through which politics affects corporate governance. The political theory is built around two axes. The first is represented by an economic approach called the "New Political Economy." The second axis is constituted by an ideological inspiration "the 'ideological' political theory."

The New Political Economy

This approach aims to provide some answers to two main questions. First, it seeks to understand why the regulations of financial systems are often imperfect and reduce the development of financial markets. On the other hand, it seeks to determine as to why some countries are characterized by inefficient financial institutions or by low-quality implementations of financial regulations. This new theory of political economy, initiated by Pagano and Volpin (2001), addresses these questions using the economic analysis methods. Originally, political economy focused on the interconnection of politics and macroeconomics. The New Political Economy, however, aims to analyze the political interference in the financial market. History bears testimony to the fact that political intervention in the financial market is not confined to periods of crisis and economic depression. Indeed, political pressure from special interest groups and politicians' concerns about the progress of their careers lead to specific policy interventions in financial markets such as nationalization, privatization, etc. (Pagano and Volpin 2001). This raises the question of how interest groups can influence the rules of law through the

political process. The LLSV's legal argument is based on the idea that law is an exogenous factor of financial development. The policy approach outlined by Pagano and Volpin (2001) however requires that the degree of protection of investors and the quality of enforcement of the law cannot be considered as exogenous variables. In fact, the political process has an impact on the development of legal rules and on their applications. Thus, in order to maximize their economic performance, voters (individuals or firms) form interest groups. These groups influence politicians who will therefore introduce the legal reforms required and suggested by the interest group voters. These reforms will therefore influence the economic results in favor of interest groups. (We can hold the same reasoning within corporations: stakeholders can prefer a low level of legal protection of minority investors; actually they can expropriate them easier. Then, they can appeal to the help of employees in order to require a weak protection to minority shareholders. This can work only when employees are not shareholders of the firm (Pagano and Volpin 2001).) On the other hand, holding a significant stake in the firm by the directors may constitute a solution to the problem of expropriation of minority shareholders by the stakeholders (LLSV 1999). When the manager is himself a shareholder, conflict of interest between shareholders and the manager is reduced.

An empirical study conducted on some OECD's countries classifies the sample into two groups: a group of "corporatist" countries and another composed of non-corporatist countries. In the former, the level of protection of employees is high to the detriment of investors. In contrast, the latter giving more importance to investors. Pagano and Volpin noticed that this classification is the result of the political process influences that determine the orientation of a country toward better protection in favor of employees or in favor of shareholders. The potential cost of corporatism may be a low level of external funding and underinvestment. Therefore, the arrangement creates a social inefficiency *ex ante*. Indeed, minority shareholders (with no control rights) do not want to take the risk of being expropriated of their funds by insiders. This rationing of capital

can be very expensive for new companies looking for funding (Pagano and Volpin 2000). The differences in terms of corporate governance, between continental Europe and the USA, can be explained by the fact that in the latter there are no shareholders holding control blocks. In contrast, in continental Europe, they, as employees and firms' managers, are insiders. Accordingly, the managers of American firms have more influence on politicians so they can better safeguard their revenues. This may explain the different waves of regulations in the USA to restrict the power of bloc-blocks of ownership and control of firms (Roe 1994 and Roe and Bebhuk 1999). It follows that the political forces influencing the structure of control rights and corporate finance. The balance of power between different stakeholders (the majority shareholders, minority shareholders, managers, and employees) determines the relative weight given by the firms in terms of the profits to shareholders and well-being of employees. The political factor also affects the development of banks. Policy reforms intended to increase the legal protection of creditors may induce a reduction in the work of selecting borrowers. A reform that aims to increase the efficiency of the legal system may encourage banks to reduce the frequency of verification of the results of borrowing.

The "Ideological" Political Theory

Initiated in 1999 by Roe, this approach assumes that policy choices that determine the protection of investors and the quality of its implementation are driven by ideological factors. Indeed, Roe (1999) emphasizes that the different models of corporate governance between the USA and the countries of continental Europe are the result of an incompatibility of the American ideology with specific social democracy European countries. According to Roe, European states must maintain a social pact between all classes. The question that arises is why this thesis based on the ideological factor paves a prominent place for the political process. Given the fact that ownership is concentrated in the most developed countries, Roe (2001) assumes that the political factor has an impact on the ownership structure. The

governments affect the concentration of ownership of firms in countries of continental Europe. In contrast, in the USA, the low level of government's intervention has eased the development of the managerial firm. According to Roe, in a social democracy, the social factor is clearly present. Indeed, politicians encourage the directors of firms to ignore the interests of shareholders in order to preserve those of employees and not dismiss them. In these circumstances, the directors submit easily to political interference and pressure. However, this intervention does not lead to the same result when capital is concentrated. Indeed, majority shareholders do not yield so easily to the pressures of politicians because of their financial goals tied to their firms. By influencing corporate governance, political leaders of socially democratic countries induce more agency costs (compared to control costs incurred by minority shareholders of the managerial firms). As a result, minorities opt for the concentration of ownership in order to be blockholders of control. To test these predictions, Roe undergoes an empirical study based on 16 countries of different political outlooks. The main conclusions are the influence of political positioning is confirmed, and countries with "left" politically position have corporations with high level of ownership concentration. In the USA, a low level of concentration of ownership is a striking characteristic of the firms.

The results of the ideological approach of the political theory demonstrate that the continental European countries, characterized by a social democracy, seem less effective than the USA in terms of dispersion of ownership. The governments of socially democratic countries may create gaps between shareholders and managers and employees, and consequently managerial firms find raising external capital relatively cumbersome in these countries. It is noteworthy that empirical studies found out that until the early twentieth century, a capital market more developed than the US capital market characterized France. However, this trend was reversed in the 1980s. In recent years, the differences between the two countries are not clearly evident. Hence, the discrepancies that may exist between these two

countries may not necessarily determine the superiority of one over the other.

Roe notes that in the socially democratic countries, stability may compensate the loss in efficiency caused by the weak protection of shareholders and that over a long period these countries have a high level of productivity. The author explains the specificity of social democracies using an analysis in terms of “path dependence.” He believes that continental European countries cannot neglect the social factor because of the influence of past wars and encourages them to opt for social stability. The ideological approach is focused on the fact that the solution to problems, caused by the opening of capital of firms and the willingness to develop market economy, can provide stability to their political and social structure. Solving these problems is not only limited to setting up a legal environment. Legal reforms must also be accompanied by political reforms to further develop the financial markets and foster economic growth.

The ideological approach of the political theory differs from that of the New Political Economy. According to the latter, the degree of legal protection of investors is the result of the economic interests, as it assumes that political decisions are influenced by economic interests and not by ideological and social factors. However, it is difficult to distinguish those ideological political choices from economic political choices.

Conclusion

The “law and finance” literature integrates the legal factors into the analysis and thus provides answers for three basic questions: the explanation of the differences between (i) the structures of ownership and control, (ii) modes of corporate governance, (iii) and the development and organization of financial systems. The several empirical studies conclude the existence of two types of legal origins (civil law and common law). Based on the idea of the superiority of common law, the seminal papers (LLSV) deduce a number of

results on corporate governance and the development of financial systems. First, they challenge the widely held model described by Berle and Means (1932) by showing that it is only representative of few number of countries. Indeed, they highlight the existence of a strong propensity of companies with a concentrated family shareholding. Second, the proponents of the “law and finance” theory note that the dispersion of ownership and control is a proof of the strong protection of minority shareholders.

Third, this legal approach shows that common law countries are characterized by a model of corporate governance that is exercised primarily through the control of external shareholders. In civil law countries, the companies are controlled internally by its main shareholders.

Fourth, the legal approach empirically verifies the hypothesis of a relationship between the legal system and financial development. It highlights that those countries that provide better protection to shareholders (common law countries) as compared to those who do not (civil law countries) enjoy a better-developed market economy.

Fifth, the legal approach leads to the conclusion that legal protection of investors must be represented by not only the content of legal rules but also the efficiency and quality of their enforcement.

Finally, the legal approach of LLSV is the subject of several theoretical and empirical critics. The main one is the political theory. The political theory is built around two axes. The first is represented by an economic approach called the “New Political Economy.” The second axis is constituted by an ideological inspiration “the ‘ideological’ political theory.” The main argument of the first is that the degree of investor protection and the quality of enforcement of the law cannot be considered as exogenous variables in the functioning of the financial system. The second approach excludes the legal factor and explains the legal differences between countries in terms of financial development, by ideological factors.

We note also that the framework of law and finance theory is extended around many other axes. Indeed, many studies are based on “stakeholder” perspective that extends the analysis to

include the legal protection of employees and bondholders and not only shareholders. In addition, we include within the “law and finance” literature many studies that provide a more comprehensive view of distressed firms and bankruptcy. These studies are based on new methodology that provides several legal indexes used to describe bankruptcy law. Thus, they provide a more comprehensive view of bankruptcy than that proposed by LLSV, who were built upon four legal indexes that mainly focused on secured creditors’ rights and studied only the reorganization procedure, while the majority of bankruptcies end up in liquidation procedure (see Blazy et al. (2012, 2013)).

References

- Berle A, Means G (1932) The modern corporation and private property. Macmillan, New York
- Blazy R, Boughanmi A, Deffains B, Guigou J-D (2012) Corporate governance and financial development: a study of the French case. *Eur J Law Econ* 33(2):399–445
- Blazy R, Chopard B, Nigam N (2013) Building legal indexes to explain the recovery rates: an analysis of French and UK bankruptcy codes. *J Bank Financ* 37(6):1936–1959
- Holderness C (2006) A contrarian view of ownership concentration in the United States and around the world. American Finance Association 2006 Boston Meetings
- Hyytinen A, Kuosa L, Takalo T (2003) Law or finance: evidence from Finland. *Eur J Law Econ* 16:59–89
- La Porta R, Lopez-de-Silanes F, Shleifer A, Vishny R (1998) Law and finance. *J Polit Econ* 106(6):1113–1155
- Levine R (1998) The legal environment, banks, and long-run economic growth. *J Money, Credit Bank* 30:596–613
- Modigliani F, Perotti E (2002) Security markets versus bank relationships? Cross country evidence. *J Financ Intermed* 9:26–56
- Pagano M, Volpin P (2000) The political economy of corporate governance. London Business School papers
- Pagano M, Volpin P (2001) The political economy of finance. *Oxf Rev Econ Policy* 17(4):502–519
- Pistor K, Raiser M, Gelfer S (2000) Law and finance in transition economies. *Econ Transit* 8:325–368
- Rajan RG, Zingales L (2001) Financial system, industrial structure and growth. *Oxf Rev Econ Policy* 17(4):467–482
- Roe M (1994) Strong managers weak owners: the political roots of American corporate finance. Princeton University Press, Princeton
- Roe M (1999) Political preconditions to separating ownership from control: the incompatibility of the American public firm with social democracy. Working paper, Columbia Law School
- Roe M (2001) Les conditions politiques au développement de la firme managériale. *Finance Contrôle Stratégie* 4(1):123–182
- Roe MJ, Bebhuk LA (1999) A theory of path dependence in corporate ownership and governance. *Stanford Law Rev* 52:127
- The World Bank (2004) Doing business in 2004: understanding regulation. A copublication of the World Bank, the International Finance Corporation, and Oxford University Press, p 211. <http://rru.worldbank.org/Documents/DoingBusiness/2004/DB2004-full-report.pdf>

Law Enforcement by Government

► Public Enforcement

Law Merchant

► Lex Mercatoria

Law of the Fair

► Lex Mercatoria

Legal Aid

Myriam Doriat-Duban¹ and Eve-Angéline Lambert^{2,3}

¹Department of Economics, BETA UMR 7522, Université de Lorraine, Nancy, France

²BETA, CNRS, University of Strasbourg, Strasbourg, France

³BETA UMR 7522, Université de Lorraine, Nancy, France

Definition

Legal aid is a type of financial aid in which the State bears the costs of litigation to allow the

poorest people access to the courts. It therefore reduces the litigation costs borne by the beneficiary litigant. It may thereby have an impact on the litigants' behavior (decision to sue, dispute resolution procedure, precaution taking, incentive to engage in a criminal activity). It may also have an impact on lawyers' activities and the effort they devote to defending their clients' interests. It may therefore be useful to compare this method of financing access to justice with other possible alternatives, such as contingent/conditional fees and legal expenses insurance.

Abstract

Legal aid is a financial aid of the State to make access to justice for poor people easier. Law and economics scholars study its impact on the litigants' behaviour (decision to sue, dispute resolution procedure, precaution taking, incentive to engage in a criminal activity) but also on lawyers' activities. Finally, legal aid is compared with other possible alternatives, such as contingent/conditional fees and legal expenses insurance.

Introduction

Legal aid (LA) conforms to the principle of solidarity and the obligation, imposed at European level by the European Court of Human Rights and the Court of Justice of the European Union, to allow all citizens to defend their rights through the courts. The methods used to grant aid vary from one country to another but are generally based on the beneficiary's financial resources (upper limit of income) and, under certain circumstances, the merit of the case. LA may be proportional, if it covers a percentage of the expenses incurred by beneficiaries for the defense of their rights (percentage of up to 100%), or granted on a flat-rate basis if beneficiaries receive a fixed amount (corresponding to the payment of a certain number of hours to the lawyer or to funding a specific stage of the litigation) and then have to cover other additional expenses. It may cover the

costs incurred for a dispute in civil as well as criminal matters.

The economic literature devoted to LA is quite insubstantial, and the studies mainly concern England and Wales (Rickman et al. 1999), Scotland (Stephen 1998, 2001), France (Doriat-Duban 2001; Ancelot et al. 2012), and occasionally European countries (Lambert and Chappe 2014). All of these studies are based on the premise of a growing and continuous increase in spending on LA in the justice budget. They mostly have two separate aims: to explain the growing recourse to LA and to understand how LA affects the behavior of the parties (beneficiary, lawyer, and defendant). The insight they provide may thus be useful for public policies seeking either to reform LA or replace it with other methods of financing access to justice (contingent/conditional fees or legal expenses insurance).

Impact of LA on Access to Justice and Effectiveness of the Law

LA may be granted in civil or criminal cases.

In civil matters, the studies focus on the influence of LA on incentives to sue, dispute resolution procedure, and damage prevention. In a model based on the optimism of the parties, Doriat-Duban (2001) shows that LA increases the number of legal proceedings by reducing the costs of access to justice. Nevertheless, if it allows for the payment of lawyers' fees when a transaction is concluded before the seizing of a court, then the frequency of prosecutions will fall, and the prevention and compensation of damage will rise. As far as the dispute resolution procedure is concerned, the impact of LA seems uncertain because two effects are exerting opposite influences: the drop in negotiation costs increases the chances of an amicable settlement and the decrease in litigation costs increases the probability of judgment.

More recently, Lambert and Chappe (2014), in a model with asymmetric information, have studied the effect of LA on incentives to take precautions, the decision to take legal action, and the amount of the expenditure incurred by the plaintiff in proceedings. They show that flat-rate or proportional LA has a positive effect on

incentives to sue and, in advance, on incentives for potential perpetrators of damage to take precautions. However, the impact on the number of litigation cases is unclear because everything depends on the dominant effect, from a rise in number of prosecutions to a drop in the number of accidents.

In criminal matters, Garoupa and Stephen (2004) mention that three theoretical arguments against LA are traditionally put forward: it may have a negative effect on deterring people from committing crimes by reducing the costs of their defense; it could be a source of inequalities in the sanctions expected for the same offense; and it could be replaced by a more effective income-based subsidy. Nonetheless, the authors present an argument in favor of LA. To this end, they consider it not merely as an expense in the State budget but as a public subsidy contributing to the optimal application of justice. More specifically, in the event of miscarriages of justice, LA improves the situation of all defendants (innocent and guilty) because it reduces the marginal cost of their defense; but as a guilty party is presumed to be more likely to plead guilty than an innocent party, its effect is to provide more help for the defense of innocent parties. If this effect was sufficiently pronounced, then governments wishing to maximize social well-being would be well advised to implement a LA system. This argument is strengthened if subsidizing the expenditure on defense reduces the sanction incurred by innocent parties by allowing them to improve their defense.

Impact of LA on Lawyers

A recurrent question on LA concerns its growing budgetary burden: whereas certain authors mention the very steep rise in expenditure on LA (Gray 1994; Stephen 1998, 2001; Goriely et al. 1997; George 2006), others explain this rise as a problem of moral uncertainty linked to its attribution methods in the majority of countries in which the legal institution (the principal) delegates to the lawyer (the agent) the power to decide whether or not to defend a case and the number of hours to devote to it. However, the combination of asymmetric information and different incentives for

these two parties generates opportunistic behavior in lawyers. Furthermore, LA could remove the incentives for clients to monitor the amount of effort their lawyer devotes to their defense (Bevan et al. 1994). The massive rise in expenditure on LA that began in the 1980s would therefore be due to a supply-induced demand, with the lack of opportunities for lawyers in privately financed cases having increased the relative attractiveness of cases financed by LA (Bevan 1996).

Extending this analysis, Gray et al. (1999) sought an empirical link between the increased expenditure on LA and the supply of lawyers' services. In particular, they set out to determine whether lawyers adjust the level of their service according to their opportunity costs. The attention paid to lawyers' efforts is justified for two reasons: one is theoretical linked to the aforementioned moral uncertainty, and the other is empirical linked to the concomitant rise in public expenditure on LA and in the number of lawyers (in England and Wales). In this way, lawyers may have been able to make upward adjustments to the services provided in the context of LA, due to the drop in the opportunity costs that might have been represented by relinquishing activities previously carried out on a monopolistic market that is now open to competition. In other terms, as the other cases became less lucrative, those covered by LA became relatively more attractive. Different elements are then identified as possible explanations for the rise in spending on LA, including the volume of disputes and the unit cost of each case, for a given region. The volume of disputes depends on the number of cases covered by LA and the number of lawyers (or firms) handling these cases. The unit cost of the cases depends on the number of hours of work provided by the lawyer and the hourly rate applied. If this rate is assumed to be constant, then only two variables can explain variations in the unit cost: an "input" effect reflecting the variation in the number of hours devoted to a case and a "volume" effect reflecting the change in the number of cases handled by a firm, in the presence of diseconomies of scale at the regional level. In addition, the unit cost may increase for two other reasons: the

arrival of relatively inexperienced lawyers on the market which reduces economies of scale and experience effects and the fact that in response to heightened competition, firms are accepting cases of lesser merit to which they are devoting greater effort. This empirical analysis therefore shows that lawyers may increase the budgetary cost of LA, irrespective of the preferences of their clients and the public authority, especially if little monitoring is carried out.

LA Versus Other Methods of Financing Access to Justice

In view of the different effects of LA identified by the economic literature, consideration should be given to its possible alternatives. This raises the question of the consequences of transferring State-funded litigation expenditure to another agent: the lawyer in a system of contingent or conditional fees and the insurer in a legal expenses insurance system.

Concerning transfer to lawyers, Lambert and Chappe (2014) compare the effects of LA with those of contingent/conditional fees. They show that the probability of prosecution is higher in the second case, insofar as the plaintiff bears no expenditure and consequently no risk, which is borne by the lawyer. However, for a given expenditure borne by the plaintiff or lawyer, LA allows for an increase in expenditure on the proceedings, which increases the probability of the plaintiff winning and, in advance, also increases the incentives for the potential perpetrator of damage to take precautions.

Concerning transfer to insurers, both LA and legal expenses insurance tend to strengthen the plaintiff's negotiating position: by covering some or all of the litigation costs, they encourage him/her to demand more in order to abandon the litigation. Nevertheless, as explained by Ancelot et al. (2012), this heightened risk of congestion of the courts is reduced if account is taken of the agency relationships forged in adversity and the possible benefits for lawyers of coming to an amicable settlement. In addition to the comparison of the two systems, the consequences of replacing LA with legal expenses insurance and its impacts on the litigation resolution mode are

identified. The arrangements accepted by plaintiffs covered by a legal expenses insurance policy appear to be higher than those demanded by plaintiffs benefiting from LA, irrespective of their risk aversion. Consequently, LA appears to be more favorable to arrangements than legal expenses insurance because it makes plaintiffs less demanding. The development of legal expenses insurance could therefore appear to be desirable for plaintiffs but not for the legislator wishing to reduce legal proceedings.

Summary

Legal aid has seldom been explored by the economic literature. The studies – mainly European – provide interesting results regarding the avoidance of conflicts, in terms of the incentives to prosecute, to come to an amicable settlement, and to take precautions. They also shed new light on the behavior of lawyers. These analyses may prove to be particularly useful with the prospect of reforms seeking to reduce the burden of LA in government budgets, perhaps by replacing it with other methods of financing access to justice (contingent/conditional fees or legal expenses insurance). However, empirical studies that confirm or invalidate the main theoretical results remain sorely lacking.

Cross-References

- [Access to Justice](#)
- [Litigation and Legal Expenses Insurance](#)

References

- Ancelot L, Doriât-Duban M, Lovat B (2012) Aide juridictionnelle et assurance de protection juridique: coexistence ou substitution dans l'accès au droit. *Rev Fr Econ* XXVII(4):115–148
- Bevan G (1996) Has there been supplier-induced demand for legal aid? *Civ Justice Q* 15:98–114
- Bevan G, Holland T, Partington M (1994) Organising cost-effective access to justice. In: Paterson A, Goriely T (eds) *Resourcing civil justice: oxford readings in socio-legal studies*, Social Market Foundation, London

- Doriat-Duban M (2001) Analyse économique de l'accès à la justice: les effets de l'aide juridictionnelle. *Rev Int Droit Econ* 15(1):77–100
- Garoupa N, Stephen FH (2004) Optimal law enforcement with legal aid. *Economica* 71:493–500
- George JP (2006) Access to justice, costs, and legal aid. *Am J Comp Law* 54:293–315
- Goriely T, Tata C, Paterson A (1997) Expenditure on Criminal Legal Aid: Report on a comparative pilot study of Scotland, England and Wales, and the Netherlands, Central Research Unit, The Scottish Office, Edinburgh, HMSO.
- Gray A (1994) The reform of legal aid. *Oxf Rev Econ Policy* 10:51–67
- Gray A, Rickman N, Fenn P (1999) Professional autonomy and the cost of legal aid. *Oxf Econ Pap* 51:545–558
- Lambert E-A, Chappe N (2014) Litigation with legal aid versus litigation with contingent/conditional fees. *Rev Law Econ* 10(1):95–115
- Rickman N, Fenn P, Gray A (1999) The reform of legal aid in England and Wales. *Fisc Stud Inst Fisc Stud* 20(3):261–286
- Stephen FH (1998) Legal aid expenditure in Scotland: growth, causes and alternatives. Law Society of Scotland, Edinburgh
- Stephen FH (2001) Reform of legal aid in Scotland. *Hume Pap Public Policy* 8(3):23–31

Legal Cost Insurance

► Litigation and Legal Expenses Insurance

Legal Disputes

Richard E. Wagner
Department of Economics, George Mason
University, Fairfax, VA, USA

Abstract

In bringing economic analysis to bear on whether a dispute is settled without trial, the presumed institutional setting is typically one of private property where the parties are residual claimants to their legal expenses. Many disputes, however, are between private and public parties. In these disputes there is a conflict between substantive rationalities because public parties are not residual claimants. Just as

the substantive content of action can vary depending on whether the actor operates within a context of private or common property, so can the substance of dispute settlement vary. While a public actor cannot pocket legal expenses that are saved through settlement, the expenses of trial can serve as an investment in pursuing future political ambitions.

JEL Codes: D23, D74, K40

Definition

A sizeable literature exists on the resolution of legal disputes, which is summarized to good effect in Miceli (2005), and with that summary containing an extensive bibliography. That literature largely operates within the framework of a dispute between two market-based entities where both parties operate within the substantive rationality associated with private property. In many legal disputes, however, one party is a public entity which operates within the substantive rationality associated with collective or common property, as Wagner (2007) explains in his treatment of public squares and market squares. This entry explores how the economic principles of dispute resolution might be modified when one of the parties to a dispute is a public entity. To provide a point of analytical departure, I start with a quick summary of dispute resolution when both parties operate inside a framework of private property. The rest of the entry explores how the analysis of how the resolution of legal conflicts might be modified when one of the parties is a public entity. The entry closes by considering some of the constitutional-level issues that arise in light of the societal tectonics that can be generated in the presence of this clash between rationalities.

Dispute Resolution Between Private Parties

Someone who buys a dilapidated hotel hires a construction firm to renovate the hotel. The

contract calls for a series of payments corresponding to various stages of completion of the work. The construction firm hires workers and contracts for the delivery of materials. The renovation is projected to take 2 years. Shortly after renovation starts, the hotel's owner discovers a considerable miscalculation in the financial projection on which the renovation was based. Correcting the mistake, however, lowers significantly the expected value of renovation. Consequently, the hotel owner revises the planned renovation and offers the owner of the construction firm a new contract, one that calls for less work, a later starting date, and a lower price. The owner of the construction firm refuses to accept the lower price, saying that the revised work plan and the different materials involved, along with complications due to the later starting date, warrants payment of the original price. The hotel owner refuses and demands that renovation proceed under the new plan, so the construction firm sues the hotel for the \$20 million.

The logic of settlement looks at the situation from the point of view of each party and asks whether there are conditions under which both parties could gain by settling the case rather than going to trial. For the plaintiff, the object of a trial is to acquire the amount requested; for the defendant, the object is to avoid having to pay that amount. For a plaintiff, the expected gain from going to trial, P_T , is $P_T = AP_P - E_P$, where A is the amount requested, P_P is the probability of success expected by the plaintiff, and E_P is the plaintiff's expense in pursuing the trial. For a defendant, the similar relationship is $D_T = AP_D - E_D$. In this instance, $A = \$20$ million. Suppose for each party the expected cost of pursuing a trial is \$2 million. Further suppose that each party believes its chance of success at trial is 50%. Under these conditions, each party has the same net expected value from going to trial, \$8 million.

The aggregate net value of going to trial is \$16 million, but the award generated through the trial is \$20 million. The other \$4 million is dissipated through litigation. This dissipated amount illustrates the potential gain from settling the case. A settlement where the defendant paid the

plaintiff \$10 million would split the settlement range evenly, leaving each party with \$2 million more than they could expect to obtain through trial. Any payment to the plaintiff between \$8 and \$12 million would provide gains to both parties. The existence of a settlement range fits with the observation that most commercial disputes are settled without trial. But not all such disputes are settled, nor is there any reason to expect that they should all be settled. For instance, there is no necessary reason for both parties to have the same perception of the outcome of a trial. In dealing with expectations about particular events, there is no reason to think that the sum of probabilistic beliefs must add to one. Each party might be optimistic about the outcome of a trial, as illustrated by each party thinking that its chance of success was 80%. In this case, the expected value of going to trial would be \$14 million for each party. With trial now offering what the parties perceive to be an aggregate value of \$28 million when only \$20 million will be awarded, the settlement range vanishes amid the inconsistent perceptions. To be sure, there is good reason to think that such procedures as discovery and deposition operate to narrow the distance between perceptions; however, while the tendency to settle disputes is economically intelligible, settlement is not economically necessary.

What is particularly significant about this formulation is that both parties operate within a framework of private property where each entity owns the value consequences of its actions. The basis for settlement resides in the institutional framework of private property wherein the parties can pocket the legal expenses avoided by settling the dispute. This framework fits normal actions of private law involving disputes between two commercial entities. But a good number of disputes in modern societies are between private and collective entities. With collective entities, however, there is no ownership of the value consequences of collective action, at least within democratic regimes. To explore the resolution of such disputes, it is necessary to take into account differences in practical rationality between differently constituted entities.

Form, Substance, and Rationality in Practical Action

Economists typically work with a purely formal notion of rationality, as expressed by the axiom that people make consistent choices. This formal notion of rationality pays no attention to the context within which choices are made. It does this by presuming that all choices are made within a market context where people choose among options according to market prices. This context, however, applies only to a subset of all choices and interactions within a society. Political choices, for instance, are not made within this context even though public choice theorizing tended to assume that they are doubtlessly to increase analytical tractability.

With respect to the preceding example, suppose the plaintiff is not a construction firm but an environmental control agency that is seeking to force the hotel owner to change the renovation in a manner that adds \$20 million to the expenses of renovation while adding nothing of potential value to future customers. The hotel owner can sue the environmental agency. The market-based settlement calculus, however, does not apply in this context. It applies to the hotel owner, of course, but it does not apply to the environmental agency because the agency operates outside the purview of private property and residual claimancy. The agency still engages in economic calculation because it always faces choices, but the context of calculation differs through the absence of residual claimancy, which means, among other things, that there is no market valuation of the environmental agency that some owner can claim directly.

There is a formal logic of choice, and there are substantive logics of practical action, as noted cogently by Pierre Bourdieu (1990, 1998). As a formal matter, it is reasonable to presume that people prefer more of what they value over less. As a substantive matter, however, what is valued depends on the context within which action is taken. Among athletes, for instance, practice will be the form by which the athlete prepares for competition. The substance of that practice, however, will depend on the context of competition.

For someone engaged in American-style football, weight lifting will be a significant component of practice. In contrast, someone engaged in billiards might not lift weights at all. There are many substantive logics of practice that can all be rendered sensible within a covering logic of form, as illustrated by Alasdair MacIntyre's (1988) distinction between a rationality of excellence and a rationality of effectiveness.

The logic of practice must relate cost to choice along the lines Buchanan (1969) sketches. Cost is thus the value an actor places on the option that is displaced by virtue of choosing the alternative. The cost of choosing to settle a case is the value of the option that the chooser foregoes by choosing instead to go to trial. The relevant cost, however, pertains to a person who faces a choice, and that cost differs between frameworks of private property and collective property, as Ringa Raudla (2010) notes in her treatment of incorporating institutional considerations into the theory of public finance. For instance, someone who fishes on a private pond will tend to return immature fish to catch later, while someone who fishes on a common pond cannot count on those fish still being there later when they are larger.

With respect to commercial disputants, it is reasonable to relate cost to perceptions of the comparative value of an enterprise in light of the options. With respect to the value of the enterprise in the preceding illustration when both parties agree on the 50-50 assessment of prospects of a trial, the value of the enterprise is \$2 million higher with settlement than with trial. That value, moreover, accrues to the owner of the enterprise. While in a large firm it will be the director of a legal office that makes such decisions, that director will bear a clear agency relationship with the owner of the firm. With collective entities, however, at least of the democratic form, there is no market for ownership, and so there is no value of the enterprise. It is still possible as a formal matter to treat collective entities as enterprises, but there is no market-based valuation of those enterprises. The relevant cost to the director of an environmental agency is not reflected in some calculation of enterprise value because there is no such value.

For such an agency, choices to settle or go to trial must be related not to some fictional construction of firm value but to the cost-gain calculus faced by the director of the agency, as modified by other political interests that the director values highly. There is still a formal logic of settlement, but the specific decisions regarding settlement are guided by the substantive logic of practice within a particular institutional framework. To get at substance requires that choices be related the chooser's valuation of perceived options at the level of practical action. Both football players and billiard players will practice for forthcoming competitions, but the substance of their practices will differ. In similar fashion, someone who fishes on a common fishing ground will act differently than someone who fishes on a privately owned pond.

Dispute Resolution Within a Mixed Economy

In her perceptive treatment of *Systems of Survival*, Jane Jacobs (1992) explained that well-working societies operate with a delicate balance between what she described as commercial and guardian moral syndromes. While her distinction wasn't identical to the distinction between commerce and government, it was close. She also explained that when the carriers of those syndromes commingle excessively, what she termed "monstrous moral hybrids" can arise. Such commingling comprises what is described as a mixed economy (Littlechild 1978; Ikeda 1997). The central issue that such commingling must confront is whether it supports or revises the character of the economic order. While democratic systems are based on a logic of equality and mutuality, there are plenty of analytical grounds for recognizing that they also face pressures leading toward a re-feudalization of the economy. It was to forestall such re-feudalization that Walter Eucken (1952) articulated the principle that state action should be market conformable. Such market conformability in Eucken's framework would present a barrier to the generation of monstrous moral hybrids in Jacobs's framework.

When both parties to a dispute operate within a private property framework, they both speak the same language of profit and loss. For instance, the defendant might be the hotel owner after renovation, and the plaintiff might be the owner of an adjacent marina who complains that the height of the hotel blocks a view of the setting sun, thereby degrading the value of the rooftop restaurant. Both participants speak the language of profit and loss. While each would have understandable incentives to exaggerate their cases, that exaggeration is limited by their positions of residual claimants and the possibility that settlement might lead to higher net worth than trial.

The setting changes if the plaintiff is an environmental agency or any political entity for that matter. The hotel still speaks the language of profit and loss, but the environmental agency speaks a different dialect, one that refers to social value that cannot materialize in anyone's account in particular. What results instead are ideological appeals that resonate with targeted interest groups along the lines of Pareto's (1935) treatment of ideology as amplified by Backhaus (1978) and as explored further in McLure (2007). In Pareto's terms, we observe the agency filing suit against the hotel, and we witness the agency advancing derivations or justifications based on claims of capturing social value. What we don't know about are the underlying motivations that led to the agency choosing that course of action, which Pareto called residues. The agency cannot claim to be trying to capture lost profits, and in this it is speaking truthfully because there are no profits that it can capture due to its status as a nonprofit entity. But in speaking of securing the social interest, it is advancing a proposition that is incapable of falsification. The head of the agency might well aspire to run for public office and sees this suit as a means of capturing favorable publicity. If so, the suit offers the head of the agency an investment at public expense in achieving that electoral outcome, though, of course, no person with such electoral aspirations would ever make such a declaration. In a suit between market participants, both parties can be honest and open in their aspirations. But when one party is a public entity, such rectitude necessarily and understandably recedes.

Disputes involving both private and public parties entail a form of Faustian bargain. Public parties can sue private parties without having to face the constraints upon the power to sue that private property imposes on market participants. In some instances, reasonable public purposes might be secured by the deployment of that power. But that power can also be employed in the pursuit of the private advantages of those who wield that power, as perhaps illustrated by an attorney general who is able to parlay legal expenditures into investments in an effort to seek higher elected office.

It is doubtful that there is any resolution to this particular Faustian bargain. This form of bargain bears a family resemblance to the problem of statistical decision theory as illustrated by Jerzy Neyman's (1950) treatment of the problem of the lady tasting tea. In that problem, a lady claims to be able to tell whether milk is added to tea that is already in the cup or whether tea is added to milk. It is in the nature of the problem setting that perfection is impossible. The greater the effort made to avoid granting the lady's claim when she really can't distinguish between the methods, the greater will be the frequency with which her claim will be rejected even though she can distinguish between the methods.

This problem setting has relevance for constitution of dispute resolution within society. Along the lines of Epstein (1985), a public agency might take private property under eminent domain, claiming that it had good public reason for doing so. After all, what else could it claim? The Fifth Amendment to the US Constitution requires that such takings be only for public purposes and that any such taking must be accompanied by just compensation. It is easy enough to give a Coasian gloss on this procedure. The ability of the hotel to build upward might destroy scenic opportunities elsewhere that are valued more highly than what is created by adding to the height of the hotel. There is a simple market test for this proposition. But in the absence of a market test, the burden falls on the public processes through which agency actions are determined. Some of those processes might be more market conformable than other processes. In any case, relationships between market-based and public-based entities create a source of turbulence that is

not present in relationships between market-based entities, due to differences in the substance of rational action.

References

- Backhaus JG (1978) Pareto and public choice. *Public Choice* 33:5–17
- Bourdieu P (1990) *The logic of practice*. Stanford University Press, Stanford
- Bourdieu P (1998) *Practical reason: on the theory of action*. Stanford University Press, Stanford
- Buchanan JM (1969) *Cost and choice*. Markham, Chicago
- Epstein RA (1985) *Takings: private property and the power of eminent domain*. Harvard University Press, Cambridge
- Eucken W (1952) *Grundsätze der Wirtschaftspolitik*. J. C. B. Mohr, Tübingen
- Ikedo S (1997) *Dynamics of the mixed economy*. Routledge, London
- Jacobs J (1992) *Systems of survival*. Random House, New York
- Littlechild SC (1978) *The fallacy of the mixed economy*. Institute of Economic Affairs, London
- MacIntyre A (1988) *Whose justice? Which rationality?* University of Notre Dame Press, Notre Dame
- McLure M (2007) *The Paretian school and Italian fiscal sociology*. Palgrave Macmillan, Basingstoke
- Miceli TJ (2005) Dispute resolution. In: Backhaus JG (ed) *The Elgar companion to law and economics*, 2nd edn. Edward Elgar, Cheltenham, pp 293–403
- Neyman J (1950) *First course in probability and statistics*. Henry Holt, New York
- Pareto V (1916 (1935)) *The mind and society: a treatise on general sociology*. Harcourt Brace, New York
- Raudla R (2010) Governing the budgetary commons: what can we learn from Elinor Ostrom? *Eur J Law Econ* 30:201–221
- Wagner RE (2007) *Fiscal sociology and the theory of public finance*. Edward Elgar, Cheltenham

Legal Formants

Bianca Gardella Tedeschi
DiSEI, Dipartimento per lo Studio dell'Economia e dell'Impresa, Università del Piemonte Orientale, Novara, Italy

Definition

Legal formants are the different components that concur to build any given legal system. The

comparative law scholar can understand the dynamics that shape any legal system through the analysis of their interactions. Legal formants analysis of legal systems have been developed by Rodolfo Sacco, and it is now a recognized and accepted methodology by comparative law scholars.

A Dynamic Approach to Comparative Law

Legal formants are the different components that concur to build any given legal system. The comparative law scholar can understand the dynamics that shape any legal system through the analysis of their interactions. The expression “legal formants” have been transplanted in comparative law from phonetics by Rodolfo Sacco, an Italian law professor based in Torino. “Legal formants” made their appearance in Sacco’s work since the early 1970s (Sacco 1974) and have been subsequently included in his seminal book on comparative law (Sacco 1980). Legal formants theory have been available to the English reader in 1991 (Sacco 1991a, b).

Comparative law is the academic subject that identifies the “plurality of rules and institutions . . . in order to establish to what extent they are identical or different” (Sacco 1991a: 5). To achieve this intellectual exercise, comparative law scholars have to collect legal rules, in order to be able to analyze and compare them. Sacco’s theory rests on the question “what is a legal rule?”, as the comparative law scholar has to extract the lesser object of his study from the wholeness of the legal system. From a national point of view, the answer to question, “what is a legal rule” may be simple: national jurist in civil law jurisdictions is willing to say that the legal rule is what a statute says, whereas in common law countries would be what case law affirms. Sacco, however, considers wrong to search for “the legal rule” on a specific subject. This attitude is, in his words, “the typical view of an inexperienced jurist” (Sacco 1991a: 21). In fact, any trained lawyer, “who proceeds from the axiom that there can be only one legal rule in force, recognizes, even only implicitly, that the living law contains many different elements:

statutory rules, the formulations of scholars, and the decisions of judges” (Sacco 1991a: 22). The national jurist tends to keep separate all these elements in the back of his mind. But still: “The statutes are not the entire law. The definitions of legal doctrinal scholars are not the entire law. Neither is an exhaustive list of all the reasons given for the decisions made by the courts. In order to see the entire law, it is necessary to find a suitable place for statute, definition, reason, holding and so forth. More precisely, it is necessary to recognize all ‘legal formants’ of the system and to identify the scope proper to each” (Sacco 1991a: 29).

“The number of legal formants and their comparative importance varies enormously from one system to another. For example, in some areas of English law, statutes are wholly lacking” (Sacco 1991a: 32). So, what can constitute a legal formant? Constitutions, statutes, case law, and scholarship are legal formants. They are so as texts, as well as work of legal actors, for they include the social practices of each particular group of interpreters. Even the way law is explained to law students in the classroom is a legal formant as teaching concurs to the modelling of the legal system. Interpretation is itself a legal formant: “Whatever influences interpretation is a source of law. To discover what influences interpretation various methods can be used. For example, one can examine the sources that an interpreter uses, be he lawyer or judge, when he advocates or adopts an interpretation of a rule” (Sacco 1991b: 345).

In Sacco’s understanding, “the reasons and the conclusions given by judges and scholars” (1991a: 30) are also legal formants. “Strange as it may sound, the reasons that judges and scholars give are different “legal formants” than their conclusions. The reasons have a life of their own, independent from the conclusions they supposedly support” (Sacco 1991a: 30).

“The propositions about the law that are put forward as conclusions by scholars, legislators, or judges are another legal formant” (Sacco 1991a: 31).

Reasons and propositions about the law are what Sacco considers “declamatory statements,” whereas “operational rule” designates the legal

rule, as it is applied. Sometimes “declaratory statements . . . make explicit an ideology, be it the ideology that actually inspire[s] the system in question or the one that a given authority believes to have inspired it or the one this authority wishes people to think inspired it. In civil law and in common law countries, declaratory statements are often made in accordance with the background of *jus naturalism*. Declaratory statements, for example, may insist that contracts are made by consent, while the operational rule requires not only consent but a reason or cause for the enforcement of the contract. . . Or, in common law countries, there are declaratory statements that property is transferred by the will of the parties while the operational rules require an additional element: consideration or delivery or, for the transfer of immovable property, a conveyance” (Sacco 1991a: 31).

National jurists assume that all legal formants in the same legal system have an identical content, that’s why the task for the national jurist is the quest for “the legal rule” within the system. Sacco’s theory, however, shows that the legal formants, within a given legal system, are never in complete harmony. “Comparison recognizes that the “legal formants” within a system are not always uniform and therefore contradiction is possible. The principle of non-contradiction, the fetish of municipal lawyers, loses all value in an historical perspective, and the comparative perspective is historical par excellence. Despite the search of unity and certainty of law, the comparative law scholar is able to show that the different legal formants are not in harmony, but rather in conflict” (Sacco 1991a: 24). An example comes from the interpretation of code provisions that are adopted with identical wording in two different legal systems. It may happen that, despite the exact wording of the statutory provisions, the two judges interpret the provisions in different ways and therefore apply different rules: the statutes alone do not determine the rules followed by the judges.

The dynamic approach, through the analysis of multiple legal formants, is opposed to the static approach based on dogmatic. The static analysis of a legal system looks for the “the rule” that must exist and, if not evident, can be deduced logically

from the official sources of law. The dynamic approach allows the observer to understand the whole picture, to take into account the different forces that are at work in shaping legal systems and to acknowledge the work that all the legal actors provide. The dynamic approach focuses on law as a social activity and refutes those “scientific” methods of legal reasoning that do not measure themselves against practice, but formulate definitions that are supported solely by their consistency with other definitions” (Sacco 1991a: 25).

The dynamic approach is at odds with positivism. A dynamic approach enables the observer to distinguish between the description that the legal system offers of itself (declaratory rule) and the rule that is in fact enforced (operative rule). A system, for instance, may grapple around the declaratory rule that affirms total and absolute “privity of contract” and then, in specific circumstances, allows compensation for interference with contract. If the set of facts that allows compensation are recurrent, we have an operative rule that is different from the declaratory rule: in specific circumstances, the contract is relevant for third parties that are obliged to pay damages in case of interference.

Sacco’s approach to law through legal formants is deemed a structuralist approach, as structuralism, in law, debunks the myth of the legal rule (Mattei 2001). Structuralism, as well as Sacco’s dynamic approach to comparative law, “show [s] that each rule is a complex structure composed of different “formants”: the rule formulated by the legislature, the rule as interpreted by courts, the implementation of the rule by administrative agencies, the discussion of the rule by law professors, etc. It also makes plain that each rule consists of both an operative prescription and a justification for that prescription, thereby emphasizing the role played by ideology and rhetoric” (Di Robilant 2016: 1326–1327).

Tacit, or Implicit, Legal Formant: The Cryptotype

Beside explicit legal formants, such as statutes, case law and scholarship, there are implicit legal

formants that Sacco calls “cryptotypes.” The terminology is imported from linguistics (Legrand 1995: 953–955). When we speak, very few of us would be able to formulate the linguistic rule we follow when we say “three dark suits” and not “three suits dark.” According to Sacco, the same applies to law. It happens that jurists follow and apply rules not explicitly formulated or enforce rules they are not aware of. In our experience, the law has become a theoretical exercise, especially since we have, within the society, a particular class of professional jurists. But even in our times, social order and performance are always extant in the shaping of rules and legal systems and they always keep a different and practical dimension. Customs, the social practices, and rules of conduct that people follow within social groups, that ultimately guide the actions of the members of a community, are what Sacco defines as “cryptotypes.” Judges, lawyers, interpreters are equally bound to the implicit formants that, therefore, are very influential in the modelling of different legal systems.

It is not easy to identify cryptotypes. It is even harder for national jurists, as they involuntarily follow rules they are unaware of. “Normally, a jurist who belongs to a given system finds greater difficulty in freeing himself from the cryptotypes of his system” (Sacco 1991b: 387). This is why comparative lawyer is privileged in detecting cryptotypes, as she is an outsider who doesn’t share the same implicit rules. “Some cryptotypes are more specific, others more general. The more general they are, the harder they are to identify. In extreme cases they may form the conceptual framework for the whole system” (Sacco 1991b: 386).

Legal Formants and Factual Approach: The Common Core Project

The Legal formants approach to comparative law is part of the methodology applied by the large group of scholars that participate in the Common Core Project. The Project, established in Trento in 1993, by Ugo Mattei and Mauro Bussani, aims at “unearth the common core of the bulk of

European Private Law, that is, of what is already common, if anything, among the different legal systems of European Union Member States” (Bussani and Mattei 2002: 1). Two methodologies were drawn together by Bussani and Mattei for the Common Core Project: the factual approach to comparative law, conceived by Rudolf Schlesinger (Schlesinger 1968) and the dynamic approach to comparative law, elaborated by Sacco (Mattei 2001).

The first source of inspiration has been the Cornell seminars held by Rudolph Schlesinger in the 1960s. Schlesinger’s endeavor was to describe how the rules on formation of contracts were functioning in various legal systems (Sacco 1991a: 29–30). Instead of asking participants to describe their own legal systems, he preferred to draw a questionnaire based on specific-case scenarios. His project launched what is now known in comparative law as the “factual approach.” Schlesinger’s work provides an accurate comparative overview, because an inquiry on how different legal systems solve the same practical problems better describes the functioning of a legal system. This approach is known by now as a functional approach, meaning that legal rules are best described by their function (Graziadei 2003).

Together with the factual approach, Common Core Project was inspired by Sacco’s theory on legal formants. This theory allowed national reporters as well as editors in charge of drawing the final comparative overview to have the full picture of the dynamic that do operate within a given legal system. In the instructions given to participants about how to answer the questionnaires (Bussani and Mattei 2002: Annex transform in 1), we read: “On a first level (‘operative rules’), the national reporters are asked to indicate how the case would be solved according to case law, legislation, legal doctrine, custom and usage” (Hesselink and Cartwright, 3). They should indicate as well “whether these formants are concordant. . . from an internal point of view and from a diachronic point of view. On a second level (‘descriptive formants’), the reporter is to indicate the reasons why lawyers feel obliged to adopt the solutions described at the first of the inquiry. Finally, on a third level (‘metalegal formant’, or

cryptotypes), the reporters are invited to indicate any other element that might affect the solutions mentioned at level 1, such as policy considerations, economic factors, social context and values, and the structure of legal process” (Hesselink and Cartwright 2002: 3; cf. Cartwright and Hesselink 2002 for a description and critiques of Common Core Project).

Legal Formants in Critical Legal Theory

Critical legal thought praised the legal formants approach to the analysis of law. Duncan Kennedy in a Critique of Adjudication (1997: 92) states the importance of Sacco’s approach, as a formidable tool in the scholars’ hands. Legal formants, showing the dynamics within a legal system, refute the idea of the inevitability of deduction. “A comparative method can thus provide a check on the claim of jurists within a legal system that their method rests purely on logic and deduction” (Sacco 1991a: 24). There is no obliged logical deduction from a legal rule. Duncan Kennedy bases his praise for Sacco’s work on the description of how identical code provisions are applied differently. For Kennedy, this is the evidence of the possibility for the judge, the scholar, and more broader the interpreter, to choose among various possible application of the rule. The judge, the scholar, and the interpreter cannot hide behind the neutrality of the law, neither behind the logic of the law. Any application of the rule, within any legal system, is a specific choice for the interpreter. Legal formants, therefore, are extremely effective in conducting an internal critique, to debunk the neutrality of the legal system using the same devices that the various legal actors use.

At the same time, Duncan Kennedy (2012) urges the comparative law scholar to take into account ideology as a legal formant that operates in an implicit way, therefore a cryptotype. The mainstream comparative law has not been particularly focused in this direction. According to Kennedy, if the deductive process of the judge, the scholar, and the jurists community is not determined, constrained, or forced by legal necessity, it is clear that the community of the interpreters

are actually choosing the solution. How do they determine the rule that they are going to apply is the central question of Duncan Kennedy’s scholarship. Confronted with legal formants methodology, he is willing to unveil ideology (Du. Kennedy 1997: 157–179) as one of the components of legal systems, that implicitly influences decisions. Ideology is therefore one of the most effective examples of cryptotypes, because it influences jurists’ decisions in an unconscious way. Classic comparative law, though, never considered politics and ideology as part of the legal system, because classic comparative lawyers preferred to situate themselves on a scientific level, where law is a technical and autonomous subject that has a separate life from politics.

Comparative Law and Economics

In the 1990s, Ugo Mattei, a comparative law scholar who studied law and economics in the United States, proposed a new approach to legal research (Mattei 1997; cfr. Mattei and Cafaggi 1998; Caterina 2006). Comparative law and economics was “the new frontier” (Mattei and Cafaggi 1998: 346) of legal scholarship that could “enable scholars to tackle major challenges of globalization” (Mattei and Cafaggi 1998: 346). Comparative law, the science that addresses the study of “similarities and differences between legal systems, is conjugated with law and economics,” (Mattei and Cafaggi 1998: 346) the science that evaluates legal rules and institutions from the point of view of economic efficiency. The final goal of this new academic challenge was to assess which legal system was more efficient. “Comparative law and economics seeks to explain in precise terms the convergence of legal rules by using efficiency as a key metric” (Di Robilant 2016: 1387). The legal formants approach is paramount in this endeavor, as it takes into account all the forces at work in a legal system, and not only black letter law. In addition, a dynamic approach sheds light on how and in which measure the rules and the community of jurists are actually working, consciously or

unconsciously, to build a more efficient legal system.

Comparative law and economics, in the hope of the founder, could work both at the level of positive and normative analysis. At the positive level, comparative law and economics can provide useful insights for understanding the reasons of the differences and the peculiarities of each legal system vis à vis the more universalistic efficient model. At a normative level, the scholar can suggest policy changes that move any legal system towards efficiency every time that a distance from the universalistic model is not justified. Moreover, comparative law and economics could give important contributions to the pattern of legal change, especially through the analysis of competing legal models and their inclusion in the legal system through the various legal formants, including path dependency and implicit legal formants (cryptotypes). It is well known that in the globalized world, legal systems are subject to changes, sometimes imposed, sometimes voluntary. Comparative law and economics can provide an understanding on which legal formant takes the responsibility for the change.

The development of comparative law and economics proceeds at slow pace in the academic legal analysis for several factors. Ramello (Eisenberg and Ramello 2016a: 3) identifies the causes in the lack of a specific journal that gives “dignity” to the discipline, and in the European character of comparative law and economics, uninteresting to the USA, and therefore global, scientific environment. This said, we can add that very few scholars are equipped for the task: it requires a good preparation in both disciplines, both comparative law and economics, usually very distant one from the other; secondly, the methodology is not always clear, same for expected results. Ramello and Eisenberg are willing to revive this particular angle of the discipline, bringing more comparison and a more cross-boundaries viewpoint into law and economics. Their recent book (Eisenberg and Ramello) is definitely heading towards this direction.

It is still unclear if legal originalists approach to legal systems, that led the way to World Bank

“Doing Business” reports, falls into the broader category of “comparative law and economics” (Eisenberg and Ramello 2016b: 8–9; Michaels 2009b). Through this program, the World Bank established its own method of assessing economic efficiency of legal systems (Djankov 2016; Michaels 2009a, b). The reports provide “objective measures of business regulations and their enforcement across 190 economies” <https://nation.com.pk/29-Jan-2018/conducive-environment-for-business>.

The reports “gather and analyz[e] comprehensive quantitative data to compare business regulation environments across economies and over time” <https://nation.com.pk/29-Jan-2018/conducive-environment-for-business>.

The “Doing Business” reports “encourage economies to compete towards more efficient regulations; they offer measurable benchmarks for reform and serve as a resource for academics, journalists, private sector researchers and others interested in the business climate of each economy” <https://nation.com.pk/29-Jan-2018/conducive-environment-for-business>.

Cross-References

- [Economic Analysis of Law](#)
- [Legal Origin](#)
- [Legal Transplants](#)

References

- Bussani M, Mattei U (2002) The common Core of European common law. Kluwer Law International, The Hague
- Caterina R (2006) Comparative law and economics. In: Smits JM (ed) Elgar Encyclopedia of comparative law. Edward Elgar Publishing Ltd, Cheltenham, pp 161–171
- Czuy K, Hogarth M (2018) Circling the square: Indigenizing the dissertation. *Emerging Perspectives*, 2(2):1–19
- Di Robilant A (2016) A symposium on ran Hirschl’s comparative matters: the renaissance of comparative constitutional law: big questions comparative law. *Boston Univ Law Rev* 96(4):1325–1345
- Djankov S (2016) The doing business project: how it started: correspondence. *J Econ Perspect* 30(1): 247–248

- Eisenberg T, Ramello GB (2016a) Introduction. In: Eisenberg T, Ramello GB (eds) *Comparative law and economics*. Edward Elgar Publishing, Cheltenham
- Eisenberg T, Ramello GB (2016b) *Comparative law and economics*. Edward Elgar Publishing, Cheltenham
- Graziadei M (2003) The functionalist heritage. In: Legrand P, Munday R (eds) *Comparative legal studies: traditions and transitions*. Cambridge University Press, Cambridge, pp 100–130
- Hesselink M, Cartwright J (2002) Introduction. In: Hesselink M, Cartwright J (eds) *Precontractual liability in European private law*. Cambridge University Press, Cambridge
- Kennedy Du (1997) *A Critique of Adjudication fin de siècle*, Harvard University Press, Cambridge, Mass.
- Kennedy D (2012) Political ideology and comparative law. In: Bussani M, Mattei U (eds) *The Cambridge companion to comparative law*. Cambridge University Press, Cambridge, pp 35–56
- Legrand P (1995) Questions à Rodolfo Sacco. *Revue Internationale de Droit Comparé* 47:943–971
- Mattei U (1997) *Comparative law and economics*. University of Michigan Press, Ann Arbor
- Mattei U (2001) The comparative jurisprudence of Schelsinger and Sacco: a study on legal influence. In: Riles A (ed) *Rethinking the masters of comparative law*. Hart Publishing, Oxford, pp 238–257
- Mattei U, Cafaggi F (1998) *Comparative law and economics*. In: Newman PK (ed) *The new Palgrave dictionary of economics and the law*. Stockton Press, New York
- Michaels R (2009a) The second wave of comparative law and economics? In: 59 *University of Toronto law Journal*, pp 197–213
- Michaels R (2009b) Comparative law by numbers? Legal origins thesis, *Doing Business* reports, and the silence of traditional comparative law. *Am J Comp Law* 57:765–796
- Sacco R (1974) Les buts et les méthodes de la comparaison du droit. In: *Reports Nationaux Italiens au IX Congrès International de droit comparé*, Teheran, 1974. GiuffrèTeheran, Milano. pp 113–131
- Sacco R (1980) *Introduzione al diritto comparato*. G Giappichelli, Torino
- Sacco R (1991a) Legal formants: a dynamic approach to comparative law. *American Journal of Comparative Law* 39:1–34
- Sacco R (1991b) Legal formants: a dynamic approach to comparative law. *American Journal of Comparative Law* 39:343–401
- Schlesinger R (ed) (1968) *Formation of contracts. A study of common Core of legal systems*. 2 vols. Oceana Publications, New York/London

Websites

- World Bank Doing Business Project. www.doingbusiness.org. Accessed 3 Oct 2018
- The Common Core of European Private Law. www.common-core.org. Accessed 3 Oct 2018
- Captain Zafar Iqbal, The Nation, 2018. <https://nation.com.pk/29-Jan-2018/conducive-environment-for-business>

Legal Insurance

► Litigation and Legal Expenses Insurance

Legal Origin

Maria T. Brouwer
Amsterdam, The Netherlands

Abstract

Flexible common law is considered to be more conducive to external investment than more rigid civil law. However, both systems can either adapt to new situations or become less flexible over time. Legal systems function best, when there is sufficient room for revision and review.

Definition

The concept of legal origin stems from the law and economics literature. It analyzes the effects of different legal traditions on economic performance. Common and civil law systems have spread over the globe and impacted the organization of economic life in Western Europe and former colonies.

Law Systems

Common law is of English origin and prevails in the UK and former British colonies. Judges have a great degree of discretion in common law systems. Common law evolves gradually through decisions in specific disputes that are upheld by higher courts. It differs from civil law that depends on legal statutes and comprehensive codes. The legislator in civil law systems wants to provide rules for as many situations as possible. Judges can only use discretion, if the legislator either overlooked the problem or left it deliberately open for the courts (Buetter 2002, 29). History has shaped how courts take decisions. French

judges became bureaucrats employed by the state, while English judges gained considerable independence from the Crown in the Glorious Revolution of 1688 (Mahoney 2001). Common law is said to support market outcomes, while civil law takes directions from the state. Civil law codes originate in the French Code Napoleon of 1804. The Code Napoleon abolished feudal rights and offered equality under the law to everybody (but not to married women). The Napoleonic Code was imposed on conquered territories in Italy, Spain, the Rhineland, and the Low Countries. Many of these countries voluntarily adopted the Napoleonic Code after Napoleon's defeat. The Dutch Civil Code of 1838 and the German Civil Law of 1898 are offshoots of the Code Napoleon (Meijer and Meijer 2002). The German Civil Code differs somewhat from French civil law and constitutes a different class. Scandinavian law differs and constitutes a class of its own. Legal origin theory, therefore, divides countries in four groups: (1) common law countries, (2) French civil law countries, (3) German civil law countries, and (4) Scandinavian civil law countries. Former colonies adopted the legal code of their former rulers after gaining independence.

Legal Origin and Financial Development

The law and finance literature found that financial development is related to legal origin. Common law countries make more extensive use of external finance (LaPorta et al. 1998). Both shareholders and creditors are better protected in common than in French civil law countries. Cross-country research found that French civil law countries have smaller stock markets, less active public offering markets, and lower levels of bank credit as a percentage of GDP than common law countries (LaPorta et al. 1997). A predominance of debt over equity finance characterizes civil law countries. This applies with the greatest force to Germany and Japan (LaPorta et al. 1997). Debt finance is ill suited to cope with entry and exit of firms. Both Europe and Asia showed little turbulence in their populations of quoted firms due to the scarcity of new quotations. This differed from the USA, where new firms appeared in droves on

the NYSE and Nasdaq after 1980. Failing firms in German civil law countries could cause the collapse of lending banks. A greater reliance on equity in common law systems offers more opportunities for continuation in bankruptcy proceedings. More US bankruptcies resulted in reorganization and continuation than in continental Europe (Brouwer 2006). The distinction between civil and common law countries overlaps that between relationship and market-based financial systems (Rajan and Zingales 1998). Anglo-Saxon countries have market-based systems, while continental Europe and Asia feature relationship-based systems. Rajan and Zingales found that market-based systems grow faster. Anglo-Saxon finance is more hospitable to entrant firms and therefore better suited to finance innovation (Djankov et al. 2002). Relationship banking, by contrast, mainly funds tangible assets that can be used as collateral to loans. The higher risk associated with debt finance prompts financial intermediaries in continental Europe and Japan to bet on the same risks. All are rescued by the state, if disaster strikes (Rajan and Zingales 2003, 18).

Legal Origin and Economic Development

Common law seems better suited to further economic development than civil law. Former English colonies like the USA, Canada, and Australia illustrate this narrative. However, not all former English colonies in Africa and Asia show superior economic performance. Japan that adopted German law outstripped economic growth in many former English colonies. Moreover, Belgium, France, and Germany achieved rapid economic growth after 1945. The empirical evidence is, therefore, not completely in favor of common law countries. Common law countries grew more rapidly in the 1960–2000 period than French civil law countries. German civil law countries, however, grew fastest (LaPorta et al. 2008). High costs of litigation and judicial arbitrariness are mentioned as reasons for slow growth in common law countries. The relative good performance of France, Belgium, and the Netherlands is ascribed to the more flexible use

of French civil law by judges in those countries. Judges in Italy and former French colonies stuck more to the letter of the code (LaPorta et al. 2008). Jurisprudence was recognized as a source of law in German courts, which enhanced flexibility. The capacity of legal codes to adapt themselves to new circumstances might, therefore, be more significant in explaining economic development than origins. Common law systems are said to evolve toward efficient rules through sequential decisions by appellate courts that can reverse decisions. The trial and error character of common law systems promotes flexibility. Napoleon did not allow for judicial review by judges. The centralized judiciary he installed was deemed superior to the arbitrariness of local judges of the ancien regime. French judges were villains in constitutional development, while English judges were heroes (Mahoney 2001). The English king was not above the law in contrast to French civil law. But, France installed administrative courts to review rules and decisions made by the state.

The Debate

The legal origins hypotheses put forward by LaPorta et al. have sparked an intense debate between proponents and critics. Critics pointed at the good economic performance of civil law countries like Japan, Belgium, and Germany. Culture, politics, and history were more important than legal origin in their view (LaPorta et al. 2008). Intermittent variables instead of legal origin might explain the differences between the different law systems. Years of schooling are sharply higher in common law than in French legal origin countries. Some authors point to religion to explain differences between countries. Protestant England was more conducive to market-led growth than Catholic France. Politics are also mentioned as explanation for divergent economic performances. Social democracy took root in continental Europe but is absent in the US-A. Others point at the reversibility of financial markets over time. Latin American countries of French legal origin were more financially developed than the USA and the UK in 1913 (Rajan and Zingales 2003). LaPorta et al. refute these critiques in an

elaborate response. They argue that the distinction between French civil and common law elucidates a host of phenomena from financial development, to regulation and state ownership of banks and companies. The beneficial effects of common law systems derive from the smaller hand of the state in economic affairs than in civil law countries. Heavy government interference entails more corruption, a larger unofficial economy and higher unemployment (LaPorta et al. 2008).

Anglo-Saxon Finance Came to Europe

Common law countries leave more to markets, while civil law countries turn to the state to solve economic problems. However, recent developments have put several characteristics of common law countries into question. The spread of Anglo-Saxon finance over the globe in the past decades has largely eliminated the differences in financial development among western European countries. Bank loans to the private sector as a percentage of GDP in continental Europe were almost equal to that of the USA in 2000. Stock market capitalization of continental European countries increased sharply from 17% in 1980 till 60% of US and UK capitalization in 2000 (Rajan and Zingales 2003). However, financial catching up did not boost entrepreneurship in continental Europe. There were few technological start-ups in Europe due to a lack of early stage venture capital. However, Europe became familiar with Anglo-Saxon private equity and hedge funds. These vocal shareholders demanded changes in executive boards and the breakup of companies. Many corporations were acquired by private equity or were prompted by hedge funds to sell themselves out to increase shareholder value (Brouwer 2008, Chap. 7). Globalization of finance might spell the end of relationship banking and the spread of market-based transactions. But, it is questionable, whether this financial convergence will stimulate economic growth. Anglo-Saxon financial innovations were a main cause of the financial crisis of 2008/2009 that led to the near meltdown of the world financial system. Securities based on mortgages lost most of their value, when home prices dropped and foreclosures abounded. The US securities were sold to

domestic and foreign banks, whose balance sheets deteriorated, when the losses appeared. Banks tended to collapse, when trust evaporated and interbank lending came to a halt. States had to step in to prevent banks from failing in both the USA and Europe. Financial markets are only stable if investor opinions on projects and people differ. No securities would change hands if investors would all agree on the value of a security (Brouwer 2012, Chap. 4). Security prices take a dive when investors are all looking for the exit and boom when investors enter financial markets in herds. Superior performance of Anglo-Saxon finance rests on independent, uncorrelated investor decision-making. But, investors came to rely heavily on opinions of rating agencies. Herd-like behavior caused a worldwide financial crisis. Banks participated in derivatives trade that did not create value but distributed gains and losses among buyers and sellers of these securities. Moral hazard problems appeared, when losing banks needed to be rescued by central banks. Both financial markets and legal courts are methods to solve disputes and differences of opinion. Judges take decisions after hearing both parties. Markets bridge the difference between buyer and seller valuations of a security. Security prices change continuously on secondary markets reflecting changes of investor opinion. Court decisions can be reversed on appeal or review. Landmark decisions by higher courts confirm new interpretations of the law. Financial markets and courts stir development, if initial differences of opinion are resolved through discourse.

Conclusions

The relationship between legal tradition and economic performance is not time and place invariant but depends on the quality of discourse and decision-making in markets and judiciaries. Both common and civil law systems can function well, if they allow dissent and overcome differences of opinion through appeal procedures and judicial review. Common law systems are more conducive to discussion, but discourse stops if unanimous opinion prevails.

References

- Brouwer M (2006) Reorganization in US and European Bankruptcy law. *Eur J Law Econ* 22(1):5–20 (is original publication)
- Brouwer M (2008) Chapter 7: Executive pay and tenure; founding fathers, mercenaries and revolutionaries. In: Governance and innovation. Routledge Studies in Global Competition; Executive Pay and Tenure; Founding Fathers, Mercenaries and Revolutionaries, pp 120–140
- Brouwer M (2012) Ch. 4. Corporate finance and the theory of the firm. In: Organizations, individualism and economic theory. Routledge Frontiers of Political Economy; Corporate Finance and the Theory of the Firm, pp 64–83
- Buetter M (2002) Cross border insolvency under English and German law. *Oxford University Comparative Law Forum*
- Djankov S, LaPorta R, Lopez-de-Silanes F, Shleifer A (2002) The regulation of entry. *Q J Econ* 117(1):1–37
- LaPorta R, Lopez-de-Silanes F, Shleifer A, Vishny R (1997) Legal determinants of external finance. *J Financ* 52(3):1131–1150
- LaPorta R, Lopez-de-Silanes F, Shleifer A, Vishny R (1998) Law and finance. *J Polit Econ* 106(6):1113–1155
- LaPorta R, Lopez-de-Silanes F, Shleifer A (2008) The economic consequences of legal origins. *J Econ Lit* 2:285–332
- Mahoney PG (2001) The common law and economic growth; Hayek may be right. *J Leg Stud* 30(2):503–525
- Meijer G, Meijer S (2002) Influence of the code civil in the Netherlands. *Eur J Law Econ* 14:227–236
- Raghuram R, Zingales L (1998) Financial dependence and growth. *Am Econ Rev* 3:559–586
- Raghuram R, Zingales L (2003) The great reversals: the politics of financial development in the 20th century. *J Financ Econ* 69(1):5–50

Legal Positivism and Law and Economics

Régis Lanneau

Law School, CRDP, Université de Paris Nanterre, Nanterre, France

Abstract

Law and economics scholars are rarely addressing traditional legal theory. Nevertheless, legal theory could help to better assess the limits of law and economics and the limits of legal theory. Moreover, legal

positivism largely paved the way for the emergence of law and economics. This entry will stress the points of convergence and the complementarities between these two approaches.

Introduction

Law and economics scholars are rarely addressing traditional legal theory and are often ignorant about legal positivism (with some exception of course, like Richard Posner (2003) or Lewis Kornhauser (2006, 2010) and Krecké (2016)). Legal theorists are frequently considering law and economics as an exteriority with its relevance often dubious for their own work; being “impure” (in a Kelsenian meaning of the word), it is not considered as “legal” knowledge. From this point of view, it is possible to consider that the legal positivism legacy could explain the transatlantic divide in the reception of law and economics (even if through a deeper analysis, this reality is not a necessity, Lanneau (2016)) and why law and economics remains controversial in the European legal academia. Despite this apparent mutual ignorance, both approaches are deriving from the same concern and appear complementary at many levels. Reciprocally legal positivism could enrich traditional law and economics. Being in an encyclopedia of law and economics, this entry will focus mainly on the former enrichment.

Legal positivism is hard to define precisely, and some could say that there are as many positivismisms as positivists. First, because the definition would change depending on the way it is conceptualized: legal positivism is different if it is considered as a methodology (a way to inquire into legal phenomena using the methods of natural sciences), as a theory of law (an inquiry into an object called “law”), or as an ideology (the primacy of positive law over all other normative systems). Second, because in each of these apprehension different schools exist: normativism, analytical school, post-positivism (for more details, see Grzegorczyk et al. 1993). Despite all these differences, it is possible to consider that positivists agree on three main theses. First, they

consider that law is the product of human beings; it is not something immanent. For that reason, law is perceived as an instrument. Second, and except when positivism is approached as an ideology, it is trying to develop a scientific approach of law derived from the methods of natural sciences and thus rejecting the classical distinction between episteme and praxis. Third, they are emphasizing the distinction between what is and what ought to be considering that the proper function of a science of law is to inquire into the is, not the ought (the ought would then define doctrinal approaches).

Before analyzing some of the points of divergence which are also explaining the complementarity between the two approaches (and its possible fertility), it is required to highlight the points of convergence to stress the common grounds of these approaches.

Points of Convergence

Legal positivism and law and economics are sharing two important features despite an under-theorization of law and economics. First, at the methodological level, both approaches are trying to develop a scientific approach to law. Second, at a value level, both are rejecting an immanent conception of law through a clear separation of law and morals.

Developing a Scientific Approach to Law

In order to develop a scientific approach of law, it is first required to distinguish between two levels of discourse: the level of the object (the law) and the meta level of the science (the science of law). The purpose of the science is then to describe its object (“what and how the law is, not how it ought to be” (Kelsen 1967). For that, it is required to follow a “scientific method,” a method based on causes and consequences since most legal positivists are fascinated by the methods of natural and empirical sciences. Thus, it is rejecting propositions which cannot be said either true or false (ideally on empirical grounds), it is rejecting the idea that there is an essence of things (since this essence cannot be the subject of scientific testing),

it is rejecting all forms of reason that are not episteme, and it is clearly distinguishing between the object and the science regarding this object. For Kelsen, a true science of law should “eliminate from the object of this description everything that is not strictly law” (Kelsen 1967). Of course, this approach is more aspirational than practical.

For law and economics, law is also an object of inquiry and economics (as a methodology to approach this object) is supposed to lead to some positive knowledge. This knowledge is not about the law considered as an object (what is the law?) but of the law considered only through its function (what are the consequences of the law?). This positive knowledge could be considered in two different ways. First, since the proposition about law are derived from models, the knowledge is positive (as logically derived from hypotheses); if we are assuming that people are rational and that law is providing constraints, then some consequences could be expected. Second, law and economics is more and more trying to empirically test some of its propositions to identify an “objective” effect of some law.

The Separation of Law and Morals

The separability thesis (the separation of law and morals) does not say that there are no links between law and morals or that we should not draw any lines between them. Indeed, Hart remarked that “There are many different types of relation between law and morals,” but “there is nothing which can profitably be singled out for study as the relation between them” (Hart 1983; 1994, 185). There are no essential (Hart 1958, 601) or necessary connection between law and morals; law is not dependent on morality. Because of this feature, law is, for a positivist, not defined by its content but by its structure (a set of primary and secondary rules for Hart, a dynamical system of Kelsen).

This content-free conceptualization of law is obvious in law and economics. Indeed, since the law is merely a function, it does not have any required content. Moreover, abiding by a consequentialist approach of ethics (since economics is consequentialist) it is prone to criticize any deontological conception regarding law and regulation

(Kaplow and Shavell 2002). This separation is also obvious considering that most scholars are enjoying debunking the inefficiency of morality tendencies in positive regulations.

Points of Divergence and the Complementarity of Approaches

For legal positivist, law and economics is lacking a clear concept of law. For law and economics practitioners, legal positivists are unable to address the consequences of norms or to rationalize the content of norms. Both of these blind spots – which could easily be explained by the focus of each approach – should be considered when addressing the value of each approach which appear from a practical point of view as complementary.

The Problem of “Validity,” the Blind Spot of Law and Economics

Despite the fact that for one of the major proponents of law and economics the question of what is law “has little practical significance if, indeed, it is a meaningful question at all” (Posner 1987, 765), having a concept of law is essential for at least two reasons. First, in order to have a clear understanding of what law and economics is, it is crucial to define what the law is for law and economics – otherwise the risk is simply equating an economic analysis of law to an economic analysis of norms; which is too often the case because the concept of legality is out of the reach of economic analysis (Lanneau 2010). Second, if the specificities of the law are not taken into account by law and economics, its value rests on poor foundations. Besides, to state that the concept of law “has little practical significance” is something that needs to be proven (hence the question addressed). To abide by a predictive theory of law – what the judges will do – is insufficient because in order to understand what they are doing, it is necessary for the analysis to be “comprehensive” enough to at least understand (or often presuppose) the concept of norms and legal order, otherwise it would be impossible to think of norms as reasons for action.

The identification of “valid” law requires a concept of validity. The reason why legal positivism could then be interesting for law and economics is that this question is stressing the systematic dimension of law (law is a set of norms that are interlinked to some extent). Kelsen, for example, is stressing the fact that legal orders regulate their own creation and application: this is the dynamic aspect of law (Kelsen 1967). If law is seen as a dynamic structure, it implies that each norm is produced according to a “superior” norm (or meta-norm), or at least that the procedure to create a norm is regulated by a norm in order to earn its validity (hence the necessity of a Grundnorm at the top of the system). This feature is not without impacts on economic analysis of law. Indeed, economic analysis is quite often focusing on the consequences of a primary norm *ceteris paribus*. In that case, it restricts itself to an economic analysis of a norm in “isolation” without regards to the system in which it belongs. When this happens, it is necessary to qualify the result of such economic analyses. Furthermore, the concept of efficiency becomes harder to figure out when the dynamic structure is taken into account. If, for example, a legislative process is supposed to be efficient, it does not mean that it cannot produce inefficient primary norms. However, are these primary norms really inefficient since they are produced by an efficient structure? (Lanneau 2016).

Rationalizing the Content of Norms and Assessing Consequences, the Blind Spot of Legal Positivism

For legal positivist, law is the product of authorities which are determining the “goals” of these rules. Rationalizing the goals is considered as “impure” in the Kelsenian system because it entails the use of methodologies that are considered as external. Even more, an analysis of the consequences is required if a legal order is supposed to be purposeful. This situation “forces the question, do these legal rules achieve the objectives at which they aim, and would alternative rules do any better?” (Bix 2004); then law and economics could be used in all of its dimensions to inquire into norms, sub-systems of norms or the whole system of norms.

And Kelsen does not seem opposed to an inquiry into the reason for norms that do not necessarily appear in legal propositions; it is simply not the task that he is trying to achieve. He even recognizes that: “For the legal norm obliges the debtor not only and, perhaps, not so much in order to protect the creditor, but in order to maintain a certain economic system” (Kelsen 1967). It is then possible to suggest a rationale for norms or systems of norms, and economics is one of these rationales.

Conclusion

Understanding the proper domain of validity of both legal positivism and law and economics is a prerequisite for clear thinking. Moreover, understanding the blind spots of both approaches could help to enrich both and contribute to developing a true law and economics approach and not a mere economic analysis of law. This path remains to be fully taken but should, with time, help European academics to seize the potential of law and economics.

Cross-References

- Hayek, Friedrich August von
- Law and Economics, History of
- Rule of Law

References

- Bix B (2004) Jurisprudence: theory and context. Carolina Academic Press, Durham
- Grzegorczyk C, Michaud F, Troper M (1993) Le positivisme juridique. LGDJ, Paris
- Hart H (1958) Positivism and the separation of law and morals. *Harv Law Rev* 21(4):593–629
- Hart H (1983) Essays in jurisprudence and philosophy. Clarendon Press, Oxford
- Hart H (1994) The concept of law, 2nd edn. Clarendon Press, Oxford
- Kaplow L, Shavell S (2002) Fairness versus welfare, Harvard University Press, Cambridge, MA
- Kelsen H (1967 [1960]) The pure theory of law. University of California Press, Berkeley
- Kornhauser L (2006) The economic analysis of law. In: Stanford encyclopedia of philosophy. <http://plato.stanford.edu/entries/legal-econanalysis/#ConLaw>

- Kornhauser L (2010) *Fondements juridiques de l'analyse économique du droit*. Michel Houdiard Editeur, Paris
- Krecké E (2016) Posnerian economic analysis of law and Kelsenian legal positivism: how similar are they? *Hist Econ Ideas* 23(3):167–194
- Lanneau R (2010) *Les fondements épistémologiques du mouvement law and economics*. Librairie LGDJ, Paris
- Lanneau R (2016) To what extent did European legal theory pave the way for an economic analysis of law? *Insights from Kelsen, Hart and Del Vecchio*. *Hist Econ Ideas* 23(3):195–224
- Posner RA (1987) The decline of law as an autonomous discipline: 1962–1987. *Harv Law Rev* 100:761–780
- Posner RA (2003) *Law, pragmatism and democracy*. Harvard University Press, Cambridge, MA

borrowing is usually the driving factor in legal change, reasons for dissent, repeatedly manifested, among others, by Pierre Legrand, were rooted on the denial of the same possibility of transferring rules and laws from one legal order to another (Legrand 1996).

Regardless of the academic discourses on whether legal transplants are sustainable as a notion in the legal theory, they are common practice. Nevertheless, the degree to which new laws are stimulated by foreign examples can vary. A frequent and often justified criticism is that imported laws are not suited to a certain local context.

Legal borrowings and transplants simply occur, though they can be more or less successful and effective, more or less persistent.

Legal Protection Insurance

► Litigation and Legal Expenses Insurance

Legal Transplants

Gianmaria Ajani
Department of Law, University of Torino,
Torino, Italy

Definition

The term “Legal transplants” is commonly used to designate the dissemination of legal models from an exporting legal order to a receiving one.

In a wider perspective, reception, transplants, or borrowings may either refer to the process, or to the results of a project of legal reforms, which is in turn initiated by a plan of legal change based upon an imitation of laws, doctrines and theories, and judicial decisions, already in place in different legal orders.

Within the fabric of such terminology, the notion of legal transplants has been, for the last four decades, most central. This is also due to a successful and widely discussed book by the legal historian Alan Watson (1974), which is devoted to a specific set of borrowings within the realm of private law. While the success of Watson’s study emerged from the plain recognition that

Dynamic Approach to Comparative Law

Comparative lawyers have paid a particular attention to the phenomenon of legal transplants.

At first, the traditional approach envisioned a static mapping of the major legal systems (David 1985): following this thread, which was the custom in the first part of the last century, national legal orders were classified and combined within larger groups, on the basis of the so-called “styles of legal families,” or common traits (Zweigert and Kötz 1998). Such an approach, still influenced by strands of national positivism and by the recognition of enacted legislation as the center of the observation, certainly removed some of the emphasis previously placed on the aspect of intra-system dissemination of rules.

Where mixed systems, such as South Africa or Israel, were recognized, they were presented as exceptional to the monolithic coherence of the several state legal orders, based mainly on the historical development of national law.

A new approach to comparative law was inaugurated in a seminal study by Rodolfo Sacco (1972, 1991). Sacco, progressing from the contributions of Gino Gorla and Rudolf Schlesinger, shifted the focus from the static description of legal orders to a dynamic reading of the borrowings, which nurture them. If we consider Sacco’s starting point, comparative law presupposes the

existence of a plurality of legal rules and institutions. A comparativist is called to study them in order to establish the extent to which they are identical or different. The analysis of legal formants, that is the different formative elements of a legal rule, has the aim to discover how the “jurist concerned with the law within a single country examines all of these elements and then eliminates the complications that arise from their multiplicity to arrive at one working rule” (Sacco 1991, p. 22). A full understanding of legal formants and their interrelation allows us to ascertain the factors that affect the given solutions. This clarifies the weight that interpretative practices (grounded on scholarly writings, on legal debate aroused by previous judicial decision, etc.) have in molding the actual outcome of the rules.

It follows that circulation of norms, borrowings, and transplants, all that is considered to be part of a process of reception, is the circulation of formants. By recognizing fragmented circulation, comparative law challenges the positivistic idea that reduces legal borrowings to the dissemination of enactments. The fragmented nature of the process leads to very different situations: In one case, the rule contained in a doctrinal formant may be able to circulate and finally be embodied in a statute.

In other circumstances, the influence may remain at the level of homogeneous formants (i.e., case law to case law).

Legal Borrowings, Global Harmonization, and the Neutrality of the Law

Historically, the idea of law’s indifference toward specific social contents has been kept alive by the millennial success of Roman law.

Subsequently, colonization only served to confirm the fact that different societies may be governed by similar laws. A more recent reappraisal of the idea occurred in the case of transition from Sovietism to market, in the wide area formerly under communist rule.

The belief that law or, more precisely, some areas covered by private and business law can be

independent, or neutral, with respect to a given society is closely connected to the idea that law can have a predictable impact on economic performance. The idea is not new at all: in fact it is linked with the spread of modern rationalism, which secured the transition from natural law to rational law. Similarly, the process of searching for the best legislative practice beyond national borders in order to support the modernization of a national legal system is not a new strategy. In this respect, one may recall the famous polemic originated two centuries ago between two great German legal scholars, Anton F. Thibaut, on the one hand, suggesting that the German states had to imitate French “rational” codification of private law in order to support modernization, and Friedrich K. von Savigny, on the other hand, contrasting the proposal, on the basis of a temporary cultural inadequacy of German legal scholarship.

During the last 20 years, after the waning of ideological confrontation, the resurfacing of a search for national identities based on culture and civilization has made the idea of neutrality of certain areas of the law somehow more difficult to defend. As a result, the recognition of what can be considered more neutral, or less dependent on society, has changed.

All throughout the period of the East-West confrontation, family law was deemed to be more indifferent (and therefore the comparison was inspired by a search of similarity), while economic and commercial law was taken as absolutely unfit for comparison. Today, in the age of globalization, many commentators would subscribe the statement that it is a smoother process to harmonize, for example, company law than the legislation addressing family law issues.

In order to draw closer to assessing the relevance of the principle of indifference toward legal change, one should begin by distinguishing between two possible ramifications: indifference as a political issue and indifference as a cultural issue.

As a political question, the principle of the law’s indifference to the social context has gained strength since the end of the cold war. While the idea of the uselessness of communist law “in

its entirety” (*tabula rasa* principle) was being affirmed by the governments and the international institutions supporting legal reforms in post-communist states, the neutrality principle characteristic of Western private law was put forward.

The adoption of an ideal model of market development by the neoliberal economists has significantly influenced the actions of the European Union and the international financial institutions, such as the World Bank and the International Monetary Fund, both seeking to speed up the process of market integration. Yet, this brought the focus of the discussion away from the institutional perspective: The “best” legal transplant appears to be a denationalized one, clothed in an appearance of objectivity that thwarts local resistances based on cultural identity. This separation of the action of the law from the actual role played by institutions is central to an understanding of the many failures of legal reform agendas.

It is, in fact, rather difficult to conceive a role for a legal transplant without considering how local institutions work. For any single new rule that is introduced, role-occupants will necessarily consider, more or less consciously, a set of pre-existing conditions that are country-specific and often non- or sub-formalized.

As a cultural question, the idea of indifference is indebted with a couple of phenomena that have significantly marked the last decades of the twentieth century:

- *Postmodernism*: There is something paradoxical in the fact that by insisting on the incommensurability of phenomena such as culture and social behaviors, postmodernist theories, which were perceived as a defense of localism against the forceful nature of globalization policies, by making relative the belief in justice and by criticizing the principle of objectivity in the law, have contributed to paving the way to the spreading of globalized formal regulations.
- *Uniformization of the law*: The success (both in terms of absolute numbers of initiatives, and of the number of adhering states) of the recourse to international conventions has brought the style of international law into the process of

legal transplant. This has important effects on the issue of neutrality, as international law was traditionally characterized by indifference toward domestic diversity and cultural differentiations: the traditional approach to the cultural issue practiced by international lawyers stresses the recognition of “similarities in economic development” (in such a sense, for instance, Thailand and Kenya are similar, in spite of evident cultural difference).

In other words, the technical side of neutrality is expressed by the language of performances and functions, while the cultural difference calls for an explanation in terms of history. International law deals with the former, while the latter is left to the realm of local policies.

This approach, however, has been subjected to a set of critiques (Kennedy 2003).

A first argument points to the fact that it is almost impossible to discern local from foreign cultures; even within a short span of time what was an alien culture may become local, as illustrated by the important cases of languages and arts, during every historical era and everywhere in the world.

Second, the “cultural exception” is of relevance to the context of economic issues. In particular, there are significant cases where the State actors use cultural issues to protect what are perceived to be national economic interests.

Finally, legal culture is encompassed by the more general definition of culture; the isolation of local legal culture from the process of legal transplants verges on the absurd. As the comparative law analysis has demonstrated, the dissemination of rules and legal change occur with the active involvement of legal doctrines. This is usually classified as interpretation, but it would be more accurate to speak of “interdoctrinal legal transplants.”

Transplant of Vague Notions

In contrast to what was common practice until the middle of the last century, when the process of legal reform via legal transplants occurred mainly

from state to state, at a slow pace, and generally through the vehicle of legal scholars, an important portion of the dissemination today is entrusted to supranational institutions, such as the International Monetary Fund, the World Bank, and the European Bank for Reconstruction and Development. These institutions support intergovernmental agreements aimed at fostering legal reforms and introduce a matrix of soft laws, proposals, and recommendations for legal change.

In the past, a new statute, a code, or a constitution was adopted or emended on the assumption that the “prestige” of the imported set of rules was sufficient to legitimize the change.

Today, the process of legitimizing the legal reform cannot be built upon a simple and circular argument such as the reputation of the foreign model.

Within this new situation, we notice that the process of legal transplantation is characterized by the recourse to vague formulas, such as *due process*, *governance* (and *good governance*), *reasonableness*, *rule of law*, *transparency*, and *accountability*.

We can identify three factors that have contributed to this shift in the building of legitimacy for the phenomenon of legal transplantation:

- Functionalism
- Cultural resistance
- Recourse to “naïf language”

As far as functionalism is concerned, one may consider replacing evaluations from the side of the borrower, based upon historical ties and cultural appreciation, with assessment built on the assumed utility of a set of norms targeted toward improving economic performance.

Cultural resistance concerns the increased awareness, from both sides (importer and exporter of legal rules), that local diversity is (or can be) an obstacle to an open transfer of rules among different legal orders.

When mentioning “naïf” (i.e., nontechnical, ordinary) language, one may recall the use of general terms and concepts that characterize law-making as practiced by some institutions. As an illustration, we can consider the case of the

European Union: If we examine the style of EU laws we cannot fail to notice that, while drafting directives and regulations, the EU institutions do not feel the obligation to adhere to the system of concepts and doctrines practiced at state level by local jurists.

This approach has led, in the last 15 years, to a proliferation of the use of vague notions and to their inclusion within the rhetoric of “good governance for the market.” Furthermore, because broad formulas expressed in ordinary language are more palatable to the media, a bundle of concepts whose juridical/technical content is far from certain became the paradigm for assessing the modernization of a legal system.

Let’s consider some examples: What is, for instance, the actual meaning of the slogan: “Russian law today is based on the US model of corporate governance”?

It is easy to grasp the symbolic meaning of the statement: Those involved in reforming Russian law indicate that the Russian commercial law has accomplished its renewal from plan to market. By proclaiming the adoption of the US pattern for corporate governance (whatever its contents can be), it is also implied that the economic system is able to become as sophisticated as the ones where corporate governance based on US law is in operation.

Such a symbolic significance, however, does not shed a lot of light upon the casual links between the nature of the market and the contents of the rules: Is it from a particular market that the corporate governance has come into effect, or, the other way round, is it the particular organization of the market, with its constraints and freedoms, that originated such a model for corporate governance?

Another relevant example concerns the notion of the “Rule of Law.”

The lip service paid to the rule of law idea by many constitutions adopted, for example, in the Russian Federation or in the People’s Republic of China during the last 20 years, acts as a legitimization for legal and judicial reform and for intrusions inspired by those who formulate, know, and possess the taxonomy of reforms currently attached to the vague notion under scrutiny.

The insistence in reproducing the rule of law blueprint within constitutional texts also demonstrates that the professionals who organize legal reforms in the new post-Soviet Countries “are part of the West,” and share the so-called Shihata’s doctrine (named after the former General Counsel of the World Bank). The doctrine is based on a global generalization of Max Weber’s assessment of the role of rules and institutions for economic development.

Following such a design, the process of transplanting new laws is conducted by following two different steps: Bringing the rule of law and other related and vague notions to the attention of reformers; and conveying operational rules as necessary consequences that follow the adoption of rule of law as a “constitutional standard.” The process is manifest in the pronouncements made by major financial institutions, as well as by other institutions, such as the European Union, or the World Trade Organization.

Its recognition, however, should not lead us to the immediate identification of that two-step process as a sort of machination; using vague notions in order to ease the bypassing of cultural sovereignty could, in fact, be part of a plan, aimed to an intense homogenization of the rules that are conceived by the International Financial Institutions as conducive to liberalization of trade, but this is not the only possible explanation of the success of the rule of law notion.

Besides the desire to avoid negotiations in the process of legal transplant, other reasons can be spent, such as the spreading, among the agencies exporting legal rules, of a practice of legitimization for their action based on the recourse to broad and consensus-making formulas. And if we share the élitarian explanation for the success of legal transplants suggested by some comparativists (Watson), we will assume that a recourse to vague notions is welcomed by the lawyers in the recipient country, who share the interest of the exporters in overcoming local resistances against transplants.

Whatever the case is, strategy or serendipity, it is very reasonable to assign to the mentioned vague notions the role of reducing transaction costs arising from the process of legal transplant.

Legal Transplants and International Law

The transnational institutions’ officials that are active in supporting legal transplants have inherited an obsession with the establishment of worldwide recognized norms from international lawyers. To attain the objective of a global recognition of international norms in a world where cultural diversity claims are reaching their climax, there are only two, rather conflicting, paths. The first entails recourse to hyper-specialized language, which is assumed to be, similar to the language of technicians and engineers, neutral and unrelated to local considerations. The alternative lies with the use of ordinary language, which does not dabble with the jargon of local experts.

The rule of law has not only become a substitute for discussion on the effects of development policies, but also a means to avoid, at least during the initial stages of dialogue between donors and recipient countries, an embarrassing confrontation on the “operational meaning” of widely debatable political notions such as democracy, pluripartitism, and judicial independence. Such an opportunity has proved to be useful not only for the international financial institutions seeking to sidestep problems with their mandate, but also for the EU institutions, involved in the extremely delicate issue of outlining the conditions for the Union enlargement toward East.

It is very clear, at this point, that the problem of providing exact definitions has been simply postponed by the urgency of having to adopt legal reform programs without too much discussion about contents or possible effects on the concerned domestic legal and economic system. Vague notions are useful as long as they respond to the persistent problems of globalization: scarcity of time and lack of consent. If this is the context, the strategy of legal transplants and reception requires that the rhetoric of the vague notions be characterized by a neutral style. Normative arguments follow at a later stage, once the sovereignty barrier has been circumvented: The process of transplanting rules assumed to be necessary for a good development of the market will be presented as a natural derivation from the vague notion itself.

What lies at the core of the practice of legal transplants is the contamination by the “style” of negotiation that characterizes both international conventions and supranational legislation. Indeterminacy of language allows the reaching of consensus (when required) and obtaining agreements within a reasonable span of time. In other words, a vague language facilitates the following:

- Postponing the tackling of obscure issues and puzzles related to implementation
- Avoiding a dive into details of distributional nature, and choices connected with introduction of good governance standards
- Putting off the resolution of arguments deriving from translation
- Empowering external lawmakers, who ground their legitimacy on a set of previously adopted vague formulas
- Keeping the governance of macroeconomic decisions under the control of “external standards provider” (e.g., the EU, during the process of enlargement toward Central and Eastern Europe)
- Producing an output ostensible to the donors (codes, laws, conventions, and the like)

These facilitating factors support an uncontested transplant of laws that are considered necessary for market functioning and economic development.

At the same time, however, they produce a false image of law as a static, neutral set of rules of universal application, under the unique condition that a commitment to the rule of law is expressed by the governments.

Conclusion

Both comparative analysis and functionalist methodology tell us that different rules may lead, through interpretation, to analogous results. The actual insistence on extensive harmonization is originated by a deficit of confidence in the capability or in the willingness of the local governments to set in place the enactments that are considered to ease economic development.

A reduction of the emphasis on the absolute necessity to imitate the formal contents of the laws, and a more convinced recognition of the role played by institutions in modeling the rules: These two actions, when taken together, can reduce the friction that characterizes the process of legal transplants.

In doing this, an important contribution may derive from institutional law & economics: It is indeed essential that, within the process which prepares legal change, every single macroeconomic recipe, for instance, the recognition of freedom of competition as an element of good economic performance, be separated from the several possible legal designs that it can assume. This should be based on the recognition that, historically, a successful transition from stagnation or underdevelopment to economic development was originated by a peculiar blend of mainstream economic theories and local regulations.

Cross-References

- [Law and Economics, History of](#)
- [Legal Formants](#)
- [Rule of Law](#)
- [Transition Economies: Rule of Law and Credible Commitment](#)

References

- David R (1985) *Major legal Systems in the World Today: an introduction to the comparative study of law*, 3rd edn. Stevens, London
- Eisenberg T, Ramello GB (eds) (2016) *Comparative law and economics*. Elgar, Northampton
- Gorla G (1983) *Diritto comparato e diritto comune europeo*. Giuffrè, Milan
- Kennedy D (2003) The politics and methods of comparative law. In: Legrand P, Munday R (eds) *Comparative legal studies: traditions and transitions*. Cambridge University Press, Cambridge, pp 345–433
- Legrand P (1996) European legal systems are not converging. *Int Comp Law Q* 45:52–81
- Sacco R (1972) *Introduzione al diritto comparato*. Giappichelli, Torino
- Sacco R (1991) Legal formants. A dynamic approach to comparative law. *Am J Comp Law* 39(Part I): 1–34, (Part II): 343–401

- Schlesinger R et al (1988) *Comparative law: cases, text, materials*, 6th edn. Foundation Press, New York
- Watson A (1974) *Legal transplants: an approach to comparative law*. Scottish Academic Press, Edinburgh/London
- Zweigert K, Kötz H (1998) *Introduction to comparative law*. Oxford University Press, Oxford/New York

Further Reading

- Ajani G (2007) Legal change and institutional reforms. In: *Ret tog tolerance. Festkrift til Helge Johan Thue*. Glyndendal Norsk Forlag, Oslo, pp 473–497
- Berkowitz D, Pistor K, Richard JF (2003) The transplant effect. *Am J Comp Law* 51:163–187
- Glenn P (2000) *Legal traditions in the world today*. Oxford University Press, Oxford
- Graziadei M (1999) Comparative law, legal history, and the holistic approach to legal cultures. *Z Eur Privatrecht* 7:531–543
- Mattei U (1994) Efficiency in legal transplants: an essay in comparative law and economics. *Int Rev Law Econ* 14:3–19
- Riles A (ed) (2001) *Rethinking the masters of comparative law*. Hart Publishing, Oxford/Portland
- Trubek D, Santos A (eds) (2006) *The new law and economic development*. Cambridge University Press, New York

Legislative Television

Franklin G. Mixon Jr.
Center for Economic Education, Columbus State University, Columbus, GA, USA

Abstract

Seminal studies describe legislative television as an advancement in political information technology that is used by incumbent legislators for protection against political challengers. Today, legislative television spans much of the globe, creating both intra- and international differences that have made the study of the impact of legislative television on the political process something akin to a natural experiment. This entry discusses some of the studies that make up this branch of the law and economics literature, with particular focus on how the presence of television in a legislature impacts turnover rates, session lengths and the popularity of various parliamentary procedures.

Definition

Legislative Television refers to the televising, either in real time or on a tape-delayed basis, of a government's (local, regional, or national) legislative process.

Televising Legislatures in the U.S. and Beyond

In a study whose primary focus is *not* the televising of legislatures or politics and politicians, Cowen (2000, p. 51) cuts to the heart of the subject, stating:

Successful politicians must use television and compete with popular culture for audience attention. Leaders therefore court voters by entertaining them and making them feel good. This strategy may win popularity and increase votes. . .

It is the general theme expressed above by Cowen that seminal studies on legislative television, either from a US or international perspective, have used as a theoretical foundation for modeling the behavior of politicians in the world where their daily activities are televised for thousands, or even millions, of constituents. The seminal studies by Crain and Goff (1986, 1988), with supplemental work by Greene (1991), describe legislative television as an advancement in political information technology – one used by incumbent legislators for protection against political challengers – thus placing the subject in the broader law and economics subcategory known as the economics of information (Stigler 1961; Nelson 1970, 1974, 1976; Telser 1976).

According to the National Conference of State Legislatures (ncls.org), 32 of the 50 states in the USA now offer televised coverage of state-level legislative sessions. In terms of national governments, Canada and Mexico join the USA, where the Cable-Satellite Public Affairs Network (i.e., C-SPAN, C-SPAN2, and C-SPAN3) operates, and most of Europe in offering televised broadcasts of legislatures. These continents are, however, different from Asia and South America, where relatively few countries now provide legislative television. Thus, from an academic perspective,

these intra- and international differences have made the study of the impact of legislative television on the political process something akin to a natural experiment and, therefore, fertile ground for modern scholars of law and economics.

The comprehensive work of Crain and Goff (1988) begins with an empirical analysis of the impact of legislative television at the state level (in the USA). They find that in smaller, less diverse (in terms of population) states, wherein legislative television is theoretically useful in protecting incumbents, the presence of television cameras serves to reduce the number of winning challengers in state house elections. This result carries to the federal level, given Crain and Goff's (1988) finding that legislative television coverage, in this case C-SPAN coverage, served, in more homogeneous districts, to increase the typical US House of Representatives incumbent's vote share. Mixon and Upadhyaya (2003) carry this theme forward in finding that the presence of C-SPAN cameras in the US House has, throughout its history, reduced turnover in that legislative body. This result not only confirms the incumbency-protection hypothesis in Crain and Goff (1988), but it also supports Mixon and Upadhyaya's (2002) earlier finding that the presence of C-SPAN2 cameras in the US Senate has, throughout its history, reduced turnover in that legislative body. These results complement the reverse causality found in Crain and Goff (1986, 1988) and Tyrone et al. (2003), wherein incumbents, seeking electoral protection from challengers, choose to adopt the televising of their legislative sessions.

Digging further into how incumbents are able to use legislative television to their advantage, Mixon (2002) and Mixon et al. (2003) find that US Senate filibustering is, in the presence of C-SPAN2 cameras, more effective as a form of low-cost political advertising for incumbents and, therefore, occurs with greater frequency. Similarly, Mixon and Upadhyaya (2003) find that one-minute speeches in the US House of Representatives increased in value and frequency in the presence of C-SPAN cameras. Given these results, it is not surprising that studies indicate that the presence of television cameras lengthens

legislative sessions, which, in the case of the US Senate, has been estimated to be an additional 63 min per recorded vote (Mixon et al. 2001; Mixon and Upadhyaya 2003). As Kimenyi and Tollison (1995) show, longer legislative sessions, such as those occurring in the presence of legislative television, come with a secondary effect – more complex legislation and greater government spending.

Where does this research program go from here? The short answer is *outside of the USA*. In one of the first published studies in this category, Soroka et al. (2015) find that televising the Canadian House of Commons has not had any discernible effect on politicians' behavior. This result confirms those in unpublished studies referenced in Soroka et al. (2015) indicating that the presence of cameras in the British House of Commons has had little, if any, effect, while the televising of Canada's House of Commons has impacted only the attire and within-session attention of the members of that legislative body. Clearly, there is much yet to learn about the law and economics of legislative television.

References

- Cowen T (2000) What price fame? Harvard University Press, Cambridge
- Crain WM, Goff BL (1986) Televising legislatures: an economic analysis. *J Law Econ* 29:405–421
- Crain WM, Goff BL (1988) Televised legislatures: political information technology and public choice. Kluwer Academic, Dordrecht
- Greene KV (1991) The nature of political services, legislative turnover and television. *Public Choice* 70:267–276
- Kimenyi MS, Tollison RD (1995) The length of legislative sessions and the growth of government. *Ration Soc* 7:151–155
- Mixon FG Jr (2002) Does legislative television alter the relationship between voters and politicians? *Ration Soc* 14:109–128
- Mixon FG Jr, Upadhyaya KP (2002) Legislative television as an institutional entry barrier: the impact of C-SPAN2 on turnover in the U.S. Senate, 1946–1998. *Public Choice* 112:433–448
- Mixon FG Jr, Upadhyaya KP (2003) Legislative television as political advertising: a public choice approach. iUniverse, Inc., New York
- Mixon FG Jr, Hobson DL, Upadhyaya KP (2001) Gavel-to-gavel congressional television coverage as political

- advertising: the impact of C-SPAN on legislative sessions. *Econ Inquiry* 39:351–364
- Mixon FG Jr, Gibson MT, Upadhyaya KP (2003) Has legislative television changed legislator behavior? C-SPAN2 and the frequency of Senate filibustering. *Public Choice* 115:139–162
- Nelson P (1970) Information and consumer behavior. *J Pol Econ* 77:311–329
- Nelson P (1974) Advertising and information. *J Pol Econ* 81:729–754
- Nelson P (1976) Political information. *J Law Econ* 19:315–336
- Soroka SN, Resko O, Albaugh Q (2015) Television in the legislature: the impact of cameras in the House of Commons. *Parliam Aff* 68:203–217
- Stigler GJ (1961) The economics of information. *J Pol Econ* 68:213–225
- Telser L (1976) Political information: comment. *J Law Econ* 19:337–340
- Tyrone JM, Mixon FG Jr, Treviño LJ, Minto TC (2003) Politics and the adoption of legislative television: an analysis of the U.S. House vote on C-SPAN. *Eur J Law Econ* 16:345–355

Lender of Last Resort

Uwe Vollmer

University of Leipzig, Leipzig, Germany

Definition

A lender of last resort (LoLR) is an institution, usually the central bank (CB) or a public deposit insurer (DI), which offers emergency financial assistance to commercial banks particularly during a financial crisis. In most cases, financial assistance means provision of liquidity to a single financial institution or to the financial market as a loan (liquidity assistance), usually against first-class collateral. Sometimes, the LoLR also recapitalizes commercial banks and provides risk-capital support that has not to be paid back (bank bailout).

Origin

The need for a CB to provide liquidity assistance was first mentioned by Henry Thornton (1802/

1939) and advanced by Walter Bagehot (1873) who elaborated the 9 principles applied to a policy of an LoLR. He proposed that the Bank of England should announce in advance its readiness to lend against collateral any amount to an illiquid but solvent financial institution at a penalty rate of interest. Bagehot suggested that during a financial crisis the central bank should lend freely at interest rates higher than precrisis levels to any sound borrower. Quality standards on collateral should be relaxed during crisis, but banks without good collateral were assumed to be insolvent and should be allowed to fail (Freixas et al. 2000a).

During the twentieth century authorities in many countries have acted as LoLR (Bordo 1990). Recent examples comprise the Swedish Riksbank and particularly the Bank of Japan which during the financial crisis of the 1990s provided substantial liquidity assistance to commercial banks; moreover, in Japan the Ministry of Finance injected almost 25 trillion yen of public funds into the banking industry (Nakaso 2001; Bebenroth et al. 2009). After the collapse of Lehman Brothers in September 2008, CBs of all major OECD countries conducted expansionary open market operations and provided liquidity assistance to single banks and other authorities conducted extensive recapitalization schemes (Petrovic and Tutsch 2010).

The need for a CB to provide liquidity assistance was first mentioned by Henry Thornton (*An enquiry into the nature and effects of the paper credit of Great Britain* (ed with an introduction by FAvon Hayek). George Allen and Unwin, London; 1802/1939) and advanced by Walter Bagehot (*Lombard street: a description of the money market*. Dodo Press, London; 1873) who elaborated the principles applied to a policy of an LoLR. He proposed that the Bank of England should announce in advance its readiness to lend against collateral any amount to an illiquid but solvent financial institution at a penalty rate of interest. Bagehot suggested that during a financial crisis the central bank should lend freely at interest rates higher than precrisis levels to any sound borrower. Quality standards on collateral should be relaxed during crisis, but banks without good collateral were assumed to be insolvent and should be

allowed to fail (Freixas et al. *J Financ Serv Res* 18(1):63–84, 15 [2000b](#)).

Bank Runs and Interbank Market Failures

The need for liquidity assistance by an LoLR results from the fact that commercial banks are fragile institutions that simultaneously offer very liquid deposits to their customers and invest funds received into profitable but illiquid projects. With this kind of liquidity creation, banks are subject to the possibility of a bank run, i.e., a situation where all depositors, even those without actually facing liquidity need, want to withdraw their deposits. Because deposits are subject to a “sequential service constraint” and are paid out in full on a “first come, first served” basis, depositors have to stand in front of the line to get their deposits repaid. A run may then result from a coordination failure among depositors, i.e., when sunspot events make depositors withdraw deposits because they believe that other depositors will do so (Diamond and Dybvig [1983](#)). Moreover, a bank run can also occur by changing fundamentals such as expectations that a bank’s capital cushion may be totally spent when assets devalue (Jacklin and Bhattacharya [1988](#)). Since the bank’s assets are not ready marketable and short-run rewards from “fire sales” are small, a run always results in an insolvency of an otherwise sound bank.

Interbank markets may shield individual banks against the risk of a bank run because they allow for a reallocation of liquidity among financial institutions. As long as individual bank’s liquidity needs cancel out, banks with liquidity surpluses may lend to other banks with a deficit, provided that they are considered creditworthy. If a run on a single bank, however, results in an aggregate liquidity shortage, an additional source of liquidity from outside the banking system is needed. If interbank markets function efficiently, the CB may provide additional liquidity through open market operations, leaving the allocation of liquidity to the interbank markets. This is, however, not sufficient during a financial crisis when interbank markets do not work smoothly, because

market participants perceive increases in counterparty risk or liquidity risk (Flannery [1996](#); Freixas et al. [2000a](#); Eisenschmidt and Taping [2009](#); Heider et al. [2009](#)). Then, the LoLR is forced to provide liquidity assistance to single institutions because the failure of an illiquid bank may have systemic consequences.

Systemic Crises

A run on a single bank may spread over the banking system if depositors consider the failure of another bank as a signal that their own bank is in trouble and withdraw their funds. If this occurs because depositors suspect similarities between banks, i.e., only depositors of banks with similar types of business withdraw while others do not, the contagion is called information based (Chari and Jagannathan [1988](#)). If the failure of the first bank leads to widespread run on banks without any assessment of similarities or dissimilarities, the situation is called pure panic (Kindleberger [1989](#)). The interbank market may also be a source of financial contagion when liquidity needs of one bank are sufficiently large to create an economy-wide shortage of liquidity, which then spills over to other banks (Allen and Gale [2000](#)). With extensive interbank exposures, the failure of one bank may have immediate effects on other institutions, especially if interbank lending is unsecured.

Systemic banking crises may be tremendous costly for society because they threaten the ability of the financial system to fulfill its basic functions and to provide liquidity. While this may justify LoLR liquidity support, it may also get into conflict with the CB’s ordinary monetary policy which is aimed at controlling interest rates or medium- or long-run monetary growth and inflation (Goodhart and Schoenmaker [1995](#)). Though emergency financial assistance is meant to satisfy extraordinary short-term increases in liquidity demand of depositors, CBs often have difficulties to exit from these measures. Especially if LoLR policies are extended on terms more favorable than are available in the interbank markets, banks may have an incentive to use the CB’s

standing facilities too extensively, making it difficult to keep the control over base money supply.

Solving this problem implies separating emergency liquidity assistance from monetary policy and to assess LoLR function not to the CB but to the public deposit insurer. Since the DI has to repay deposits in case of a bank failure, its incentives to provide financial assistance differ from those of the CB, and the optimal allocation of the LoLR functions between authorities depends on the size of the liquidity shock. The central bank should act as LoLR when a bank's liquidity needs are small, but the deposit insurer should be in charge when they are large (Repullo 2000). Kahn and Santos (2005) further argue for allocating supervisory power to the deposit insurer and identify conditions under which centralizing LoLR and deposit insurance functions is inefficient. In particular, when incentives to share information are weak, the allocation of regulatory power across government agencies should be contingent on their respective comparative advantages in gathering and utilizing relevant information.

Bank Bailouts

While Bagehot proposed only to support illiquid but solvent banks, the distinction between solvent and insolvent institutions is often difficult to make, especially in a short period of time (Goodhart 1999). Moreover, it may be less costly to recapitalize an insolvent bank than allow it to fail, and the failure of an insolvent bank may have systemic effects on the financial system. Therefore, an LoLR may also consider it worthwhile to recapitalize insolvent banks in a bank bailout. Such bank bailouts include instruments that either applies off the balance sheets of banks (like guarantees) or instruments that alter their balance sheets. The latter comprise a recapitalization through the asset side of the bank's balance sheet (purchases of nonperforming loans or "toxic assets" which are transferred to a "bad bank") or a recapitalization through the liability side of the balance sheet (introducing

either fresh risk capital or junior debt from the LoLR).

Though an ex post recapitalization of insolvent banks may be welfare improving, it creates – as any form of insurance – moral hazard on the part of commercial banks and sets ex ante incentives for banks to assume excessive risks. Guarantees give banks an incentive to take similar risks, thereby reducing asset diversification inside the financial sector and increasing its vulnerability against common shocks; they also create incentives to take higher asset risks and to accept a higher leverage (Chaney and Thakor 1985; Acharya 2009). Besides, guarantees influence the banks' competitors since they allow the beneficiary bank to take higher risks thereby forcing its competitors to accept higher risks, too (Gropp et al. 2010). Recapitalizations either through the asset or the liability side give banks an incentive to invest in unduly risky assets and to reduce reserves while depositors lose incentives to monitor the behavior and performance of their banks and to exert market discipline (Kaufman 1991; Saunders and Wilson 1996; Demirgüç-Kunt and Huizinga 2004).

Penalty Rates

To prevent these adverse effects, the LoLR may provide liquidity only at a penalty rate, i.e., at an interest rate higher than the market rate. Demanding such a penalty rate, however, may aggravate the bank's solvency problem. It may also send signals to market participants that the bank is in trouble; this makes banks reluctant to apply for funds because they fear to be singled out as a weak institute. Moreover, charging a penalty interest rate may even give an incentive to bank managers to "gamble for resurrection," i.e., to invest in projects with higher risks and higher returns in the hope of surviving and getting out of trouble (Rochet and Vives 2004; Repullo 2005; Castiglionesi and Wagner 2012).

"Constructive ambiguity" may be an alternative device to constrain moral hazard on the side of banks. It can be defined as a situation where the

LoLR ex ante creates uncertainty as to whether, when, and under what conditions financial support of an individual financial institution will be provided. If the LoLR keeps secret whether or not financial support will be granted, banks will not know individually whether they will be rescued or not; moreover, this might avoid “imitation effects” within the banking sector. If the LoLR is ambiguous about the conditions of financial assistance, it keeps a bank’s shareholders and management uncertain about the costs they have to bear in the case of financial assistance.

The beneficial effects of “constructive ambiguity” assume that following a mixed strategy and randomizing the unconditional rescue of banks dominates pure strategies where the CB either always liquidates or always bails out a distressed bank. If a bank failure has systemic consequences, the optimal strategy chosen by the LoLR depends on the size of the bank in trouble. The LoLR should set an upper limit for the size of the uninsured bank debt and never rescue banks that exceed that limit; constructive ambiguity should be applied to the rest of the banks (Freixas 2000). In a dynamic setting the effect on moral hazard is complemented by adding a value effect because rescuing a bank increases the current value of future bank profits and that increases the bank’s incentives to improve loan quality. Then, a mixed strategy is never beneficial, and the LoLR should always follow a pure strategy, i.e., rescue a bank if the value effect dominates and liquidate otherwise (Cordella and Yeyati 2003).

Private Sector Solutions

If “constructive ambiguity” is ineffective as a check on moral hazard, concerted private sector lending, organized by the CB, or private bank mergers and acquisitions, promoted by bank supervisors, could form alternatives. Orchestrated liquidity support operations may overcome the market’s coordination problems by encouraging a dialogue among banks or by imposing pressure on surplus banks to lend; this may be warranted if the CB has superior information about the banks’

solvency (Freixas et al. 2000a). Promoting takeovers of weaker institutions by solvent banks results in larger rents for the incumbent banks and creates additional incentives not to invest into gambling assets and to remain solvent (Perotti and Suarez 2002; Acharya and Yorulmazer 2008). Private sector solutions, however, are difficult to implement if the number of banks is large, market entry is easy, and the degree of competition in the financial sector is high. They are more probable to occur in financial systems with limited competition in banking, as during the 1970s in, e.g., Germany where after the failure of “Bankhaus Herstatt” liquidity assistance was organized as a private club (Beck 2002).

References

- Acharya VV (2009) A theory of systemic risk and design of prudential bank regulation. *J Financ Stab* 5(3):224–255
- Acharya VV, Yorulmazer T (2008) Cash-in-the-market pricing and optimal resolution of bank failures. *Rev Financ Stud* 21(6):2705–2742
- Allen F, Gale D (2000) Financial contagion. *J Polit Econ* 108(1):1–33
- Bagehot W (1873) *Lombard street: a description of the money market*. Dodo Press, London (reprint)
- Bebenroth R, Dietrich D, Vollmer U (2009) Bank regulation and supervision in bank-dominated financial systems: a comparison between Japan and Germany. *Eur J Law Econ* 27(2):177–209
- Beck T (2002) Deposit insurance as a private club: is Germany a model? *Q Rev Econ Finance* 42:701–719
- Bordo MD (1990) The lender of last resort: alternative views and historical experience. *Fed Reserve Bank Richmond Econ Rev.* 18–29 Jan/Feb
- Castiglionesi F, Wagner W (2012) Turning Bagehot on his head: lending at penalty rates when banks can become insolvent. *J Money Credit Bank* 44(1):201–219
- Chaney PK, Thakor AV (1985) Incentive effects of benevolent intervention: the case of government loan guarantees. *J Public Econ* 26(2):169–189
- Chari VV, Jagannathan R (1988) Banking panics, information, and rational expectations equilibrium. *J Financ* 43(3):749–761
- Cordella T, Yeyati EL (2003) Bank “bailouts”: moral hazard vs. value effect. *J Financ Intermed* 12:300–330
- Demirgüç-Kunt A, Huizinga H (2004) Market discipline and financial insurance. *J Monet Econ* 51(2):375–399
- Diamond D, Dybvig P (1983) Bank runs, deposit insurance, and liquidity. *J Polit Econ* 91:401–419

- Eisenschmidt J, Tapking J (2009) Liquidity risk premia in unsecured interbank money markets. European Central Bank working paper 1025
- Flannery M (1996) Financial crises, payment system problems, and discount window lending. *J Money Credit Bank* 28:804–824
- Freixas X (2000) Optimal bail out policy, conditionality and constructive ambiguity. Netherlands Central Bank, DNB staff reports 49, Amsterdam
- Freixas X, Giannini C, Hoggarth G, Soussa F (2000a) Lender of last resort: what have we learned since Bagehot? *J Financ Serv Res* 18(1):63–84
- Freixas X, Parigi BM, Rochet J-C (2000b) Systemic risk, interbank relations, and liquidity provision by the central bank. *J Money Credit Bank* 32(3):611–638
- Goodhart CAE (1999) Myths about the lender of last resort. *Int Financ* 2(3):339–360
- Goodhart C, Schoenmaker D (1995) Should the functions of monetary policy and banking supervision be separated? *Oxf Econ Pap* 47:539–560
- Gropp R, Hakenes H, Schnabel I (2010) Competition, risk-shifting, and public bail-out policies. *Rev Financ Stud* 25(8):2343–2380
- Heider F, Hoerova M, Holthausen C (2009) Liquidity hoarding and interbank market spreads: the role of counterparty risk. European Central Bank working paper 1126
- Jacklin C, Bhattacharya S (1988) Distinguishing panics and information-based bank runs: welfare and policy implications. *J Polit Econ* 96:568–592
- Kahn CM, Santos JAC (2005) Allocating bank regulatory powers: lender of last resort, deposit insurance and supervision. *Eur Econ Rev* 49:2107–2136
- Kaufman G (1991) Lender of last resort: a contemporary perspective. *J Financ Serv Res* 5(2):123–150
- Kindleberger C (1989) Manias, panics, and crashes: a history of financial crises. Basic Books, New York
- Nakaso H (2001) The financial crisis in Japan during the 1990s: how the Bank of Japan responded and the lessons learnt. BIS papers no 6, Basel
- Perotti E, Suarez J (2002) Last bank standing: what do I gain if you fail? *Eur Econ Rev* 46:1599–1622
- Petrovic A, Tutsch R (2010) National rescue measures in response to the current financial crisis. European Central Bank, legal working paper series no 8, Frankfurt/Main
- Repullo R (2000) Who should act as lender of last resort? An incomplete contracts model. *J Money Credit Bank* 32:580–605
- Repullo R (2005) Liquidity, risk taking, and the lender of last resort. *Int J Cent Bank* 2(2):47–80
- Rochet J-C, Vives X (2004) Coordination failures and the lender of last resort: was Bagehot right after all? *J Eur Econ Assoc* 2(6):1116–1147
- Saunders A, Wilson B (1996) Contagious bank runs: evidence from the 1929–1933 period. *J Financ Intermed* 5(4):409–423
- Thornton H (1802/1939) An enquiry into the nature and effects of the paper credit of Great Britain (ed with an introduction by FA von Hayek). George Allen and Unwin, London

Leniency Programs

Karine Brisset

Centre de Recherche sur les Stratégies Economiques, University of Franche-Comté, Besançon, France

Abstract

Article 101 of the Treaty on the Functioning of the European Union prohibits cartels and other antitrust agreements that reduce or eliminate competition between firms. Leniency programs are an important investigative tool which give cartel's members incentives to report their cartel activity and cooperate with competition authorities. We present their main objectives, their development in different countries, their direct and indirect effects and how these programs could be improved.

Synonyms

Amnesty Programs

Definition

A leniency program consists in canceling or reducing any fine or other sanctions against companies, engaged in an illegal cartel, when they report strong evidence of their activity to a Competition Authority.

Leniency program's objectives

Across the world, one of the main objectives of antitrust authorities concerns the fight against illegal cartels. For that purpose, they conceive new tools to prosecute them and to increase their effectiveness. For more than 30 years, leniency programs constitute a major innovation applied by many competition authorities to fight cartels. These programs were first applied in the USA in 1978: the Amnesty Program. Currently, following the success of the American system, different

leniency programs are in effect in most advanced countries (most of the European member states, e.g., Germany since 2000 and France since 2001, Japan, Australia, New Zealand, Canada, China, South Korea, etc) and also in some developing countries. More than 50 countries have adopted different leniency programs.

For competition authorities, the main objectives behind this tool are:

- To increase their effectiveness in order to uncover cartels by relying on testimony from cartel's firms (enhance cartel detection, reduce the cost of prosecution)
- To prevent or reduce cartel formation by making cartel less stable, providing a greater incentive to deviate from cartel agreement and to immediately cooperate with authorities (enhance cartel deterrence)

History

The first American Corporate Amnesty Program, implemented in 1978, granted amnesty to the first company that gave strong evidence of its cartel before any authority's investigation. This program was unsuccessful (one case by year) for two reasons. Firstly, the amnesty depended on the discretion of the attorney general and required full cooperation of the applicant. Secondly, before any investigation by an authority, the individual risk of conviction is low, and the first American program gave no sufficient incentives to be effective. Hence, in 1993, the US Antitrust Division of the Department of Justice decided to review the program. The current program also provides full amnesty to the first company revealing information even after the authority's investigation has already begun and until the authority has sufficient evidence to condemn the cartel. Since this reform, there are one or two leniency applications per month.

In 1996, the European Union has also implemented its own leniency program. However, after 5 years of implementation, results appeared to be mixed. There was no certainty that the first applicant would obtain complete immunity. In

fact, less than half immunity was granted. In addition, once the Commission had begun an investigation, it was only possible to obtain a maximal fine's reduction of 75%. That is why the policy was first improved in 2002 and in 2006, with the objective of a greater transparency and certainty.

Current conditions are as follows:

- The Commission cancels any fines to the first applicant that provides substantial evidence of the cartel, allowing the authority to implement an investigation pursuant to Article 14 (3) Regulation No. 17 or to prove an infringement of Article 81 of the Amsterdam Treaty.
- In addition, if another company provides further information, it may receive a fine reduction (first firm to cooperate, reduction of 30–50%; second firm to cooperate, 20–30% reduction; other firms to cooperate, reduction of up to 20%).

Direct and Indirect Effects

Theoretical analyses have shown that leniency programs can increase the incentive to cooperate with authorities and may have a deterrence effect on collusion. However, they may also have a pro-collusive effect (Motta and Polo 2003). Indeed, leniency programs can increase the benefits of a cartel's member as they reduce the expected cost of the antitrust sanction. Hence, it is possible that the number of cartels increases because of leniency. Different empirical studies seem to conclude that the positive effect on cartel destabilization and on cartel deterrence dominates the pro-collusive effect. However, empirical identification is only derived with data from detected cartels (Brenner 2009; Miller 2009). Therefore, if the success of a leniency program depends on the number of uncovered cartels, it may mean that there are more current cartels and not necessarily more efficient prosecution.

To reduce this pro-collusive effect and to increase the effectiveness of a leniency policy, it is necessary to have higher sanctions against cartel's members and to have a real power of investigation. Indeed, if fines are weak and if the

probability of an investigation is low, the expected sanction will be low. Then, cartel's members will have no incentive to report to the competition authority because the risk of a conviction is weak. That is why, since 2002, the European Commission has decided to increase the level of fines against cartel's members. The maximal fine was doubled and represents 10% of a company's global turnover. This policy has increased the incentives to apply for leniency.

In contrast with the European program, the US program only grants exclusive amnesty to the first confessor. This strict program may create a "preemption effect" and may increase the deterrence effect (Harrington 2013). Indeed, even if one member thinks that the competition authority has a low likelihood to undermine the cartel, he may also think that another member may report and apply for leniency. In these conditions, he can have a strong incentive to quickly cooperate with authorities to preempt other members and to benefit from the fine's reduction. With the European program, second and third firms revealing strong evidence receive reductions in sanctions as well, but lower than the first confessor. Hence, the "preemption effect" exists but is less powerful than the one in the American program. It is also important that the gap in the reduction of fines be sufficiently great to increase the incentive to be the first to cooperate.

Can a Ringleader Apply for Leniency?

In the current US leniency program, a firm may only be eligible for amnesty when it is not a "ringleader." This clearly means that it is not at the origin of the cartel and that it has not invited any other firm to take part in the cartel (organize meetings, communication between cartel members, etc.). The first European leniency program, between 1996 and 2002, applied discrimination to the ringleader, too, but it was less strict: a ringleader was not illegible for full amnesty but could only have smaller fine reduction between 10% and 50%. However, the current program, following two revisions in 2002 and in 2006, is more

accurate and allows any cartel's firm, even a leader in the cartel, to benefit from full immunity provided it is not the cartel's instigator (which organizes and initiates the activity).

What is the economic impact of excluding the ringleader? If a firm, as a ringleader, is excluded from amnesty, it increases his potential sanction's expected cost in comparison with other members. Then, this ringleader can ask the others for a monetary compensation. This can create an asymmetry between this firm and the others and may decrease the sustainability of the cartel. Indeed, the result depends on the probability to be reviewed by the competition authority (Bos and Wandschneider 2011). If this probability is rather low, including the ringleader in the program may increase the probability of a successful prosecution. In contrast, if this probability is rather high, it may be better to exclude the ringleader to reinforce asymmetry with others.

How to Improve Leniency Programs?

Since the first American leniency programs, many measures have been adopted to improve their efficiency.

Leniency Program for Multimarket Firms

Multimarket contacts can facilitate the sustainability of collusion (Bernheim and Whinston 1990). In practice, cartels may concern many markets. That is why in 1999, the USA has adopted **the Amnesty Plus program**. Its objective is to convince a firm convicted in a first market to report evidence on a collusive agreement in another market. Hence, this firm could obtain amnesty as a cartel member in the second market if it is the first one reporting this cartel and a fine's discount as member in the first cartel. Theoretical conclusions on the Amnesty Plus reform are not clear. Indeed, it has two opposite effects on companies' incentive to collude. It may reduce the deterrence effect by increasing the sustainability of multimarket collusion. However, it can reduce the cartel's expected duration by increasing the risk of report after the detection of the first cartel

(Lefouili and Roux 2012). If the program is rather lenient, it would seem that the procompetitive effect is stronger.

Leniency Program and Fight Against International Cartels

In practice, many cartels concern international markets and may be prosecuted by different jurisdictions. In the USA, 90% of recent fines for antitrust infringement result from international cartels. Information sharing among competitive authorities of different jurisdictions could increase fight against cartels, reduce the cost, and improve the effectiveness of an investigation (reducing time to obtain strategic information, improving exchange of strategic documents). However, information sharing between different authorities could reduce the effectiveness of leniency programs as well (Choi and Gerlach 2012). Indeed, confidentiality is essential and is a necessary condition to be credible and to protect the integrity of leniency applications.

A More Incentive System: Reward for Whistleblowers

In April 2005, the South Korean Fair Trade Commission has introduced a reward system to give more financial incentive for outsider whistleblowers to reveal information and provide evidence on secret cartels. Hence, in June 2005, an anonymous person who revealed a welding rod cartel, providing strong evidence, received a reward of \$63,700. In the USA, through the False Claims Act, there is a similar financial reward mechanism to fight fraud against the government, but not for violations of competition law. In Europe, no reward system for informers exists. Theoretical predictions suggest that rewarding cartel members for reporting is a powerful tool to deter cartel formation (Aubert et al. 2006; Buccirossi and Spagnolo 2006). It increases the risk of an individual deviation, bringing more incentive to report. It can also increase the cartel's expected cost because firms will have to "bribe" individuals to prevent them to report. However, a first experimental analysis concludes that rewarding whistleblowers does not further deter cartel

formation in comparison with traditional leniency (Apesteguia et al 2004). But a second one, assuming a dynamic setting, predicts that rewards increase cartel detection due to self-reporting (Bigoni et al. 2012).

Leniency Program and Administrative and Criminal Enforcement

In the USA and in South Korea, there are criminal sanctions for individuals engaged in cartel activities (custodial sentences). In the USA, since 1994, a criminal leniency has been implemented. Moreover, since 2004, an individual involved in a cartel can be imprisoned for up to 10 years. Most of European member states have the possibility to impose criminal sanctions on individuals for cartel conduct. For example, in France, individuals can face jail time for up to 4 years and may have a criminal fine for up to 35,000 euros. Moreover, some European states (UK, Ireland, and Austria) have both administrative and criminal leniency programs. In these countries, the competition authority will not prosecute an individual if he satisfies conditions for leniency. Other member states (Belgium, France, Germany, Slovenia, Estonia, Ireland, etc.) do not have any criminal leniency program. In practice, jail sentences are indeed rarely imposed. Moreover, currently, the law of the European Union does not provide for criminal sanctions, and each member state is free to impose its own system of penalties for antitrust infringements. In contrast, in the USA, there were many jail sanctions against chairmen of companies engaged in cartel activities. The idea is that jail sentences are a great disincentive for individuals to participate in a cartel. In Europe, some officials think that high administrative fines are sufficient to deter cartel conduct. Both mechanisms are indeed complementary. With an administrative leniency, the efficiency only depends on the incentive of cartel companies. With an individual criminal leniency, the impact depends on the objectives of an employee concerned by a criminal sanction, but it may also concern his company itself. Indeed, if the company anticipates the risk of report by its employee, it could have more incentives to apply for a corporate leniency.

Cross-References

- [Cartels and Collusion](#)
- [Criminal Sanctions and Deterrence](#)
- [Experimental Law and Economics](#)

References

- Apesteguia J, Dufwenberg M, Selten R (2004) Blowing the whistle. *Economic Theory* 31(1):143–166
- Aubert C, Kovacic W, Rey P (2006) The impact of leniency and whistleblowing program on cartels. *Int J Ind Organ* 24:1241–1266
- Bernheim BD, Whinston MD (1990) Multimarket contact and collusive behavior. *Rand J Econ* 21(1):1–26
- Bigoni M, Fridolfsson SO, Le Coq C, Spagnolo G (2012) Trust, salience and deterrence: evidence from an anti-trust experiment. *Rand J Econ* 43(2):368–390
- Bos I, Wandschneider F (2011) Cartel ringleaders and the corporate leniency program. Working paper, Maastricht Research School of Economics of Technology and Organization
- Brenner S (2009) An empirical study of the European corporate leniency program. *Int J Ind Organ* 27(6):639–645
- Buccirossi P, Spagnolo G (2006) Leniency policies and illegal transactions. *J Public Econ* 90(6–7):1281–1297, Elsevier
- Choi JP, Gerlach H (2012) Global cartels, leniency programs and international antitrust cooperation. *Int J Ind Organ* 30:528–540
- Harrington JE (2013) Corporate leniency programs when firms have private information: the push of prosecution and the pull of pre-emption. *J Ind Econ* 61(1):1–27
- Lefouili Y, Roux C (2012) Leniency programs for multi-market firms: the effect of Amnesty Plus on cartel formation. *Int J Ind Organ* 30(6):624–640
- Miller N (2009) Strategic leniency and cartel enforcement. *Am Econ Rev* 99(3):750–768
- Motta M, Polo M (2003) Leniency programs and cartel prosecution. *Int J Ind Organ* 21(3):347–379
- Leliefeld D, Motchenkova E (2010) Adverse effects of corporate leniency programs in view of industry asymmetry. *J Appl Econ Sci* 5(2(12)):114–128
- Motchenkova E, Van der Laan R (2011) Strictness of leniency programs and asymmetric punishment effect. *Int Rev Econ* 58:401–431
- Spagnolo G (2002) Leniency and whistle-blowers in anti-trust. CEPR Discussion Papers 5794, C.E.P.R. Discussion Papers

Lex Mercatoria

Bruce L. Benson

Department of Economics, Florida State University, Tallahassee, FL, USA

Abstract

Lex Mercatoria, or the Law Merchant, generally refers to the customary rules and procedures developed within merchant communities to support trade in medieval Europe, without the assistance of government, although this system has had many names through its evolution. This system began to develop sometime in the early Middle Ages, but became widely recognized as commerce expanded in the eleventh and twelfth centuries. Lex mercatoria lowered transactions costs, and provided incentives to live up to promises, thereby allowing for widespread use of contracting and credit as commerce expanded. The customary law system developed behavioral rules based on customs, practice and usage within the network of merchant communities, but processes to encourage recognition, provide adjudication and generate changes in the rules. The primary impetus for recognition of the law merchant within the commercial sector were the positive incentives associated with maintaining reputations and repeated dealings, along with the potential of reputation sanctions (e.g., spontaneous ostracism) for misbehavior. Arbitration was probably the primary dispute resolution process, although merchants used other courts (particularly fair courts sometimes referred to as piepowder courts, market courts, urban courts and ecclesiastical courts) when they

Further Reading

- Brisset K, Thomas L (2004) Leniency program: a new tool in competition policy to deter cartel activity in procurement auctions. *Eur J Law Econ* 17(1):5–19
- Hamaguchi Y, Kawagoe T, Shibata A (2009) Group size effects on cartel formation and the enforcement power of leniency programs. *Int J Ind Organ* 27(2):145–165
- Harrington JE (2008) Optimal corporate leniency programs. *J Ind Econ* 56(2):215–246
- Hinloopen J, Soetevent AR (2008) Laboratory evidence on the effectiveness of corporate leniency programs. *Rand J Econ* 39(2):607–616

were willing to base decisions on the law merchant. Change was generally initiated through bargaining and contracting, followed by emulation so that new rules spread. Various general principles of the law merchant became increasingly universal over time, although the system remained polycentric in many ways. One reason for polycentric characteristics is that authoritarian legal systems (e.g., royal law, urban law) strove to gain control over commerce and its regulation by imposing additional laws on merchants that often conflicted with the law merchant, forcing changes in merchant practice and usage. The success of such efforts varied over time and space.

Synonyms

Custom of merchants; Customary law of merchants; Law of the fair; Law merchant; Merchants' law; Method of merchants; Way of trade, The

Definition

Lex Mercatoria: The customary rules and procedures developed within merchant communities to support trade in medieval Europe, without the assistance of government.

Introduction to the Law Merchant

Ninth century documents suggest that merchants in parts of Europe used different legal procedures than local communities (Kadens 2004, 43). According to Notger of St. Gallen, writing around 1000 A.D., “merchants maintain that a sale made in a fair should be binding ... since it is their custom” (Volckart and Mangels 1999, 439). Greif (2006, 70) quotes numerous eleventh century documents illustrating “that the Maghribi traders ... employed a set of cultural rules of behavior – merchants’ law.” Such institutional arrangements continued to evolve as a system of customary law used by merchants and referred to as “merchants’ law,” “custom of merchants,”

“customary law of merchants,” “method of merchants,” “the way of the trade,” and “law of the fair,” but now regularly labeled *Lex Mercatoria* or the “Law Merchant” (*Lex Mercatoria* terminology also is applied to rules and institutions supporting modern international trade, but the focus here is on the medieval system). A large literature describing and analyzing medieval *Lex Mercatoria* exists [for example, *The Little Red Book of Bristol* written sometime around 1280 (Bickley 1900; Teeter 1962; Basile 1998); Stracca (1553); Malynes (1622); Marius (1651); Zouche (1663); Mitchell (1904); Bewes (1923); Trakman (1983); Berman (1983); Benson (1989, 1999, 2011, 2014a, 2014b), and Milgrom et al. (1990)], although there also are critics of the medieval Law Merchant story [e.g., Volckart and Mangels (1999); Kadens (2004, 2012); Sachs (2006); Michaels (2012); criticisms are not addressed here, but see Benson (2011, 2014a, 2014b)].

Among factors that encouraged Europe’s advance into the “High Middle Ages” were technological innovations reducing labor requirements in agriculture, allowing specialization, and resulting in substantial expansions in production and trade. Growing numbers of merchants from widely dispersed areas of Europe traded with one another, primarily at fairs. Transactions costs were high as merchants had different cultural and ethnic backgrounds. Such transactions costs could be reduced with legal arrangements to increase credibility of merchant promises, but to the degree that states existed, they were “unable to supply the basic services of the state” (Volckart and Mangels 1999, 435), including enforcement of commercial contracts. In this context, “the basic concepts and institutions of ... *lex mercatoria* ... were formed” in eleventh and twelfth century Europe by merchants themselves (Berman 1983, 333).

Customary Law

Lex Mercatoria was customary law: a system of rules and governance processes that

spontaneously evolved within merchant communities and recognized because of trust arrangements, reciprocities, mutual insurance, and reputation mechanisms, including ostracism threats. Negotiation (contracting) was the most important source of legal change. Agreements only applied to the parties involved, but others voluntarily adopted changes if they appeared beneficial. As resulting behavior spread it became expected, and a new rule was recognized. Individuals also could unilaterally adopt behavior which others then observed, came to expect, and emulated, or an arbitrator/mediator might offer an innovative solution to a dispute, followed by voluntarily adopted by others.

Contracts and Credit

Face's (1958, 1959) examination of twelfth and thirteenth century documents demonstrates that nonsimultaneous (contractual) trade between Northern and Southern European merchants was the overwhelming dominant practice at the Champagne fairs, the most important European fairs at the time. Six annual fairs were held sequentially in different Champagne towns, each lasting about 52 days. These fairs included: (1) eight "entry" days when merchants set up their shops, (2) ten days for exclusively trading cloth, (3) eleven days to trade cordovan (leather) goods, (4) nineteen days when goods sold by weight, such as spices and dye-stuffs, were traded, and (5) four days for settling accounts and drawing up "letters of the Fair" (Face 1958, 427). Thus, Southern European "caravan" merchants bought northern wool and linen before selling their spices, dyes, or cordovan, using promissory notes or letters of credit from money lenders (or simply verbal promises) accepted by French, German, English, and Flemish merchants. These financial instruments were negotiable so northern merchants could buy goods from any other merchant or retain instruments for later use. Documents also demonstrate that caravan merchants bought spices, dyes, and cordovan on credit in Mediterranean ports and made large

numbers of cloth sales to *draperii*, again often on credit. Similarly, Northern merchants bought and transported cloth to Champagne, sold it, and bought spices, dyes, and/or cordovan to transport and sell, almost entirely through the use of credit. Merchant also contracted with agents or partners "to act in his place, to fulfill old obligations, and in many cases to undertake new ones" (Face 1958, 431), and with professional freighters who transported trade goods to and from fairs (Face 1959). Clearly, as the *Little Red Book of Bristol* c. 1280 (Basile 1998, 11) states, "it is well known to all that merchants sell their goods and merchandize on credit . . . and also that servants and apprentices of such merchants [sell on credit] the goods and merchandise of their lords to other men in the same way" (parenthetic in original).

Merchants had incentives to live up to contractual promises and behave as expected because information spread rapidly throughout interdependent merchant communities (Face 1958; Milgrom et al. 1990; Grief 2006). Therefore, reputations for honesty became valuable, and the threat of spontaneous, uncoordinated but effective punishment for misbehavior developed: ostracism by merchants as they received information about dishonesty (Greif 2006, 66–69).

Disputes

Deterrence is never perfect, so opportunistic breaches no doubt occurred. A merchant might accuse another of misbehavior, however, and the accused might deny it. Disputes also could arise because parties disagreed about what rule applied or how to deal with an unanticipated contingency. Impartial third-party dispute resolution reduced costs of disagreements (i.e., violence) and therefore, encouraged contracting. Merchants included arbitration clauses in contracts (Face 1959, 243), and reputable merchants served as arbitrators (Malynes 1622 [1686], 447–454). Few documents from arbitrated disputes survive, however, in part because merchant arbitrators had a "determinate power to make an end of controversies in general terms, without declaration of particulars"

(Malynes 1622 [1686], 450). Merchants wanted quick solutions to disputes so they chose arbitrators with “skill and knowledge of the Customs of Merchants, which always does intend expedition” (Malynes 1622 [1686], 450).

Temporary participatory courts, often called *Piepoudre* or Pie Powder courts, were also established at fairs. Groups of merchants traveling to a fair generally chose captains or consuls to perform various administrative duties such as determining locations of merchant stalls, but they also served on the fair court (Bewes 1923, 14). As with arbitrators, Pie Powder courts did not leave significant records: their purpose was to provide quick and equitable dispute resolution so merchants could complete transactions at one fair and quickly move to the next. Indeed, Pie Powder court rulings were not binding on any merchant other than the parties in the disputes and then only for the terms of their contracts. Decisions from different merchant courts could be inconsistent, as local rules were understood to apply in a market or fair. In fact, merchants did not have to use Pie Powder courts. Other legal systems were evolving (Berman 1983) and if a nonmerchant court with a reputation for fair decisions was willing to try a case quickly, merchants certainly could and did use it. Ecclesiastical courts were often available, for instance, because many fairs were held at priories and abbeys. Furthermore, the Church was a major producer and trader so ecclesiastical commercial rules often were consistent with *Lex Mercatoria*.

Some Pie Powder courts were taken over by authorities such as “a mayor of a corporate town. Sometimes they belonged to a lord” (Holdsworth 1903: 331). Even when a manor or urban court claimed jurisdiction, however, merchants often dominated dispute resolution (Basile 1998 [c. 1280]: 20). For instant, a lord might impose his court on a fair but use merchant juries. Some local authorities also imposed some of their own rules, but the ability to do so depended on availability of other potential courts (e.g., arbitration) and trading locations. An authority with a particularly attractive fair location could capture location rents through his court. Kings often “granted” rights to hold fairs, however, and in doing so, they

frequently stated that merchant law must apply. These grants may have deterred local powers from interfering with markets. While merchant “justice during the fairs” often was recognized by royal authority, it should not be inferred that royal “backing” was necessary for *Lex Mercatoria* when it was not threatened by a coercive power.

Universal and Polycentric Law

The medieval Law Merchant was polycentric (Benson 1999, 2011, 2014a, 2014b), with parallel, interdependent, and overlapping merchant communities using their own rules and procedures, but this does not mean that the most important principles of *Lex Mercatoria* were not universal (Epstein 2004, 8–9). Indeed, even though variations in rules over time and space were common, “the law, in its broad lines, as laid down by the merchants . . . was necessarily of the international character” (Bewes 1923, 299). By 1200 merchant behavior in commercialized Europe implies recognition of universal rules such as (Benson 2011, 2014a, 2014b):

1. Respect other merchants’ property rights.
2. Respect freedom of contract.
3. Be honest.
4. Do what you promised to do in a valid agreement, unless a subsequent voluntarily agreement alters the first contract.
5. Provide truthful information about observed misbehavior of other merchants
6. Follow local practices and usage unless all parties agree to behave otherwise.
7. Provide accurate information to foreign merchants about relevant local practices and usage.
8. Treat all merchants at a fair or market equitably.
9. If a dispute cannot be resolved, call upon an arbitrator(s) or judge(s) who is either agreed to by both parties or chosen by the relevant group of merchants [e.g., those attending a fair]
10. Accept a reputable arbitrator(s) or judge(s) who is knowledgeable about the relevant customs, practices and usage.

11. Abide by the resolution proposed by an arbitrator(s) or judge(s) within the confines of the contract generating the dispute; the same dispute is not to be taken [appealed] to another adjudicator.
12. The terms of a particular contract and the resolution of a particular dispute do not impose rules that must be followed in future interactions.
13. Support reputable members of your community [guild, caravan, ethnic or national group] if called upon to protect property or assist in pursuit and collection,
14. Provide financial support to reputable merchants from your community, and if you receive such a surety loan, repay it in a timely fashion.

There probably were other universal rules, and the importance of some rules probably varied over time and space. Significantly, however, by accepting these kinds of fundamental general rules, many more specific behavioral requirements could be generated through negotiation and contracting, and observation and emulation of effective rules and procedures meant that over time different communities evolved in similar ways.

Custom Versus Authority

When most trade took place at temporary fairs, incentives for local powers to try to control commerce were relatively weak. However, as trade expanded, permanent market towns developed. Kings often “granted” legal authority to politically important urban governments just as they did for other local authorities. Market towns were governed by local merchant and craft guilds so many *Lex Mercatoria* rules were adopted in urban law. When coercive power is established, however, discriminatory rules can be imposed (Benson 1999, 2011, 2014a, 2014b). Many urban governments began discriminating against some foreign merchants while simultaneously granting others special privileges in exchange for similar privileges in the towns where those

merchants were based (Holdsworth 1903, 302). Thus, as Coquillette (1987, note 21) explains, “surviving correspondence forms between one fair court and another, even fair courts of different countries, show a formula that clearly distinguishes between the law merchant . . . and the town customs.” Kings also extracted revenues or political benefits from commerce by imposing Royal Law (Trakman 1983; Benson 1989). Thus, *Lex Mercatoria* was absorbed and changed in varying degrees over time and space. Nonetheless, it continues to apply within many trade associations (Benson 1995) and in most international trade (Trakman 1983; Benson 1999, 2014b).

Cross-Reference

► Customary Law

References

- Benson BL (1989) The spontaneous evolution of commercial law. *South Econ J* 55:644–661
- Benson BL (1995) An exploration of the impact of modern arbitration statutes on the development of arbitration in the United States. *J Law Econ Org* 11:479–501
- Benson BL (1999) To arbitrate or to litigate: that is the question. *Eur J Law Econ* 8:91–151
- Benson BL (2011) The law merchant story: how romantic is it? In: Calliess G, Zumbansen P (eds) *Law, economics and evolutionary theory*. Edward Elgar, London, pp 68–87
- Benson BL (2014a) Customary commercial law, credibility, contracting, and credit in the High Middle Ages. In: Boettke P, Zywicki T (eds) *Austrian law and economics*. Edward Elgar, London, forthcoming
- Benson BL (2014b) Yes Virginia, there is a law merchant. Working Paper, Department of Economics, Florida State University
- Berman H (1983) *Law and revolution: the formation of Western legal tradition*. Harvard University Press, Cambridge, MA
- Bewes W (1923) *The romance of the law merchant: being an introduction to the study of international and commercial law with some account of the commerce and fairs of the middle ages*. Sweet & Maxwell, London
- Bickley F (ed) (1900) *The Little red book of Bristol, c. 1280*. Bristol: W Crofton Hemmons
- Coquillette D (1987) Ideology and incorporation III: reason regulated – the post-restoration English civilians, 1653–1735. *Boston Univ Law Rev* 67:289–361

- Epstein R (2004) Reflections on the historical origins and economic structure of the law merchant. *Univ Chic J Int Law* 5:1–20
- Face R (1958) Techniques of business in the trade between the fairs of Champagne and the south of Europe in the twelfth and thirteenth centuries. *Econ Hist Rev* 10:427–438
- Face R (1959) The *Vectuarii* in the overland commerce between Champagne and southern Europe. *Econ Hist Rev* 12:239–246
- Greif A (2006) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, New York
- Holdsworth W (1903) A history of English law, vol 1. Methuen & Co, London
- Kadens E (2004) Order within law, variety within custom: the character of the medieval merchant law. *Univ Chic J Int Law* 5:39–65
- Kadens E (2012) The myth of the customary law merchant. *Texas Law Rev* 90:1153–1206
- Malynes G (1622 [1686]) *Consuetudo, vel, lex mercatoria* or, The ancient law-merchant, in three parts, according to the essentials of traffick: necessary for statesmen, judges, magistrates, temporal and civil lawyers, mint-men, merchants, mariners, and all others negotiating in any parts of the world. Printed for T. Basset, R. Chiswell, M. Horne, and E. Smith, London: Islip
- Marius J (1651 [1790]) *Advice Concerning Bills of Exchange* wherein is set forth the nature of exchange of monies, the several kinds of exchange in different countries, divers cases propounded and resolved, objections answered, &c.: with two exact tables of old and new style. Re-printed by D. Humphreys, Front-Street, near the drawbridge, Philadelphia London: Anno
- Michaels R (2012) Legal medievalism in *lex mercatoria* scholarship. *Texas Law Rev* 90:259–268
- Milgrom P, North D, Weingast B (1990) The role of institutions in the revival of trade: the law merchant, private judges, and the Champagne fairs. *Econ Polit* 2:1–23
- Mitchell W (1904) *Essays on the early history of the law merchant*. Burt Franklin, New York
- Sachs S (2006) From St. Ives to cyberspace: the modern distortions of the medieval law merchant. *Am Univ Int Law Rev* 21:685–812
- Stracca B (1553) *De mercatura seu mercatore tractatus*. No printer listed, Venice
- Teeter R (1962) England's earliest treatise on the law merchant: the essay on *lex mercatoria* from The Little Red Book of Bristol (Circa 1280 AD). *Am J Leg Hist* 6:172–210
- The Little Red Book of Bristol (1998) *Lex mercatoria*. In: Basile M (ed) *Lex mercatoria and legal pluralism: a late thirteenth century treatise and its afterlife*. Ames Foundation, Cambridge, UK
- Trakman L (1983) The law merchant: the evolution of commercial law. Fred B. Rothman and Co, Littleton
- Volckart O, Mangels A (1999) Are the roots of the modern *lex mercatoria* really medieval? *South Econ J* 65:427–450

Zouche R (1663) The jurisdiction of the Admiralty of England asserted against Sr. Edward Coke's *Articuli admiraltatis*, in XXII chapter of his jurisdiction of courts. Printed for Francis Tyton and Thomas Dring, London

Further Reading

Migne JP (1850 [1936]) *Patrologiae cursus completus*, vol LXXX. Garnier Freres, Paris. <http://pld.chadwyck.com.proxy.lib.fsu.edu/>

Lex Talionis

Mark D. White

Department of Philosophy, College of Staten Island/CUNY, Staten Island, NY, USA

Definition

The classical formulation of the view that the guilty must be punished in exact proportionate to their crimes: “an eye for an eye, a tooth for a tooth.”

The *lex talionis* is otherwise known as the view that punishment for crimes must exact “an eye for an eye, a tooth for a tooth.” It dates at least to the law of Moses and the Code of Hammurabi, and the general idea is cited in modern times by both scholars and laypeople in support of punishment that “fits the crime.” The *lex talionis* is sometimes used as a justification of *retributivist punishment*, which requires that the guilty receive their due punishment as a matter of right or justice, although it is less of a justification and more of a statement of the proper target of punishment (the guilty) and degree of punishment (proportionality).

The *lex talionis* is cited by key retributivists, including its leading proponent, Immanuel Kant, who asked “what kind and what amount of punishment is it that public justice makes its principle and measure? None other than the principle of equality. . . . Accordingly, whatever undeserved evil you inflict upon another within the people, that you inflict upon yourself. . . . But only the *law of retribution (ius talionis)*. . . can specify definitely the quality and the quantity of punishment”

(1797, p. 332). However, as modern retributivists and their critics alike realize – and as asserted definitively by Blackstone (1765–1769, bk 4, Chap. 1) – exact proportionality is either inhumane (such as in the case of rape), impossible (the case of multiple murders), or nonsensical (the case of attempted crime). Kant acknowledged this, asking “but what is to be done in the case of crimes that cannot be punished by a return for them because this would be either impossible or itself a punishable crime against *humanity* as such?” and moderated the *lex talionis* to prescribe instead that “what is done to [the wrongdoer] in accordance with penal law is what he has perpetrated on others, if not in terms of its letter at least in terms of its spirit” (1797, p. 363). Appropriately, it is the spirit of the *lex talionis* rather than its letter that has survived on in the form of modern retributivism, which emphasizes proportionality while recognizing its difficulties (Davis 1983, 1986).

The call for proportionality embodied in the *lex talionis* is often invoked against systems of punishment, such as deterrence, which focus on generating social benefits from punishment instead of enforcing the just deserts of the guilty. While optimally deterrent punishments – such as those recommended by the economic approach to crime – result in some degree of proportionality between crimes to create optimal incentives, they can also be disproportionately severe to compensate for high enforcement costs (Becker 1968). However, those who raise the *lex talionis* in argument are usually less concerned with disproportionately high punishment and more with disproportionately *low* ones, such as those resulting from plea bargains or judicial acts of mercy – the first of which can be justified as a regrettable compromise in the face of resource constraints (Cahill 2007) while the second is consistent with retributivism in general (Holtman 2009). At its worse, it can be used – incorrectly – to justify private acts of vengeance and “vigilante justice,” which are distinct from state-sponsored punishment (Nozick 1981, pp. 366–368; Brooks 2012, pp. 16–18). Since more refined accounts of retributivism are available that acknowledge the subtleties of

punishment in both theory and practice, the *lex talionis* is rarely invoked in scholarly debate today.

Cross-References

- [Crime and Punishment \(Becker 1968\)](#)
- [Criminal Sanctions and Deterrence](#)
- [Retributivism](#)

References

- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Blackstone W (1765–1769) *Commentaries on the law of England*. Clarendon, Oxford
- Brooks T (2012) *Punishment*. Routledge, Abingdon
- Cahill MT (2007) Real retributivism. *Wash Univ Law Rev* 85:815–870
- Davis M (1983) How to make the punishment fit the crime. *Ethics* 93:726–752
- Davis M (1986) Harm and retribution. *Philos Public Aff* 15:246–266
- Holtman S (2009) Justice, mercy, and efficiency. In: White MD (ed) *Theoretical foundations of law and economics*. Cambridge University Press, Cambridge, pp 119–135
- Kant I (1797) *The metaphysics of morals* (trans: Gregor M). Cambridge University Press, Cambridge (1996 edn)
- Nozick R (1981) *Philosophical explanations*. Belknap, Cambridge, MA

Liability and Information

Florian Baumann¹ and Tim Friehe²

¹Center for Advanced Studies in Law and Economics (CASTLE), University of Bonn, Bonn, Germany

²Public Economics Group, Philipps-University Marburg, Marburg, Germany

Abstract

The allocation of information is of paramount importance for the efficacy of liability law as an instrument to address externalities. This entry discusses several possible repercussions of imperfect information, such as inefficient

caretaking resulting from the injurer's imperfect information about the standard of care under negligence or distorted decisions about buying liability insurance. In addition, the contribution presents results from the literature on the incentives to acquire information about risk induced by liability law. Decision makers usually have the opportunity to improve their state of knowledge at a cost, and liability law will impact on the incentives to actually do so and to possibly share this information. Starting from Shavell (1992), we discuss – *inter alia* – literature analyzing the incentive for accurate harm assessment and the use of hindsight information in court, the effects of disclosure rules, or adjustments in due-care standards to further the generation of information in a process of learning-by-doing.

Introduction

The availability of information about risks is critical to the performance of liability rules in terms of inducing the efficient outcome. In this entry, we first describe certain repercussions of taking imperfect information as given. In order to provide a full understanding of the functionality of liability rules, we also discuss the incentives that such rules create with regard to the acquisition of information – that is, we consider information allocation as an endogenous outcome that can be influenced by liability rules. This entry presents results from selected works in this field.

Liability and Exogenous Information

The repercussions of taking imperfect information as given have been described for a wide range of important aspects in the context of liability. For example, erroneous assessments of harm by the court *ex post* that can be anticipated by the tortfeasor *ex ante* necessarily distort both care and activity incentives under strict liability and may also do so under negligence (e.g., Endres 1989). In the case of negligence, the choice and communication of the due-care standard are

important variables. In order to set the appropriate standard of care and to be able to compare actual care with the standard, the court needs to know all aspects relevant to the cost minimization and to observe care without error, as otherwise excessive or suboptimal caretaking may result (e.g., Craswell and Calfee 1986; Shavell 1987, Chap. 4). Moreover, when risk-averse tortfeasors are uncertain about the workings of the legal system or individual accident risk, this may affect their decision to buy liability insurance (see, e.g., Crocker and Doherty 2000; Shavell 2000). For the alternative care model, Dari-Mattiacci and Garoupa (2009) show that the least-cost avoider rule can lead to inefficient outcomes when there is imperfect information about the other party's cost of care. In the domain of product liability, how well consumers are informed about product risks is a key factor in determining which liability rule ought to be used (see, e.g., Daughety and Reinganum 2013; Miceli et al. 2015; Shavell 1987, Chap. 7). Liability rules also influence decisions about innovation and the adoption of technologies with differing characteristics in terms of the cost of care or expected harm. In this context, the policy-maker may not have all relevant information on hand (e.g., about the technologies available to firms or the level of firms' adoption costs); this means that the choice and design of the liability rule are decisive for the eventual outcome (see, e.g., Dari-Mattiacci and Franzoni 2014; Endres and Bertram 2006). In practice, when courts lack information about the relevant accident technology, they may refer to industry custom as the relevant negligence standard, potentially hindering innovation and favoring the replication of established technologies (Parchomosky and Stein 2008). For many innovations, the lack of experience will make it impossible to arrive at a reasonable estimate of the accident probability, which could, *inter alia*, undermine the insurability of risks (Skogh 1998).

Liability and Endogenous Information

The assumption that imperfect information is inevitable disregards an important factor in the

incentive effects of liability rules. Shavell (1992) provides a seminal contribution on the influence of different liability rules on the injurer's incentives to acquire information about risk and on whether they are aligned with social incentives. In his setting, information about whether or not the activity in question is risky can be acquired at a positive fixed cost. The social incentives to acquire information about risk depend on a comparison of the social value of information (stemming from the expected adjustment of care away from the level that is optimal based solely on expectations about the riskiness of the activity) and the social cost of information. When considering the performance of liability rules, Shavell (1992) employs the common assumptions that injurers are risk neutral, that the level of damage is equal to the level of harm, and that there are no issues regarding the establishment of causation. In such a scenario, strict liability results in the perfect alignment of private and social incentives to acquire information about risk, since the injurer's cost minimization problem is the same as that of the planner. For negligence, various versions of the liability rule may apply, as different behavioral dimensions may be included in the negligence standard. A comprehensive negligence rule – that is, one that holds injurers liable (a) when they did not obtain information even though it was socially desirable to do so or (b) when they did not exercise the level of care that minimizes social costs contingent on the socially optimal level of information – will induce the socially optimal outcome. Importantly, Shavell (1992) establishes that a negligence rule that uses only criterion (b) for the level of care also induces socially optimal choices by the injurer with regard to both care and information acquisition, demonstrating that one standard of behavior suffices for efficient incentives under negligence in his setup. In fact, the private value of information exceeds the social value of information, as a tortfeasor completely avoids paying the level of harm by being nonnegligent in the risky state. However, other due-care standards – namely, (i) the level of care that is optimal given the injurer's information or (ii) the level of care that is optimal when information is obtained – cannot align private and

social incentives. Standard (i) provides insufficient incentives to obtain information about risk because the injurer can avoid any liability by choosing the appropriate care in view of expectations about the riskiness of the activity; standard (ii) may also lead to inadequate information acquisition and to inefficient care.

Bajtelsmit and Thistle (2015) extend the results of Shavell (1992) by considering risk-averse tortfeasors in a framework in which liability insurance is provided by perfectly competitive insurance companies. Insurance premiums can be contingent on the observable level of care and the injurers' information about the riskiness of their activity, implying that insurance premiums match expected harm as a function of care. Bajtelsmit and Thistle find that the resulting classification risk entails a significant cost of acquiring information about the activity's riskiness, potentially encouraging individuals to avoid becoming informed; it may also render the social value of information negative. This classification risk means that injurers who acquire information about risk may discover that their expected harm is either high or low, such that information acquisition is associated with an uninsurable lottery; the alternative of remaining uninformed and acting according to the *ex ante* expectation about risk precludes any variation in income. As is true in Shavell (1992), injurers' incentives to acquire information about risk are aligned with social incentives under strict liability, whereas the results for negligence are more complex.

Kaplow and Shavell (1996) explore the social justification for the accurate measurement of harm in court; this is often a key issue in litigation and is in all likelihood responsible for a significant share of litigation expenditures. As implied by the social value of information formulated in Shavell (1992), the accurate measurement of harm *ex post* may be socially valuable only when the information can be anticipated *ex ante*. Only in this case will an injurer be subject to strict liability increase (decrease) the level of care – relative to the benchmark level without information acquisition – in states in which the expected harm exceeds (falls below) the average expected harm. By implication, acquiring information

about the level of harm in court is not socially valuable when the injurer cannot or should not obtain such information before making the choice about the level of care. Significantly, the court's decision of whether or not to accurately assess the level of harm is relevant for injurers' decisions about the acquisition of information *ex ante*.

Related to Kaplow and Shavell (1996), Ben-Shahar (1998) explores the implications of using information about the harmfulness of products that was not available when the product was marketed in the court's assessment of producer liability *ex post*. The analysis focuses on how the use of hindsight information (in comparison to a state-of-the-art approach) alters incentives for safety investments before and after the product's release, incentives for developing new technologies, and incentives to become better informed about product risks *ex ante*. The author finds, *inter alia*, that hindsight increases incentives for remedial actions after product release but distorts *ex ante* safety decisions. The impact of hindsight on the incentive to invest in new technologies is ambiguous; generally speaking, neither a state-of-the-art regime nor a regime based on hindsight results in first-best decisions regarding the acquisition of information. For environmental law, Wagner (2004) argues that it actually weakens incentives to acquire information about product risk and instead motivates attempts to discredit public research on the matter, as the information is likely to be used in support of additional regulations or standards.

In Polinsky and Shavell (2012), firms with different products and heterogeneous information acquisition costs choose whether or not to obtain information about their product risks and – contingent on having obtained that information – whether or not to truthfully transmit it to consumers. Such information may be socially valuable because it allows consumers with heterogeneous consumption benefits to make informed choices. The policy issue analyzed is whether forcing firms to reveal any information they possess (i.e., mandatory disclosure) is socially preferable to leaving disclosure decisions to firms (i.e., voluntary disclosure). Because information costs are considered private information,

the regulation of the acquisition of information is not included in the model (unlike Shavell 1992). Under the assumption that the firm's level of care must be decided before information about risk can be obtained, the authors show that mandatory and voluntary disclosure of product risks are equivalent under strict liability (as consumers' choices are not responsive to information in the case of complete compensation). In contrast, under negligence, incentives to acquire information are greater under voluntary disclosure, but the actual disclosure choice made given the information acquired is more desirable under mandatory disclosure.

Baumann and Friehe (2016) depart from Shavell (1992) by assuming that information about risk can only be acquired via learning-by-doing. In their two-period framework, the policy-maker has a prior about the true accident technology (i.e., the level of risk and how it responds to care) and uses the accident history in the first period to update the assessment of the accident technology in the second period. In this setup, because the signaling value of the accident history depends on care, the level of care taken in the first period plays a role in both the minimization of social costs and the acquisition of information about the accident technology. Accordingly, optimal care in the first period may deviate from the level that minimizes the sum of care and expected harm costs. There are circumstances in which strict liability and negligence both fail to induce the socially optimal level of care in the first period.

The idea of learning from an accident history is also explored in Feess and Wohlschlegel (2006). In their setup, some injurers know the true accident technology, whereas other injurers and the courts may be uninformed; moreover, actual injurer care is erroneously reconstructed by the court. Under these circumstances, the court will use actual injurer care as an informative signal about the true accident technology and will dynamically adjust the due-care standard accordingly. The planner thus transmits a better understanding of the accident technology (obtained from observed levels of care in negligence trials) to the uninformed injurers by adjusting the due level of care.

Most of the literature in law and economics uses the standard assumptions of rational choice theory – for example, that risk-averse subjects maximize expected utility and that future utility is incorporated using exponential discounting. Behavioral aspects may introduce other distortions with regard to what constitutes privately optimal information acquisition. For example, Chemarin and Orset (2011) discuss whether present-biased agents may acquire less information about risk than individuals with time-consistent preferences.

References

- Bajtelsmit V, Thistle P (2015) Liability, insurance, and the incentive to obtain information about risk. *Geneva Risk and Insur Rev* 40:171–193
- Baumann F, Friehe T (2016) Learning-by-doing in torts: Liability and information about accident technology. *Econ Lett* 138:1–4
- Ben-Shahar O (1998) Should product liability be based on hindsight? *J Law Econ Org* 14:325–357
- Chemarin S, Orset C (2011) Innovation and information acquisition under time inconsistency and uncertainty. *Geneva Risk and Insur Rev* 36:132–173
- Craswell R, Calfee J (1986) Deterrence and uncertain legal standards. *J Law Econ Org* 2:279–303
- Crocker KJ, Doherty N (2000) Why people buy liability insurance under the rule of simple negligence. In: Baye MR (ed) *Industrial organization (advances in applied microeconomics)*, vol 9. Elsevier, Amsterdam, pp 133–148
- Dari-Mattiacci G, Franzoni LA (2014) Innovative negligence rules. *American Law Econ Rev* 16:333–365
- Dari-Mattiacci G, Garoupa N (2009) Least-cost avoidance: the tragedy of common safety. *J Law Econ Org* 25:235–261
- Daughety A, Reinganum J (2013) Economic analysis of products liability: theory. In: Arlen JH (ed) *Research handbook on the economics of torts*. Cheltenham: Edward Elgar Cheltenham, UK.
- Endres A (1989) Liability and information. *J Inst Theor Econ* 145:249–274
- Endres A, Bertram R (2006) The development of care technology under liability law. *Int Rev Law Econ* 26:503–518
- Feess E, Wohlschlegel A (2006) Liability and information transmission: the advantage of negligence based rules. *Econ Lett* 92:63–67
- Kaplow L, and S. Shavell, 1996. Accuracy in the Assessment of Damages. *Journal of Law and Economics* 39, 191–210.
- Miceli TJ, Segerson K, Wang S (2015) Products liability when consumers vary in their susceptibility to harm and may misperceive risk. *Contemp Econ Policy* 33:468–476
- Parchomosky G, Stein A (2008) Torts and innovation. *Mich Law Rev* 107:285–315
- Polinsky AM, Shavell S (2012) Mandatory versus voluntary disclosure of product risks. *J Law Econ Org* 28:360–379
- Shavell S (1987) *Economics analysis of accident law*. Harvard University Press, Cambridge, MA
- Shavell S (1992) Liability and the incentive to obtain information about risk. *J Leg Stud* 21:259–270
- Shavell S (2000) On the social function and the regulation of liability insurance. *Geneva Pap Risk Insur* 25:166–179
- Skogh G (1998) Development risks, strict liability, and the insurability of industrial hazards. *Geneva Pap Risk Insur* 23:247–264
- Wagner WE (2004) Commons ignorance: the failure of environmental law to produce needed information on health and the environment. *Duke Law J* 53:1619–1745

Libertarian Paternalism

Samuel Ferey

CNRS, BETA, University of Lorraine, Nancy, France

Definition

In their famous book *Nudge*, published in 2008, Thaler and Sunstein popularize libertarian paternalism, one of the normative theories inspired by behavioral economics. Contrary to a large tradition in moral and political philosophy that dates back to John Stuart Mill, Sunstein and Thaler can see no inconsistency between paternalism and liberalism. We discuss this statement, and we illustrate libertarian paternalistic policies. Then, we deal with the descriptive, prescriptive, and normative consequences of libertarian paternalism for the law. Last, some of the main criticisms against libertarian paternalism are discussed.

Libertarian Paternalism and the Law: Taking Cognitive Bias Seriously

In an executive order dated September 15, 2015, and entitled “Using Behavioral Science Insights to Better Serve The American People,” American executive

departments and agencies “are encouraged to [...] design public policies in line with the findings of behavioral academic theoretical and applied literature.” And the executive order refers explicitly to one of the famous libertarian paternalistic nudges, the “automatic enrollment and automatic escalation in retirement savings plans.”

Initiated at the beginning of the 2000s in several papers (Sunstein and Thaler 2003a, b), libertarian paternalism has been popularized by the famous book *Nudge* published in 2008 by Thaler and Sunstein (2008). In less than 15 years, libertarian paternalism has developed a large body of theoretical, applied, and experimental findings covering many fields of public policies (health, insurance, finance, fight against poverty, etc.) and became one of the inspiring thoughts for regulators and legislative bodies throughout the world.

The originality of libertarian paternalism relies on its constitutive so-called oxymoron: a normative theory according to which paternalism and liberalism are not contradictory anymore. For libertarian paternalism, the famous John Stuart Mill’s warnings stating that “the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant. He cannot rightfully be compelled to do or forbear because it will be better for him to do so” (Mill 1869) are challenged by the recent findings of behavioral economics.

Thanks to behavioral economics, scientists and regulators know better the ways people respond to laws and regulation; they are aware of the bias people suffer, of their cognitive limitations, and of the weakness of their willpower. And the law should take people for who they are and be sensitive to the ways in which they depart from rational action theory predictions.

Libertarian Paternalism and Behavioral Economics

Libertarian paternalism elaborates on hypothesis, ideas, and methods raised in behavioral

economics. First, people are humans, not econs. They are not fully rational, coherent, and optimizers. On the contrary, they make systematic mistakes, do not perfectly learn, and are bad – at least compared to the fully rational agent ideal – at assessing probabilities, events, or any other unknown facts (Thaler and Sunstein 2008). Kahneman used to employing a metaphor: mind is led by two systems, system 1 and system 2. System 1 “operates automatically and quickly, with little or no effort and no sense of voluntary control.” System 2 “allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration” (Kahneman 2011). Explaining how the two-level process works and interacts with the environment is the basic agenda of behavioral sciences.

For Kahneman, Tversky, Thaler, and many others, it is possible to capture the process of system 1 and to display behavioral anomalies by using new empirical methods like laboratory and field experiments. Consequently, the scope of behavioral analysis for economics is to explain the ways people depart from rationality – the *bias* – and to draw the “maps” of bounded rationality. Among the main bias identified, overconfidence, loss aversion, status quo bias, anchoring, and representativeness are the most popular.

From Debiasing to Libertarian Paternalism

These new findings on how people actually behave have two normative consequences. First is debiasing: legal norms, laws, and regulations are expected to enhance individual rationality and to remove or to alleviate rationality failures. Second is libertarian paternalism. Rationality failures and bias are used to enhance people welfare. As Thaler and Benartzi state “Libertarian paternalism is a philosophy that advocates designing institutions that help people make better decisions but do not impinge on their freedom to choose. Automatic enrollment is a good example of libertarian

paternalism” (Thaler and Benartzi 2004). A benevolent policy maker, who is aware of the bias suffered by individuals, could decide to implement laws and policies which softly induce people to behave in lines with their true preferences. Under libertarian paternalistic institutions, systematic errors from system 1 are the best servant of system 2.

Let’s consider the previous example of automatic enrollment, and let’s take seriously that people suffer a status quo bias. Imagine two contexts (context 1, option A is chosen by default; context 2, option B is chosen by default). Last, suppose there is no economic or financial cost to change the initial option. In that case, because of the bias, people are more likely to choose the default option than the other. Such a mechanism is meaningful in a world where rationality is bounded. Because of the bias, the choice architect – the one who draws the default option – is able to induce people to behave in a certain way. The main issue of libertarian paternalism is to know how such a framing could be made. For Thaler and Sunstein, the choice architect should systematically select the default option which is in line with the best preferences of people: the automatic processes of system 1 are used to increase the savings, that is, what the long-term self (the far-sighted self) would have chosen. And this effect emerges without any intrusive policies and without depriving people of their rights and liberties: people stay free to choose any of the options. Here is the heart of the libertarian paternalist doctrine – also called *nudging* even though all nudges are not libertarian paternalistic (Mongin and Cozic 2017).

Libertarian Paternalism for the Law

A lot of nudges may be found in our daily life. One of the reasons why the eponymous book *Nudge* has been so popular lies in the examples provided by Thaler and Sunstein: the strives on a highway which induce people to operate the brakes before starting a dangerous turn, the alarm clock “clocky” that “runs away and hides

you if you don’t get out of bed” to struggle against lack of willpower, the default rule about your presumed consent for organ donation, etc. All these mechanisms have in common to use a bias in order to better achieve individuals’ true preferences. And legal and institutional nudges are of particular interest.

Regarding legal issues, three meanings of libertarian paternalism may be distinguished. First is a purely descriptive meaning. Libertarian paternalism makes sense of many existing legal provisions that would be meaningless if they were analyzed through the glasses of rational action theory: the cooling-off periods and delays in consumer law that protect people against their impulsive choices, the default rules that use inertia to better serve long-term self-interests, the legal framing that displays information by using salience bias, the legal self-constraints, etc.

Second, libertarian paternalism has a normative scope by offering a toolbox for new laws and public policy recommendations testable by experiments (Shafir 2012). For example, the randomized experiments designed by J-Pal experts (Poverty Action Lab) testing the best ways to struggle against poverty are sometimes explicitly driven by libertarian paternalistic views (Banerjee and Duflo 2011).

Third, libertarian paternalism is helpful to better understand adjudications by courts. This is becoming more important as more courts refer implicitly or explicitly to behavioral arguments. For example, in the famous antitrust case *Microsoft vs. Commission*, the European Court of First Instance extensively discusses the default rule and eventually considers that it is in the best interests of consumer to have a set of default options without Windows Media Player.

Is Libertarian Paternalism a New Paradigm for Law and Economics?

Libertarian paternalism is said to be a syncretic normative theory that liberals as well as conservatives could support. However, many criticisms

have raised. First, policy makers, legislators, and judges following libertarian paternalistic recommendations are assumed to be benevolent. What happens if they are not? Using behavioral sciences would give expertise in manipulation to public agents and contradict the neutrality of the State. Manipulation threatens individual autonomy, and paternalistic views threaten liberalism. Second, policy makers themselves suffer bias and are perhaps less competent than expected. One of the most important results of behavioral science is precisely to warn experts against their overconfidence and their illusion of validity and to show how experts' judgments may depart from rationality. Third, it is argued that market is able to deal with cognitive bias (Sugden 2008). But this point about the spontaneous debiasing of people by the market is still disputable in the empirical and theoretical literature (Gabaix and Laibson 2006).

These criticisms lead Sunstein to be more specific about political, ethical, and economic justifications of libertarian paternalism (see Sunstein 2014, 2015). Three arguments are of particular interest for the law. First, the choice architecture is inevitable. Any legal system offers choices under a specific frame. In that case, why not choosing the best frame for individuals? The argument is sound and convincing. Second, the threat of manipulation does not undermine libertarian paternalism as soon as libertarian paternalistic policies are under constant public scrutiny. According to Sunstein, transparency and democratic control are sufficient to avoid manipulative nudges. Third, libertarian paternalism could be considered as a set of devices enhancing autonomy because "we might identify autonomy with people's reflective judgements, and many nudges operate in the interest of autonomy, so understood" (Sunstein in Alemanno and Sibony 2015; Sunstein 2015).

To conclude, libertarian paternalism offers new smart insights on the role of the law in the economic system. But some further developments have to be made to know whether libertarian paternalism could be a general theory of the law and regulation.

Cross-References

- [Cognitive Law and Economics](#)
- [Experimental Law and Economics](#)
- [Nudge](#)

References

- Alemanno A, Sibony AL (eds) (2015) *Nudge and the law. A European perspective*. Hart, Oxford
- Banerjee A, Duflo E (2011) *Poor economics. A radical rethinking of the way to fight global poverty*. Public Affairs, New York
- Gabaix X, Laibson D (2006) Shrouded attributes, consumer myopia, and information suppression in competitive markets. *Q J Econ* 121:505–540
- Kahneman D (2011) *Thinking, fast and slow*. Farrar, Straus & Giroux, New York
- Mill JS (1869) *On liberty*, 4th edn. Longman, Roberts & Green, London
- Mongin Ph, Cozic M (2017) Rethinking nudge: not one but three concepts. *Behav Public Policy* 1 (forthcoming)
- Shafir E (ed) (2012) *The behavioral foundations of public policy*. Princeton University Press, Princeton
- Sugden R (2008) Why incoherent preferences do not justify paternalism. *Constit Polit Econ* 19:226–248
- Sunstein C (ed) (2001) *Behavioral law and economics*. Cambridge University Press, Cambridge
- Sunstein C (2014) *Why nudge? The politics of libertarian paternalism*. Yale University Press, New Haven
- Sunstein C (2015) The ethics of nudging. *Yale J Regul* 32:413–450
- Sunstein C, Thaler RH (2003a) Libertarian paternalism is not an oxymoron. *Univ Chicago Law R* 70: 1159–1202
- Sunstein C, Thaler RH (2003b) Libertarian paternalism. *Am Econ Rev* 93:175–179
- Thaler RH, Benartzi S (2004) Save more tomorrow: using behavioral economics to increase employee savings. *J Polit Econ* 112:164–187
- Thaler RH, Sunstein C (2008) *Nudge*. Yale University Press, New Haven

Further Reading

- Arieli D (2009) *Predictably irrational*. Harper, London
- Camerer C, Issacharoff S, Loewenstein G, Rabin M (2003) Regulation for conservatives: behavioral economics and the case for "Asymmetric Paternalism". *Univ PA Law Rev* 151:1211–1254
- Saint-Paul G (2011) *The tyranny of utility*. Princeton University Press, Princeton
- Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. *Science* 185 (4157):1124–1131

Liberty

Francois Facchini

Faculté Jean Monnet, University of Paris 11
(RITM) and Associate Economist, Centre
d'Economie de la Sorbonne, University of Paris 1
(France), Sceaux, France

Abstract

Freedom is the power to do what I want to do. The laws of nature and/or human laws limit this power. The laws of nature impose necessities. I cannot choose the speed of my fall (law of gravitation). Here I cannot have a physical meaning. Human laws limit also my power to choose, but in another meaning. They impose obligations. The law of adultery, for instance, forbids to have sexual relationships with its children. I must not in a moral sense. I do not have the right. So, human laws determine artificially the limits of my power to choose. They limit the infinite freedom of the will and leads to the study of conditions of concrete freedom possessed by human beings in society. Concrete freedom is limited by the will of others and expressed through law partially originates in a process of mutual recognition. Then, concrete freedom is based on consent.

Synonyms

Freedom

Definition

Freedom is “the power to do what I want to do.”

Liberty, Laws of Nature and Human Laws

Freedom puts the question to the relationship between human beings and nature by means of a question concerning determinism and, similarly with the relationship between one human being

and another, by means of a question concerning duty or obligation. The response of human beings to the constraint that nature places upon the will is exemplified in technology. The response of human beings to the constraints that can be imposed by other human beings is exemplified by law.

What Is a Free Action?

In order to understand why the notion of freedom or liberty involves such questioning, we must look at our experience of action and the distinction freedom makes between an action performed under constraint and a free action. Freedom is the power to choose what action one carries out. Freedom characterizes a type of action. Fundamentally, a free action is something done that could have been done in a different way. It is distinguished from a reflex action. Let's take an example. I can close my eyes in order to avoid something thrown in my direction. I can open my eyes in order to admire a landscape. When I open my eyes for that purpose, I make a choice. However, when I close my eyes to keep from getting hit in the eye by something thrown at me, the movement of my eyelids, i.e., the change in the initial situation in this case, is not the result of a choice. I never had to form the intention of closing my eyes, but my body is so constituted that it reacts automatically to the approach of such a projectile, and my eyelids close. Even if I had wanted to keep from closing my eyes, I would not have been able to prevent them from reacting, because the movement of my eyelids is determined by the way my body's nervous system works. The other situation, in which I open my eyes to admire a landscape, is different. My eyes open because I decide to open them. I decide to open them because I have good reasons to want to do so. The movement of my eyelids in reaction to the approach of a projectile is not a free act. To the contrary, the movement of my eyelids when I open them in order to admire a landscape is indeed motivated by the desire to admire the landscape. Thus, this is the description of an action that is experienced as conscious and as free. At

this point we can affirm that a free act or action has three characteristics: it is intentional (an act of will), it comes with a justification in the form of reasons for acting or motives (one wishes to enjoy looking at a landscape), and it is not the determinate result of some other act or action. This definition views human beings as responsible for their actions without requiring that actions all succeed in their purpose. Not all intentions are carried out.

Each of the three characteristics of free action calls for further explanation. Free actions are intentional. I open my eyes because I have the intention of admiring the landscape. My action is limited to carrying out this intention. My project, what I intend, is the “why” of me opening my eyes. In this sense an intention is part and parcel of reasons that justify or explain an action. Reasons for acting or motives for action are talked about in this explanation, instead of causes. An approaching projectile causes my eyelids to close together, but the beauty of the landscape is a motive or reason for my action, and not a cause. If we were speaking of final causes, beauty was the ultimate cause of my action. I can give myself a teleological explanation of the fact that I open my eyes to admire the landscape, but I give myself a physical (and *ex post facto*) explanation of the fact that I may be forced to shut my eyes in order to protect them from an approaching projectile. After the fact, I take note of the fact that something was flying toward my eye, which explains me flinching and closing my eyelids together. The explanation of a forced or determinate action is not like the explanation of a freely performed action. In one case a physical force moves a physical body. In the other case, a physical action that could not have been physically predicted is actually freely executed or motivated by a mental intention, which is further mediated by a predilection toward beauty. In the latter case, the free action is a cause of movement; in the former case, physical motion is the consequence of an external force. The cause of movement is the projectile, not the intention (Cowan 1994). The explanation of a reflex action is always *ex post facto*. I must have observed that something was going to hit my eye, and this caused me to react. The explanation of a free act is always *ex*

ante. I justify my decision to look at the landscape *ex ante*, with reference to my taste in landscapes.

The difference between a physical (instrumental) cause and a teleological (final) cause indicates the singular nature of free actions. Free action is an action that has its own cause; it determines itself. The result of this is that one may impute an action to a person or assign responsibility for the action to its “author.” This is the origin of the concept of free will as developed by Saint Augustine (Augustine of Hippo) in his treatise *De libero arbitrio*. God is not responsible for evil; human beings are, since “*Dieu a conféré à sa créature, avec le libre arbitre, la capacité de mal agir et par-là même, la responsabilité du péché*” (*De libero arbitrio*, I, 16, 35). In this regard, the will is what makes an action one's own, placing the burden of responsibility on the one performing the action (*De Libero Arbitrio* I.11). Human beings, on another hand, are not responsible for their eyes shutting when a projectile appears to be heading toward their heads. They are responsible for the act of opening their eyes to look at the landscape. In this sense, a free action involves the individual as moral person, the one who is able to respond by saying, I am the one who did this or that action. The individual is the subject who chooses or decides to look at the landscape. Freedom makes human beings into actors; they are not like stones, which cannot act but only be acted upon (Voltaire 1987, Chapitre 13, tome 62, p. 44). Voltaire makes use of Locke's theory of freedom. Freedom here is synonymous with power. This concept is distinguished from the freedom of the free will, which only has to do with the power of self-determination. A moral action is free if its consequences can be imputed to the person who chooses to do it. Inversely, an action under constraint makes human beings into objects that are acted on by forces they do not control.

In neither case is the success of the action guaranteed. Though my eyes shut reflexively, they may still be damaged by the projectile; the landscape I open my eyes is not guaranteed to be beautiful. Freedom does not protect individuals against errors in judgment that may involve the means chosen to carry out one's projects, in order to fulfill one's intentions. And the errors

committed in the one and the other case are quite different. I would not blame my own reflexes if something flew up and damaged my eye, but rather my own carelessness, or perhaps mere accident; the manner in which my body reflexively moves to avoid an external threat is not a matter of my free choice. It is a given. But a human actor can modify either a project or the means he or she mobilizes to realize the project, following an initial failure. Human beings learn from failure because they fear it. This may seem an obvious point, but it is important because it moves us away from any definition of freedom as the “maximum expansion of my personality” (Berlin 2002a, p. 179) or as itself a means for the realization of projects (Sen 1999, pp. 14–15, p. 37). Such definitions in fact confuse a free action with an action that is successfully performed or carried out. The result of a free action is not determined. This does not mean that I cannot judge freedom by its results. I can, for example, try to find out if a free society is more prosperous than the one that is not free. But such a judgment is not a definition of freedom itself or of the conditions under which it may exist.

Under What Conditions Can a Free Action Exist?

The Absence of Determinism

Regarding free action, as we have just characterized it: it is not immediately certain that it exists in reality. Human existence confirms intentionality, the existence of reasons for acting, the feeling of responsibility, and the possibility of failure. But it is not certain that the act of will that led me to open my eyes upon a landscape was not itself determined by some characteristics of the environment (of the action). Perhaps my reasons for acting or my motives in acting were after all determined by my conditions of existence. If I am not the master of my own actions, my acts are constrained in the sense that they have been determined by the conditions of my existence. Thus, human beings and their actions become again objects of nature, not subjects.

As an object, human action has the same status as a rock, a plant, or an animal. It is determined by the conditions of our existence as living beings. Genes have a desire to reproduce themselves (the law of egoistic genes, to sociobiological determinism of (Dawkins 1976)), human beings have physical needs (sleep, nutrition, etc., indicating physiological determinism), there is a law of natural selection operating (the social Darwinism of Herbert Spencer), the law of egoism or the maximization of utility (social physics), the nature of soils and climates (geographical determinism, Montesquieu, 3^e partie, Livre XIV, chap. X.; Diamond 1997), laws of the unconscious (psychological determinism, Plato and Freud), income levels (economic determinism or materialism), and/or membership in a social class (historical determinism). The free action would thus be an illusion, because the conditions of its being accomplished are not all present. This illusion stems from the fact that human beings are conscious of their actions, but not of the causes that determine that they will act. Thus, the only freedom human beings have is that they can know that they have acted in accordance with the necessity associated with their nature (Spinoza *Ethica* IV, Proposition 68). Freedom would be the recognition of necessity (Berlin 2002b, Chapitre Hegel). If I want to build an airplane, it would be suicidal to attempt to violate any of the laws of aerodynamics. Fatalism is the moral (prescriptive) consequence of this definition of freedom which associates determinism and freedom as the recognition of necessity.

Applied to the institutions of human societies, to morality and law, this may mean that human beings are able to want to change their society's institutions, but there is a natural order which is imposed on human beings by institutions. Human beings may desire to change their institutions and not be able to do it. They find themselves in the position of someone who would like to alter the speed at which he is falling. The universal law of gravitation discovered by Isaac Newton states that two bodies in the universe attract each other with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Human

beings do not choose the values of physical constants. In an analogous manner, all the laws of nature that are the basis of different forms of determinism (as mentioned above) place human beings in this world of necessity. The projects and intentions they conceive are always determined from outside. There is always something that motivates people to do what they do. In this sense, this position is fundamentally empirical. Reasons for acting are externally determined. There is a necessity attached even to the projects expressed in action. Freedom consists precisely in knowing this. Thus, it is only a single step, which takes us from the consciousness of our being subject to necessity, to cynicism.

The Consequences of Determinism With Regard To the Identification of the Conditions of Free Action

Freedom as the Power to Choose The first consequence of determinism is that freedom is defined as opposed to the absence of choice. I have not the choice of speed when I fall. Nor do I have the power to choose not to nourish myself or not to sleep, assuming I want to stay alive. In the world we live in, life only sustains itself by fighting against death (Boulgakov 1912, 2000, Chapter I, II). Basic needs (for sleep and food) must be satisfied; these are conditions of human beings' biological existence. Human beings are in a sense slaves to their own bodies, which have to be looked after. Staying alive is a choice that determines action. One must choose to live in order to choose a particular project of action. The project of survival is conditioned by work, labor, inasmuch as it is a condition of life from an economic point of view. Work is the result of the threat that nature poses to human beings. It is a necessity. It places human beings in a state of needfulness (poverty). This is Adam's curse. But since work allows human beings to get control of basic necessities, it is also the means of extricating oneself from it. Work is a source of redemption, not of enslavement.

The Power to Choose and Social Physics in the Economics of Institutions

The second consequence of determinism is that it leads to the denial of the existence of non-necessitating purposes ("ends," as in Aristotle), in other words the fact that things might be otherwise than they are. In the economics of law and of institutions, this turns out to have important consequences with regard to the way institutional dynamics are modeled. At first, there is a tendency on the part of the theory of cultural evolution to accept a Panglossian economics (Whitman 1998). In such perspective, which is, is rational and what is rational is efficient. In such a world, there is no place for change or for the change agent, the entrepreneur. The same discussion can take place when we apply the law of egoism or the principle of the maximization of profit in order to explain law. Without law, human beings are not at liberty to desire something other than the maximization of profit. This is a given, but not a variable. That which varies is the manner in which individuals exercise their ability to calculate, rather than the objective aimed at by the calculation. The entire art of legislation, as Helvétius had already said (Berlin 2002b, Chapitre Helvétius), is therefore *"de faire que l'individu trouve plus d'intérêt à suivre la loi qu'à la violer."* The entire contribution to the theory of inducements to the legislator is to suppose that the economist can furnish legislators with the means to predict what individuals will do if the rules of the game are changed, that is, if the incentive structure is manipulated, and the division of costs and benefits for each alternative. Such an ambition supposes that the law of egoism always applies and that altruism and disinterested acts are only illusions.

The theory of action teaches us to consider free acts as determinate acts. A free act is founded upon the existence of non-necessitating ends or purposes. Only ends of this type allow human beings to remain entirely free (Gilson 1997, p. 315). The law of gravitation cannot be broken, but human laws, moral or legal, can be involved in a choice. A human law, in fact, is obligatory without being necessary.

The Power to Choose and Nonempirical Approaches in the Economics of Institutions

Affirming the existence of non-necessitating ends

has therefore several consequences. Each of these consequences explains the originality of non-empirical approaches to the law and institutions. (1) The existence of non-necessitating ends first of all restores the determination of the self by the self. This self-determination is a characteristic of the free act. It explains why a certain number of authors insist on the absence of domination (Petit 2001, p. 132) or noninterference (Carter 1999, p. 237) in the definition of freedom. (2) The existence of non-necessitating ends also has the consequence of restoring the entrepreneur to his place as a change agent in the analysis of the dynamics of institutions. The institutional entrepreneur (Yu 2001) does not react to the evolution of constraints (transaction costs), but is at the beginning of his own movement. He acts, he does not react. (3) Non-necessitating ends also rehabilitate human beings' responsibility within history. They modify people's attitude toward reality. When I believe I am responsible for my own destiny and develop a strong feeling of personal effectiveness (*self-efficiency*), I have a tendency to become a change agent (Harper 2003). (4) More generally, the entrepreneur now has space in which to maneuver regarding all the laws that do not establish necessitating ends. If human laws, that is, morality and law, have this characteristic, then human beings can liberate themselves from laws that constrain them by refusing to apply them. They have the power to stand apart from their conditioning. They can overcome internal obstacles to freedom, such as addictive behavior (to tobacco, alcohol, coffee, morphine, opium, cocaine, sexual activity, even tyranny (de la Boétie 1549)). To be free is to be capable of subordinating one's action to the law of duty (a non-necessitating end). In this world of duty, passions and emotions no longer enslave human beings. In the face of danger, a soldier necessarily has a feeling of fear. It is his duty to fight. He must overcome his fear and refuse to flee or hide. The will allows us to overcome inner constraints. The will also allows us to go beyond external constraints, that is, we are able to choose whether or not to have recourse to the formal or informal institutions that structure the social order. These institutions limit a world of possibilities, but they

only define obligations. They are not necessary. It is always possible for people to avoid them or to disobey them (*civil disobedience*) (Thoreau 1849). (5) The last consequence of the existence of such ends is that law and moral rules are not similar to the laws of gravitation, although they claim to be as much in force. So laws that forbid stoning an adulterous woman artificially institute a necessity between two events that are not at all connected in nature. They artificially create determinations by instituting obligations. The fact that it is possible not to conform to them does not make them any less destructive of liberty. It is not because I can leave my own country to escape the oppression of a dictator or taxes that the law protects my power to choose. The word "power" changes its meaning here. In the case of the law of gravitation, "I can not" has a physical meaning. In the case of the law on adultery, "I must not" in a moral sense. I do not have the right. The law artificially determines the limits of my power to choose. It limits the infinite freedom of the will and leads to the study of the concrete freedom possessed by human beings in society (Hegel 1821, §4, §30).

Respect for the Principle of Self-Determination by the Self

The Ideal of Contractual Law

It is not only laws of nature that limit freedom. There are also human laws, morality, and the law of the courts. The origin of this limit has to do with a confrontation between two wills. The infinity of free will, that is, the possibility of willing, even the impossible becomes concrete freedom when two wills (Hegel 1821) or two individual claims (*pretesa/pretesan*, Léoni 1961) oppose each other. This explains why possession becomes property and takes on a legal character to the extent that the other, or all others, recognize that a thing I have made mine is mine, as I recognize others' possessions as their own. Concrete freedom that is expressed through law partially originates in a process of mutual recognition (Facchini 2002), the other part dealing with connections of a contractual type. Law guarantees the conditions of the free will if it is the result of that confrontation of wills, which accept the task of mutually

limiting one another all in taking account of the wills of others. Law limits the power to choose. It exercises a constraint without for all that violating the principle of self-determination by the self. This represents the fact that human beings can limit their own power to choose through laws that they freely support and which they apply thanks to the implementation of trust rules and solidarity rules (Vanberg and Buchanan 1990). These laws constitute an order based on rules (order of rules, Hayek 1973) that change as a function of the relationships between the wills of different members of a group. Concrete freedom is based on consent. Outside the contract, law becomes the death of liberty, for it is imposed against the will of human beings.

Freedom, Tyranny, and Paternalism

In the great conflict of wills, the other may also decide on a constraint for me to labor under. The law, here, is chosen against my will by another will than mine. The principle of determination of self by self is no longer being respected. Law is no longer creating the conditions necessary for the existence of a free act recognized as such by everyone. Only the one who produces the law is free. For my own happiness, he subjects me to his will (paternalism) – or if it is for his own happiness, he is a tyrant.

The tyrant has the face of a bad person in Pascal's sense (1650, 1982, p. 125). The bad person has power and uses his force to impose his will on me. The confrontation of wills no longer can be solved through contracts, but only through the application of the law of strength, of the stronger party. The stronger party will oppress the weaker (Pascal 1650, 1982, p. 127). This transforms a factual situation into a law. What the strong possess is transformed into property rights. This means that at the beginning there is no agreement about the rights of each, but there is usurpation (Pascal 1650, 1982, p. 125). The law makes the strong free and places the weak in a situation of necessity, because freedom without power is impotent. Under these conditions the infinite freedom of the will of the weak will never receive a proper concrete expression in the law. The weak can band together to overthrow the

strong and impose their own laws, but they will never be liberated unless they reverse the relationships of force, making yesterday's strong people the weak and the oppressed of yesterday the strong. The law is necessarily that of the strongest, if without force the will is powerless. The social and political conditions of a free society will never all be present. The infinite freedom of the will is therefore only an illusion, since law is always the law of the stronger. He supports the freedoms of some but oppresses the freedom of others.

The legislator can also assume the form of a benevolent father. The strong man ceases to be a tyrant. He places his power in the service of the Good. The original form of paternalism consists in helping individuals to keep their promises, that is, to make the will of the weak-willed strong. Let us suppose that a weak person does not permit himself to commit adultery, but through the weakness of his will, he succumbs to temptation just the same. Such a situation may justify intervention on the part of the strong man. His intervention will be like the chains holding Ulysses to the mast as the Sirens sing, in Homer (Elster 1984). A second form of paternalism consists in deciding on the extent of the means that individuals give themselves in order to realize their goals. A man who wishes to be in good health should not smoke. The strong person can justify constraint with reference to the incoherence of the weak person. The weak person will not be allowed to smoke, as a means of helping that person reach their personal goals. A third form of paternalism, *soft paternalism*, prohibits nothing and uses no force but tells weak or poorly informed people about the risks various behaviors are associated with. They remain free to do as they like. But they are obliged to hear out the morality of the strong. A fourth, hard paternalism, is moral. It determines the ends of action not because the strong man wants it that way, but because the good can be objectively determined. The strong man knows what is good and seeks to promote it through politics, in which the ultimate aim is the happiness of men in society (Humboldt 1851, III). In the name of this principle, this stance gives itself the freedom to act as a tyrant in the name of the Good. Thus, we have here to do with a benevolent tyrant.

The Ideal of Contractual Law and the Role of the State

The introduction of the figure of the strong person in a contest of wills is equivalent to a consideration of the role of the State in the formation of law and the protection of individual liberty (freedom).

The law that is generated by the benevolent attitude of the strong person changes as a function of the strong person's knowledge of that which is good. It establishes the strong person as a legislator and exposes society to two kinds of risk. Happiness from this perspective dominates the principle of non-domination of one individual by another. The legal conditions of a free act are not guaranteed. The law of the legislator then risks becoming unstable because it evolves as a function of what the legislator learns concerning what is to be done or not done in order to bring about the happiness of human beings in society. This instability of the law reduces the quality of what agents are able to anticipate and increases the cost of their coordination. It opens the door to higher costs for political transactions, since everyone is attempting to influence the decisions of the legislator and to impose their own conception of the Good. The law of the legislator, in addition, no longer mobilizes the group of kinds of tacit knowledge that agents have when they are constrained only on a contractual basis. The law therefore has a good chance of being poorly adapted to many particular situations and for this very non-applied reason.

The law of the tyrant serves the tyrant's will. It includes the tyrant's conceptions of the Good. Therefore, it also has an unstable, arbitrary, and incomplete nature.

The law of a contractual nature is to the contrary freely consented to and based upon the tacit knowledge of agents. It may nonetheless be unstable, because it is never certain that one of the parties to a contract may not decide at one moment or other to refuse to keep his or her promises. A person may in fact decide that he or she is in a position of strength and that it is no longer in that person's interest to continue to be bound by agreements that were freely agreed to in earlier negotiations. This risk is real. It is normally limited by

the existence of rules involving confidence and solidarity that are made specifically to prevent such behavior by instituting dissuasive mechanisms such as guilt, shame, a bad reputation, exclusion, and/or ostracism. All the individuals of the group band together against the deviant, that is, the individual who does not wish to keep his or her word. If these rules are sufficient, the State has no role. Its existence is nothing but a useless fiction through which the freedom of some people is extended to the detriment of others' freedom, that is, through which "*tout le monde s'efforce de vivre aux dépens de tout le monde*" (Bastiat 1863, Tome IV, pp. 327–341). Only when the State arrogates to itself a monopoly on violence and the production of the law (Léoni 1961) does law become the law of the strongest. If risk is not obviated through the application of these rules and mechanisms for imposing sanctions, human beings may have recourse to force, in other words may make agreements so that power ensures the maintenance of law and order. The State is a regalian State. It ensures the enforceability of contracts against external enemies and/or internal strife (Humboldt 1851, 1969, IV, p. 45). This means taxation is one condition of a free society, for it is the condition for the financing of the operations of the police and for the defense of the State's boundaries. The financing of the police from this perspective is the single issue regarding political freedom. These freedoms guarantee to the individual the power to choose his level of taxation, his rules of allocation, and the people who will manage everything.

References

- Bastiat F (1863) Collected works of Frédéric Bastiat in 6 volumes, drawn mainly from the 7 volume Guillaumin edition published in the 1863, *Oeuvres Complètes*
- Berlin I (2002a) Two concepts of liberty. In: Hardy H (ed) *Liberty*. Oxford. pp 166–217
- Berlin I (2002b) Freedom and its Betrayal. *Six Enemies of Human Liberty*, edited by Henry Hardy with a new foreword by Enrique Krauze, Princeton University Press
- Boulgakov S (1912, 2000) *Filosofia Hozaïstava*, translated, edited and with an introduction by Catherine Evtuhov. Yale University Press, New haven/London

- Carter J (1999) A measure of freedom. Oxford University Press, Oxford
- Cowan R (1994) Causation and genetic causation in economic theory. In: Boettke PJ (ed) *The Edward Elgar companion of Austrian economics*. Edward Elgar, Aldershot, Brookfield, pp 63–71
- Dawkins R (1976) *The selfish gene*. Oxford University Press, New York
- de la Boétie E (1549) *Discours de la servitude volontaire*, Paris, Flammarion, réédition 1933, *The politics of obedience: the discourse of voluntary servitude*, translated by Harry Kurz for the edition that carried Rothbard's introduction. Free Life Editions, New York 1975
- Diamond J (1997) *Guns, germs, and steel*. W.W. Norton, New York
- Elster J (1984) States that are essentially by products. In: *Imperfect rationality: Ulysses and the sirens*. Cambridge University Press (Part II), Cambridge, New York, Melbourne, Madrid, Singapour
- Facchini F (2002) Complex individualism and legitimacy of absolute property rights. *Eur J Law Econ* 13:35–46
- Gilson E (1997) *Le Thomisme*. Introduction à la philosophie de Saint Thomas d'Aquin, études de philosophie médiévale, librairie philosophique J. Vrin, Paris
- Harper D (2003) *Foundations of entrepreneurship and economic development*. Edward Elgar, London/New York
- Hayek F (1973) *Law, legislation and liberty*, vol 1. University of Chicago Press, Chicago
- Hegel, GW (1821) *Grundlinien der Philosophie des Rechts*, elements of the philosophy of right, tr Knox TM, 1942; tr Nisbet HB (eds) Allen W. Wood, 1991. Oxford University Press, Oxford
- Leoni B (1961) *Freedom and the law*. Expanded 3rd edn, foreword by Arthur Kemp Liberty Fund 1991, Indianapolis. http://oll.libertyfund.org/?option=com_staticxt&staticfile=show.php%3Ftitle=920&Itemid=27
- Montesquieu, L'Esprit des lois, *The spirit of laws*, translated by Thomas Nugent, revised by Prichard JV, based on an public domain edition published in 1914 by G. Bell & Sons, London
- Pascal B (1650, 1982) *Pensées*, Jean de Bonnot, Paris, établi à partir de la dernière édition de Léon Brunschvicg de 1951, translation in english first published 1966 revised edn 1995, London/New York
- Pettit P (2001) *A Theory of freedom: from the psychology to the politics of agency*, Oxford University Press, Oxford, New York
- Sen A (1999) *Development as freedom*. Random House, New York
- Thoreau D (1849) *Resistance to civil government: a lecture delivered in 1847*. Aesthetic papers, selected, edited, and published by Elizabeth Palmer Peabody <https://archive.org/details/aestheticpapers00peabrich>
- Vanberg V, Buchanan J (1990) Rational choice and moral order. In: Nichols J Jr, Wright C (eds) *From political economy to economics and back?* Institute for Contemporary Studies, San Francisco
- Voltaire, *Le philosophe ignorant*, Chapitre 13, édition R. Mortier, *Les œuvres complètes de Voltaire*, tome 62, p.44, Voltaire Foundation, 1987, Oxford
- von Humboldt W (1851) *Ideen zu einen Versuch die Grenzen der Wirksamkeit des Staats zu bestimmen*, English translation *The limits of state action*, edited with an introduction and Notes by Burrow JW, Cambridge University Press, 1969, Cambridge/New York, etc. http://oll.libertyfund.org/?option=com_staticxt&staticfile=show.php%3Ftitle=589&chapter=45492&layout=html&Itemid=27
- Whitman DG (1998) Hayek Contra Pangloss an evolutionary systems. *Constit Polit Econ* 9:45–66
- Yu Tony Fu-Lui (2001) An entrepreneurial perspective of institutional change. *Constit Polit Econ* 12:217–236

Lies and Decision Making

Alessandro Antonietti¹, Barbara Colombo² and Claudia Rodella²

¹Department of Psychology, Catholic University of the Sacred Heart, Milan, Italy

²Department of Psychology, Catholic University of the Sacred Heart, Brescia, Italy

Abstract

Within the economics and law fields, many decisions must be based on what is reported by another person, who can deceive the decision maker by lying. Thus, discovering if our interlocutor is sincere or not is crucial in order to make good decisions. Research highlighted that the spontaneous strategies that we use to identify possible lies are often misleading. Our moods and personality, together with the level of trust between speakers, are all factors that can influence the detection of lies. However, the ability to discover lies may increase with appropriate training and experience of dealing with people in contexts where the probability of being deceived is quite high. Regardless of our actual skill in discovering lies, our attitude toward lying can influence the decision-making process. For example, when we are aware of the possibility that a lie occurs, a suspicious attitude can lead to a wrong judgment. Similarly, perceiving an alleged lie, whether it is real or not, can prompt the use of

emotional heuristics linked to the perceived feelings of antipathy, anxiety, or anger. This can lead to decisions aimed at creating disadvantages for the partner. Under these circumstances, it is necessary to take into account the variety of human behavior, not relying on stereotypes to identify a lie. Being used to interact in particular contexts where the risk of being deceived is high may definitely help to sharpen the ability to find out who is lying to us, increasing the likelihood of taking rational choices based on reliable cues.

Lying in the Field of Law and Economics

Within the economics and law fields, many decisions must be based on what is reported by another person. This happens, for example, when a judge has to issue a sentence on the basis of statements made by the witnesses or when a broker has to decide to invest money by evaluating the reliability of a company on the basis of what an analyst tells him about it. In these situations, it is generally assumed that the other party is sincere. However, people may lie, for various reasons: for personal interest, to defend the interests of others, for idealistic reasons, and so on. Therefore, it becomes important to know how to identify when a person is lying and manage the decisions accordingly.

Psychological knowledge can be helpful. Although there is no “truth machine” that is able to establish with certainty when one is lying, there are research data that provide guidance in this regard. The studies which have been carried out allow us to identify what causes people to lie, which are the situations when this is more likely to happen, what are the personal characteristics that distinguish liars, and how lying may affect ethical and economic situations. Being aware of these aspects may allow one to take appropriate decisions when other people lie.

Why People Lie

Lying occurs when there is no correspondence between what one says and what he thinks,

knows, and feels. This can occur when the individual does not have full knowledge of how things really are or when his/her communication is not adequate. In these cases, the person tells a lie, but it is a falsehood due to ignorance or error. Lying, therefore, has more to do with the *truthfulness* (i.e., what the person believes to be truthful) than with the *truth* per se. We are lying, as we are discussing here, when people deliberately attempt to induce others to believe that things are different from what we believe they actually are.

All of us are aware of how we tend to lie during our everyday life. We lie for many different reasons, from the most trivial to the most important ones. Studies showed that during a normal conversation people make use of deceitful assertions in 61.5% of cases (Turner et al. 1975). In general, we tell minor or white lies, which require a limited mental commitment to be planned and communicated and do not cause excessive stress to the actor because, even if the lie is discovered, the consequences will be not heavy. In other situations, however, lying can have serious consequences on the decisions that are taken, and this happens quite often when economic or legal aspects are involved.

Despite some research supporting the notion that the number of lies told decreases with age (Jensen et al. 2004), other studies showed that lying is an essential part of communication among adults (Camden et al. 1984; DePaulo and Kashy 1998; DePaulo et al. 1996; Hample 1980; Turner et al. 1975). Actually, what changes with adulthood is the social desirability of a lie, which is less accepted as a suitable mean to achieve a goal. In addition, children are more naïve and believed to be always credible, even if their lies are less elaborated and complex. This is why it is usually much easier to discover a child lying, compared to an adult, who has many more resources and an advanced ability to lie. For this reason, we may wonder if children actually lie more or if they are simply discovered more often and have less fear of confessing.

Many theorists argued that deceitful communication is linked to survival: men used to lie to get what they needed when they were lacking of resources. Like other behaviors that are

maintained over time for this reason, lying has a different meaning today. We do not lie only to obtain the resources needed for our survival but also to obtain superfluous goods (tangible and intangible), to look better, to deceive others, to protect those we care about, and so on.

Survival is also used to explain why people believe the lies. If we had a cautious attitude toward what others say in every moment of our lives, we would run the risk of spending a lot of time and energy in order to assess the evidence of the communication of others, in order to determine the authenticity of what they are saying. People would be too suspicious, not allowing social relationships to be lived fully and peacefully.

Universally speaking, human beings tend to give credit to what others say. This trend presents two levels: on one hand, if the speaker does not have any particular reason to lie, he/she will say what he/she thinks is true; on the other hand, if the listeners do not have any particular reason to doubt the speaker, they will accept as true what the speaker communicates. These statements refer to Grice's principle of cooperation: according to this principle, you have to "make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice 1989, p. 26). This truth bias is an important cognitive heuristic, i.e., a system that allows one to evaluate complex stimuli with a reduced cognitive commitment (Caldwell 2000). This is only one of a number of heuristics that humans commonly use. Another example can be the heuristic according to which it is easier to falsify facts than feelings: therefore it is more likely that people evaluate as false factual statements than emotional utterances.

Identifying Lies

Human beings tend to overestimate their ability to identify a false communication. In fact, decades of research have shown that humans are modest debunkers of the lies (Hartwig and Bond 2011). One main reason why humans have limited skills

in detecting lies is related to the lack of certain, absolute, and universal hints that can be used to identify lies. The second reason is related to the high level of complexity of human communication. Even when we tell the truth, we use a wide range of communication strategies. This means that when we want to lie, we just have to slightly change part of the communication configuration. As a consequence, differences between truth and falsehood are difficult to be detected. Thirdly, people have stereotypical theories on liars. Since such theories are not based on empirical evidence, it is not surprising that they can easily lead people to errors of judgment (believing that one is lying when he/she is telling the truth or thinking he/she is sincere when in fact he/she is a liar). As a last point, as mentioned above, social conventions that guide interpersonal relationships induce individuals to not have a constant suspicious and inquisitive attitude. Doubting everything and constantly accusing others of being a liar would prevent any relationship of intimacy and trust to develop.

It is interesting to note that there are some categories of people who are better able at discovering liars. This is the case of criminals, spies, secret services' employees, and those who are well trained to lie or have to deal with lies on a daily basis, as well as those whose life depends on their lying skills (Vrij and Semin 1996). Moreover, some professionals are more experienced than others, such as police officers and clinical psychologists (Ekman et al. 1999). Different studies (DePaulo 1994; DePaulo and Pfeifer 1986; Ekman and O'Sullivan 1991; Kraut and Poe 1980; McCornack and Levine 1990; Rosenthal and DePaulo 1979) reported that the experience in dealing with lies increases the confidence in the ability to identify them. This, however, does not correspond to an actual accuracy of the assessment. The only exception seems to be the American secret services (Ekman and O'Sullivan 1991). The possible influence of a deep relationship – like the one between partners, family members, or close friends – in detecting lies has also been explored. However, even in this specific case, the only variable that changes significantly is the declared level of confidence in being able to discover the lies of those who are

close to us. Yet, this does not show any positive correlation with the accuracy of the judgment (McCornack and Levine 1990). Rosenthal and DePaulo (1979) also investigated possible differences due to gender. The only significant result they reported is that men tend to be more suspicious than women, even if this does not mean that they are more skilled at exposing a lie. Toris and DePaulo (1984) conducted a study in order to verify if people were better at spotting a lie if they were notified of the possibility that they were lied to. Results showed that individuals become more suspicious, believing more people to be liars. Once again, yet, they were not more accurate in their judgments.

Starting from this evidence, we can conclude that a direct discovery of falsehood is impossible, mainly because we do not have the ability to read the minds of other people. However, we are potentially able to detect lying in an indirect way, relying on more or less reliable indices, although this ability is very complex and is usually rare to possess without a specific training. It is important to remember that no tool and no method is fool-proof and that the best procedure involves the integration of multiple systems (analysis of non-verbal behavior and of content and consistency of communication, attention to contextual and cultural cues as well as to speaker's personality, and so on). Along this line, it has been proved that even when there are clues that are sufficiently clear to detect lying, most people do not use them (O'Sullivan 2009). However, it is important to stress that a proper training can significantly increase the ability to recognize clues that are associated with lying and to discern between truthful and false discourses. Paul Ekman has been studying for many years the most relevant behavioral clues related to lying. His studies helped him in developing a program to teach people to be able to recognize the micro-expressions that, within a communication flux, provide information about how the other speaker is going to behave. The micro-expressions are particularly useful to predict threatening and/or dangerous behaviors (Ekman 2009). Micro-expressions are closely linked to emotions. Ekman himself, along with other authors, showed

that aphasic patients are particularly sensitive to these clues and this allows them to detect more accurately a lie (Etcoff et al. 2000). Emotions can challenge the liar, too: he/she may be betrayed by his/her emotions or may fail in the construction of the lie because of the influence of emotions. These two errors lead to qualitatively different behavioral indices as they are linked to different emotions: fear of being caught rather than guilt (Ekman and Frank 1993).

Deciding When Exposed to Possible Lies

We can assume that when we are facing a situation where there is the possibility that the person we are talking with is lying, we must determine whether we can trust him/her before taking any decision (Riva et al. 2014). To do this, we have to make inferences about his/her intentions. This can be done either through an immediate and holistic process or through a slower and analytical one (Iannello and Antonietti 2008). In other words, this may occur either through a rapid intuition or through a detailed and systematic examination of the person and the situation at hand. In the first case, the process resembles the formation of impression, whereas in the second case, the process involves a more logical assessment (Iannello et al. 2014). In this respect, intuition and analysis are conceived as different decision-making approaches (Stanovich and West 2000).

Slovic and colleagues (2002) have proposed a decision model based on the presence of two interacting cognitive processes: the analytical system, which is based on rules and on the decision maker's explicit control, and the intuitive system, based on impressions that arise automatically, without any special effort or intention. This system is activated immediately, together with the first reaction to a stimulus, which is often an affective reaction. In this regard, it is relevant to point out how the intuitive-affective reaction plays a key role in the perception of risk and benefit: if the people "like" a stimulus, they tend to underestimate the risks and overestimate the benefits; if they "do not like" it, they will probably assess the risk as very high and while considering

the benefits as being quite low. The process of judgment based on the affect feeling associated with the stimuli has been called by Slovic *affect heuristic* and is considered the key component of an intuitive decision.

Kahneman (1994) is another author who distinguished the choices based on emotional feeling (choosing by liking) from the comparative analysis of options typical of the normative choices (choosing by dominance). Whereas the last typology of choices considers the nature of the different options, the first one, based on the pleasantness, is mostly determined by emotional feelings associated with the alternative choices. We activate this strategy especially when lacking information to perform an effective assessment of the value of the stimulus (Hsee et al. 2005).

Some studies have shown that the individual mood may influence the decision-making process (Schwarz 2002). When we experience a negative mood, the decision-making process takes longer and is characterized by a greater attention to each attribute. When we are in a positive mood, instead, we tend to evaluate with greater shallowness, increasing the use of intuitive strategies.

An Example of a Complex Approach to Decision Making and Lying

As we have seen, when we have to take a decision while we may be deceived because our partner is lying, we tend to use both the intuitive processes (through which we try, by relying on our impressions and emotions, to understand if our partner is trustworthy) and the analytical processes (through which we examine the information we have about our partner from a rational standpoint). Personal characteristics play a major role, since they lead an individual to rely more on the former or the latter of the two kinds of processes. This specific decision-making process is therefore a complex process that requires a complex approach to be appropriately investigated. As an example of a psychological investigation of this process run by applying a complex approach, we report the case of a recent study (Colombo et al. 2013).

The experimental study of the relationships between lies and decision making requires the use of simplified tasks with respect to the daily life situations. In this specific study, a variant of the Ultimatum Game has been used. The Ultimatum Game (Powell 2003) is a task where a subject A (proposer) is given a sum of money and is asked to split it with another subject B (responder). B has the choice to accept or reject the proposed division. If the proposal is accepted, the split of the money becomes effective; if it is refused, both players receive nothing. This task involves both moral (fairness of the split) and economic (maximizing personal benefit but also reciprocity) considerations. According to the theory of rational choice, B should accept any sum A proposes, following the logic “better than nothing.” Actually, research shows that many unfair offers are declined according to the principle of aversion to inequality. B, therefore, generally does not accept an offer that falls below half the amount.

In the experiment we are discussing, the participants, playing as proposers, saw six videos where six different people (the responders), balanced by gender and age, gave information about themselves. Information sometimes was false. In addition to deciding how to split the amount of money with the different responders, subjects were asked to express an opinion on the truthfulness of what each partner said while introducing himself/herself. Each character gave the same amount of relevant (e.g., “I am person who does not compromise”) and nonrelevant (e.g., “I believe in values such as friendship and family”) information about himself/herself. Only half of the sample has been given prior information concerning the possibility that respondents may have been lying. While the subjects performed the task, eye movements were recorded using an eye tracker, a noninvasive tool that uses infrared technology to study the relationship between eye movements and information processing. Before the experiment, participants were tested to assess specific personality traits (such as impulsivity) and their preferred decision-making style (whether intuitive or analytical).

A first result concerns how people tend to consider information they receive. Participants

judged as true what partners said about themselves in the majority of cases, confirming what already reported in the literature. Another data that corroborates findings of previous research is the difficulty in identifying a false communication. The subjects failed in this task in about half the cases. The type of instructions received, however, affected the behavior. As mentioned above, half of the sample did not receive any information about the possibility that some of the respondents could lie, whereas the other half was aware of this possibility. As expected, the subjects who were aware of the possible presence of liars changed their behaviors: on one side, they seemed to be more suspicious, while on the other side, this awareness seemed to promote positive feelings toward the partner who has been considered sincere. In any case, however, knowing about the possibility of being deceived did not improve the performance in terms of accuracy of the judgment of truthfulness.

Data also showed that impulsivity activates stereotypes – as is also reported in other studies (e.g., Baldi et al. 2013) – or distracts from the search of clues of deception, leading to bid higher to the responders. More reflexive people are more meticulous and more responsive to information that they receive. In this specific experiment, this led to propose a lower sum of money to the respondents. The influence of personality seemed to result in a greater emphasis on the perception of similarity or difference with respect to respondents. When participants were able to “get into the shoes” of the respondent or somehow considered him/her similar to themselves (maybe for similar age and same gender), people adopted a more rational approach in splitting the money. From the opposite perspective, the greater the perceived distance from the responder, the greater the difficulty to make a logical decision.

Another hypothesis of this study concerned the fact that the visual behavior could be unconsciously influenced by the variables mentioned above. Data from the eye tracker highlighted that the main attentional focus was always on the “person” (eyes, mouth, face, and torso) when individuals were looking for a lie. They seemed to focus mainly on the eyes, probably because

they are considered as a point of reference to test whether the communication addressed to them is genuine or false.

Summarizing the data obtained from this study, it can be concluded that human being is by nature inclined to believe a communicational partner to be sincere and that, even if he/she is alerted about the possibility of receiving a false communication, he/she does not become more skilled or accurate in spotting a lie. Specific characteristics linked to personality and decision-making style appeared to play an important role: a greater impulsivity leads to the activation of a more automatic reasoning, diminishing the possibility of using a logical and rational thought.

In the same study, Colombo and colleagues (2013) also tried to modulate the choice through a technique of noninvasive brain stimulation, using the transcranial direct current stimulation (tDCS), which was applied on the prefrontal cortex. This area is well known for being involved in decision making. In previous studies (Antal et al. 2007; Kuo et al. 2013; Lang et al. 2004; Nitsche and Paulus 2001), it was reported that the application of this kind of stimulation is associated with a modulation of cortical excitability, which leads to an inhibition or activation of the stimulated area. The prefrontal area is specifically linked to inhibitory mechanisms (Bembich et al. 2014; Ding et al. 2014; Lee et al. 2014), so it was expected that the stimulation of this area could induce a change in the level of impulsivity and, consequently, in promoting the intuitive rather than the analytical approach (Iannello et al. 2014). Indeed the subjects, as a result of the stimulation of the prefrontal cortex, exhibited significantly different behavior. To be more specific, they seemed to become more rational. In the control (namely, no stimulation) condition, irrational behaviors emerged, like offering no money to the respondent to “punish him” for being dislikeable, not considering that acting this way they would not get anything, since the responder would obviously reject this partition. The stimulation of the prefrontal cortex, instead, led participants to rely on the relevant data supplied by the respondents, making the most of each detail.

Conclusions

Lying is something we deal with on a daily basis, even though most of the times we do not invest energy in order to discover either our interlocutor is sincere or not. Whenever we communicate with other people, sometimes we wonder if they are lying, because we are aware that lying is a characteristic of human beings. Nevertheless, the automatic and spontaneous hypotheses that we generate to assess the truthfulness of a statement are often misleading and do not lead to an accurate assessment. Our moods and personality, together with the level of trust between speakers, are all factors that can influence the detection of lies. Similarly, a careful assessment of information and how it is communicated does not guarantee the accuracy of judgment. Having said that, the ability to discover lies may increase with appropriate training and experience of dealing with people in contexts where the probability of being deceived is quite high.

In those fields – like law and economics – where the analysis of information received is critical, it is crucial to understand how the perception of lying can influence the judgment and, ultimately, the decision-making process. Regardless of the actual presence of a deceitful communication, when we are aware of the possibility that it occurs, a suspicious attitude can lead to a wrong judgment. Similarly, perceiving an alleged lie, whether it is real or not, can prompt the use of emotional heuristics linked to the perceived feelings of antipathy, anxiety, or anger. This can lead to decisions aimed at creating disadvantages for the partner. Under these circumstances, it is necessary to take into account the variety of human behavior, not relying on stereotypes to identify a lie. Being used to interact in particular contexts where the risk of being deceived is high may definitely help to sharpen the ability to find out who is lying to us, increasing the likelihood of taking rational choices based on reliable cues.

Cross-References

► [Credibility](#)

► [Efficiency](#)
 ► [Good Faith and Game Theory](#)

References

- Antal A, Terney D, Poreisz C, Paulus W (2007) Towards unravelling task-related modulations of neuroplastic changes induced in the human motor cortex. *Eur J Neurosci* 26:2687–2691
- Baldi PL, Iannello P, Riva S, Antonietti A (2013) Socially biased decisions are associated to individual differences in cognitive reflection. *Stud Psychol* 55:265–271
- Bembich S, Clarici A, Vecchiet C, Baldassi G, Cont G, Demarini S (2014) Differences in time course activation of dorsolateral prefrontal cortex associated with low or high risk choices in a gambling task. *Front Hum Neurosci* 24(8):464
- Caldwell S (2000) Romantic deception – the six signs he’s lying. Adams Media Corp., Holbrook
- Camden C, Motley MX, Wilson A (1984) White lies in interpersonal communication: a taxonomy and preliminary investigation of social motivations. *West J Speech Commun* 48:309–325
- Colombo B, Rodella C, Riva S, Antonietti A (2013) The effects of lies on economic decision making. An eye-tracking study. *Res Psychol Behav Sci* 1(3):38–47
- DePaulo BM (1994) Spotting lies: can humans learn to do better? *Psychol Sci* 3(3):83–86
- DePaulo BM, Kashy DA (1998) Everyday lies in close and casual relationships. *J Pers Soc Psychol* 74:63–79
- DePaulo BM, Pfeifer RL (1986) On-the-job experience and skill at detecting deception. *J Appl Soc Psychol* 16:249–267
- DePaulo BM, Kashy DA, Kirkendol SE, Wyer MM, Epstein JA (1996) Lying in everyday life. *J Pers Soc Psychol* 70:979–995
- Ding WN, Sun JH, Sun YW, Chen X, Zhou Y, Zhuang ZG, Li L, Zhang Y, Xu JR, Du YS (2014) Trait impulsivity and impaired prefrontal impulse inhibition function in adolescents with internet gaming addiction revealed by a Go/No-Go fMRI study. *Behav Brain Funct* 30:10–20
- Ekman P (2009) Telling lies: clues to deceit in the marketplace, politics, and marriage. WW Norton, New York
- Ekman P, Frank MG (1993) Lies that fail. In: Lewis M, Saarni C (eds) *Lying and deception in everyday life*. Guilford Press, New York, pp 184–200
- Ekman P, O’Sullivan M (1991) Who can catch a liar? *Am Psychol* 46:913–920
- Ekman P, O’Sullivan M, Frank MG (1999) A few can catch a liar. *Psychol Sci* 10(3):263–266
- Etcoff NL, Ekman P, Magee JJ, Frank MG (2000) Lie detection and language comprehension. *Nature* 405(6783):139–139
- Grice HP (1989) *Studies in the way of words*. Harvard University Press, Boston
- Hample D (1980) Purposes and effects of lying. *South Speech Commun J* 46:33–47

- Hartwig M, Bond CF Jr (2011) Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychol Bull* 137:643–659
- Hsee CK, Rottenstreich Y, Xiao Z (2005) When is more better? On the relationship between magnitude and subjective value. *Curr Dir Psychol Sci* 24:234–237
- Iannello P, Antonietti A (2008) Reciprocity in financial decision making: intuitive and analytical mind-reading strategies. *Int Rev Econ* 55:167–184
- Iannello P, Colombo B, Antonietti A (2014) Non-invasive brain stimulation techniques in the study of intuition. In: Sinclair M (ed) *Handbook of research methods on intuition*. Edward Elgar, Northampton, pp 130–143
- Iannello P, Balconi M, Antonietti A (2014) Analytical versus intuitive decision making modulates trust in e-commerce. *Neuropsychol Trends* 16:31–49
- Jensen LA, Arnett JJ, Feldman SS, Cauffman E (2004) The right to do wrong: lying to parents among adolescents and emerging adults. *J Youth Adolesc* 33:101–112
- Kahneman D (1994) New Challenges to the Rationality Assumption. *J Inst Theor Econ*, 150(1):18–44
- Kraut RE, Poe D (1980) Behavioral roots of person perception; the deception judgments of customs inspectors and laypersons. *J Pers Soc Psychol* 39:784–796
- Kuo HI, Bikson M, Datta A, Minhas P, Paulus W, Kuo MF, Nitsche MA (2013) Comparing cortical plasticity induced by conventional and high-definition 4×1 ring tDCS: a neurophysiological study. *Brain Stimul* 6:644–648
- Lang N, Siebner HR, Ernst D, Nitsche MA, Paulus W, Lemon RN, Rothwell JC (2004) Preconditioning with transcranial direct current stimulation sensitizes the motor cortex to rapid-rate transcranial magnetic stimulation and controls the direction of after-effects. *Biol Psychiatry* 56:634–639
- Lee I, Byeon JS (2014) Learning-dependent changes in the neuronal correlates of response inhibition in the prefrontal cortex and hippocampus. *Exp Neurobiol* 23:178–189
- McCornack SA, Levine TR (1990) When lies are uncovered: Emotional and relational outcomes of discovered deception. *Communication Monographs* 57:119–138
- Nitsche MA, Paulus W (2001) Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. *Neurology* 57:1899–1901
- O’Sullivan M (2009) Why most people parse palter, fibs, lies, whoppers, and other deceptions poorly. In: Harrington B (ed) *Deception: from ancient empires to internet dating*. Stanford University Press, Berkeley, pp 74–91
- Powell K (2003) Economy of the mind. *PLoS Biol* 3:312–315
- Riva S, Monti M, Iannello P, Pravettoni G, Schulz PJ, Antonietti A (2014) A preliminary mixed-method investigation of trust and hidden signals in medical consultations. *PLoS ONE* 9(3):e90941. <https://doi.org/10.1371/journal.pone.0090941>
- Rosenthal R, DePaulo BM (1979) Sex differences in eavesdropping on nonverbal cues. *J Pers Soc Psychol* 37:273–285
- Schwarz N (2002) Feeling as information: moods influence judgments and processing strategies. In: Gilovich T, Griffin D, Kahneman D (eds) *Heuristics and biases: the psychology of intuitive judgment*. Cambridge University Press, New York, pp 534–547
- Slovic P, Finucane ML, Peters E, McGregor DG (2002) The affect heuristic. In: Gilovich T, Griffin D, Kahneman D (eds) *Heuristics and bias: the psychology of intuitive judgment*. Cambridge University Press, New York, pp 397–420
- Stanovich KE, West RF (2000) Individual differences in reasoning: implications for the rationality debate? *Behav Brain Sci* 23:645–665
- Toris C, DePaulo BM (1984) Effects of actual deception and suspiciousness of deception on interpersonal perceptions. *J Pers Soc Psychol* 47:1063–1073
- Turner RE, Edgley C, Olmstead G (1975) Information control in conversations: honesty is not always the best policy. *Kans J Sociol* 11:69–89
- Vrij A, Semin GR (1996) Lie experts’ beliefs about non-verbal indicators of deception. *J Nonverbal Behav* 20:65–80

Lighthouses

Elodie Bertrand

CNRS and University Paris 1 Panthéon-Sorbonne (ISJPS, UMR 8103), Paris, France

Abstract

From being a symbol of the necessity for publicly financing public goods, since Coase’s 1974 article “The lighthouse in economics” the lighthouse has instead become a symbol of the private sector’s ability to provide public goods, and subsequently an archetype of a mistaken economic policy justified by “market failures.” However, recent studies in economics, law, and history have uncovered different mixes of private and public finance in the provision of lighthouse services over time, exhibiting the practical problems involved in providing public goods.

Introduction

The lighthouse has long been cited as a locus for public intervention – at least since John Stuart

Mill (1848, p. 968), who referred to the difficulty of obtaining payments for such services. The example was taken up by Henry Sidgwick (1901, p. 406), who saw it as a free-rider problem (explained by a failure of appropriation). Arthur Cecil Pigou (1932, p. 184) reinterpreted it as a problem in which the social product is greater than the private product (nowadays called a positive externality). Paul Samuelson (1964, p. 159) cited the lighthouse as an example of external effect and public good (absence of exclusion and of rivalry). Kenneth Arrow (1969, pp. 146–147) added that even if these problems were solved, there would remain a problem of bilateral monopoly in the absence of competitive prices.

But in 1974 Ronald Coase described a system of lighthouse financing in England and Wales, from the sixteenth to the nineteenth century, in which private individuals embarked on financing, building, and maintaining numerous lighthouses, and obtained payments for this service. He thus laid claim to the lighthouse as an example of a standard, mistaken, approach in economic policy and, more generally, of economists' lack of concern for the real world. The lighthouse thus became the classic example of "the failure of market failure" (Zerbe and Mc Curdy 1999).

While widely repeated, Coase's conclusion was not totally accepted by Samuelson (who still cites the lighthouse as an example of a public good – see Samuelson and Nordhaus 2010), and it has systematically been called into question by lawyers, historians, and economists (Van Zandt 1993; Taylor 2001; Bertrand 2006): the old English system, it is argued, was neither truly private nor really efficient. This has renewed the debate over the nature and efficiency of the English lighthouse system among economists and historians (Barnett and Block 2007, 2009; Bertrand 2009; Block 2011; Carnis 2013, 2015; Lindberg 2013; Levitt 2016), and studies of other historical systems have flourished (Bertrand 2005; Lai et al. 2008a, b; Poder 2010; Lindberg 2015). This is a controversy about the different modes of financing and maintaining lighthouses that have actually existed (private, public, and other hybrid modalities) and their relative efficiency pursued in a truly Coasean spirit.

Coase and the Old Lighthouse System: A Possibility of Efficient Private Financing?

As related by Coase (1974), the history of English lighthouses is one of private entrepreneurs embarking on building and maintaining lighthouses in order to remedy public failures. Trinity House, a private charity in charge of lighthouses, was founded in 1514. At the beginning of the seventeenth century it was building very few lighthouses, and private individuals, supported by petitions from sailors and ship-owners, obtained patents from the Sovereign allowing them to build lighthouses at specific places and collect corresponding payments (and excluding others from doing so). Similarly to the lighthouses of Trinity House, these "light dues" were collected by Customs officers in harbors on a ship's arrival, and their amount depended on the route and tonnage of that ship. From 1679 onwards, Trinity House itself, having obtained a patent, could grant its right to a private individual, generally on a rental basis. During the eighteenth century, it was mainly private individuals who built lighthouses. But as early as 1820, Trinity House began to buy "private" lighthouses at the request of the House of Commons; this process was completed in 1842. The House of Commons justified the centralization in the first half of the nineteenth century by the complexity of the system and the high level of dues that went into private pockets, damaging British vessels' competitiveness.

In the old English system, Coase sees the possibility for a private entrepreneur to finance and maintain a lighthouse with financial gain as his sole motive:

The early history shows that, contrary to the belief of many economists, a lighthouse service can be provided by private enterprise. [...] The lighthouses were built, operated, financed and owned by private individuals, who could sell the lighthouse or dispose of it by bequest. The role of the government was limited to the establishment and enforcement of property rights in the lighthouse. The charges were collected at the ports by agents for the lighthouses. (Coase 1974, p. 375)

Reevaluation of the Nature and Efficiency of the Old English Lighthouse System

Coase's empirical material was subsequently reexamined, raising questions about his conclusions. His interpretation of the history of English lighthouses is marked by two forms of bias. First, Coase underestimates the role of public power in the old system of English lighthouses: this system was not strictly private but mixed. Van Zandt (1993, pp. 65–67) stressed that the patents granted by the Crown, or the leases granted by Trinity House, had three characteristics that distinguished them from a private system in which the Government's role is limited to the establishment and enforcement of property rights. The authorization to build and maintain a lighthouse (1) guaranteed a monopoly, (2) set the price of the service, and (3) involved the Crown in the collection of payments. These three aspects are interlinked, since price-setting was becoming necessary to protect users facing a monopoly, while the fixed price (as well as the contract duration and the sharing of the collected dues) determined the monopoly's rent and its distribution.

Second, Coase overestimates the efficiency of the direct relationship between producer and consumer. The service offered by this mixed system was considered expensive and complex, thus motivating the nineteenth century centralization (Taylor 2001). Moreover, the lighting provided by the lighthouses was often poor, and some were not even lit – since the fixed price meant that the only way to increase profits was to lower costs. In addition, the buildings themselves were sometimes erected by individuals who did not avail themselves of all the technical guarantees required. These problems were intensified by a certain form of corruption: patents were granted according to the Sovereign's goodwill, to those he or she favored or to those who offered him or her the greatest amount of money. It was even *because* the Sovereign obtained money when he or she authorized a private individual to build a lighthouse (money he or she would not have obtained if the authorization was granted to Trinity House) that he or she prevented Trinity House

from building lighthouses (until 1679). The high cost of light dues is thus explained by its being a monopoly rent of which the Sovereign collected a share (Bertrand 2006; Carnis 2013).

The nationalization that occurred at the beginning of the nineteenth century was followed by a lowering of dues and an improvement in technology, mainly thanks to the fact that research on optics was now being undertaken by engineers (the Stevenson dynasty) at a national level. They worked in close cooperation with the French government and its engineers (Augustin, then Leonor Fresnel), and with French and English industrials (Soleil, which later became Sautter; and Chance Brothers) (see Elton 2009). In fact the centralization aimed at implementing the most recent innovations in lighting, which had been developed in France but were too expensive to be imported by private English owners (Levitt 2016).

The French Lighthouse System: The Advantages of Centralization

The French lighthouse system, which had been centralized at the end of the *Ancien Régime*, confirms the potential advantages of centralization when compared to the decentralized English system (therefore, in the period 1790–1830) (Bertrand 2005).

The French Revolution gave control of lighthouses and beacons to a service within the Ministry of the Navy, and the Empire then transferred this role to a Lighthouses Service, created in 1805 and falling under the *Direction des Ponts et Chaussées* (within the Ministry of the Interior). The control of this service was eventually granted to the Lighthouse commission, a “place of scientific and political deliberation” created in 1811 and composed of engineers, scientists, and naval officers (Guigueno 2001, p. 99; compared to Trinity House, which was made up of retired ship-owners and honorary members). This commission explicitly thought of all the French coastal lighthouses as a system: its 1825 report planned a network of lighthouses along the French coasts. The French system was the source

from which the great technological innovations (like the Fresnel lens) were diffused. This technical superiority was linked to the administrative structure of the system, centralized and composed of engineers. Centralization opened the way to the testing of new techniques on several lighthouses before an authoritative application on all the coasts, which implied scale economies.

The two systems may therefore be opposed: on the one hand, the French centralized system, supervised by engineers and scientists, financed by general taxation, with building planned along the coasts according to a consistent national scheme; on the other, the pragmatic English system, more decentralized, supervised by honorary members, subject to favoritism, financed by light dues, and with the building of lighthouses determined by dramatic wrecks (Guigueno 2001, pp. 108–109).

While the French system of lighthouses is the archetype of the centralized model, it had not always been centralized. During the *Ancien Régime*, it was closer to the English dues system but with another mix of private and public that proved less efficient. We may take the example of the Cordouan lighthouse, in the Gironde estuary leading to Bordeaux, whose history is the best known (Bertrand and Guigueno 2016). This first French lighthouse was in fact the first *English* lighthouse: a tower was built there by the Black Prince, Edward of Woodstock, between 1360 and 1370. A document from 1409 testifies that dues were raised by a hermit to finance a reconstruction that had already become necessary. England would generalize the dues system for its lighthouses, as we have seen, and France would also retain it, but in a less efficient manner. During the fifteenth and sixteenth centuries dues were raised on ships entering the Gironde to finance a new lighthouse at Cordouan. Finally, after numerous complaints that dues were being raised although the tower was not being built, under the orders of the kings Henri III and then Henri IV, a new tower was designed by the architect Louis de Foix and completed in 1611: it was a Renaissance splendor, financed by dues and specific local taxes (at the municipal and regional levels). The situation rapidly deteriorated: dues were still being raised, but

the tower was already in ruins at the beginning of Louis XIV's reign. Dues were then raised by a receiver in the name of a "governor of the tower of Cordouan" but apparently did not go to the lighthouse. An entrepreneur was commissioned to maintain the tower and the light; he obtained this charge by auction to the lowest bidder. To win it, he proposed too low an amount to cover the expenses linked to the maintenance of the tower and the furniture for lighting. This is why during the seventeenth and eighteenth centuries, the tower was often not lit. The lighthouses system began to be centralized at the end of the *Ancien Régime*: Tourtille-Sangrain then Teulère were sent by the King to improve the structure and the lighting of Cordouan. And dues were de facto abandoned in the last decade of the eighteenth century.

Conclusion

At the turn of the nineteenth century, lighthouses, which were more or less privately provided with the help of the state, were nationalized not only in England and France but also in Estonia (Poder 2010) and Sweden (Lindberg 2015). A change of lighting technology may explain this change. From local lights, made excludable because the harbor was a complementary good to the lighthouse service (Varian 1993), improvements in the technology of lights made coastal lighthouses possible: the lighthouse became nonexcludable, which accounts for its nature as a "public good." Finally, the existence of "private" (local) lighthouses in the old English system, as brought to light by Coase, does not refute the assertions of economists like Mill and Samuelson that (coastal) lighthouses could more efficiently be produced by the state (Levitt 2016).

Cross-References

- Coase, Ronald
- Market Failure: Analysis
- Market Failure: History
- Public Goods

References

- Arrow KJ (1969) The organization of economic activity: issues pertinent to the choice of market versus non-market allocations'. In: Analysis and evaluation of public expenditures: the PPP system, vol 1. GPO, Washington, DC, pp 47–64
- Barnett W, Block WE (2007) Coase and Van Zandt on lighthouses. *Public Finance Rev* 35(6):710–733
- Barnett W, Block WE (2009) Coase and Bertrand on lighthouses. *Public Choice* 140(1–2):1–13
- Bertrand E (2005) Two complex lighthouse production systems: the mixed English and the centralized French systems. In: Finch J, Orillard M (eds) Complexity and the economy: implications for economic policy. Edward Elgar, Cheltenham/Northampton, pp 191–206
- Bertrand E (2006) The Coasean analysis of lighthouse financing: myths and realities. *Camb J Econ* 30(3):389–402
- Bertrand E (2009) Empirical investigations and their normative interpretations: a reply to Barnett and Block. *Public Choice* 140(1–2):15–20
- Bertrand E, Guigueno V (2016) Financer les phares, en France et en Angleterre: l'exemple de Cordouan. Mimeo
- Block WE (2011) Rejoinder to Bertrand on lighthouses. *Rom Econ Bus Rev* 6(3):49–67
- Carnis L (2013) The provision of lighthouses services: a political economy perspective. *Public Choice* 157(1–2):51–56
- Carnis L (2015) The political economy of lighthouses: some further considerations. *J Écon Étud Hum* 20(2):143–165
- Coase RH (1974) The lighthouse in economics. *J Law Econ* 17(2):357–376
- Elton J (2009) A light to lighten our darkness: lighthouse optics and the later development of Fresnel's revolutionary refracting lens 1780–1900. *Int J Hist Eng Technol* 79(2):183–244
- Guigueno V (2001) Des phares-étoiles aux feux-éclairs : les paradigmes de la signalisation maritime française au XIXème siècle. *Réseaux* 109:96–112
- Lai LWC, Davies SNG, Lorne FT (2008a) The political economy of Coase's lighthouse in history (part I): a review of the theories and models of the provision of a public good. *Town Plan Rev* 79(4):395–425
- Lai LWC, Davies SNG, Lorne FT (2008b) The political economy of Coase's lighthouse in history (part II): lighthouse development along the coast of China. *Town Plan Rev* 79(5):555–579
- Levitt T (2016) The implications of technological change for the role of the lighthouse in economics. Mimeo
- Lindberg E (2013) From private to public provision of public goods: English lighthouses between the seventeenth and nineteenth centuries. *J Policy Hist* 25(4):538–556
- Lindberg E (2015) The Swedish lighthouse system 1650–1890: private vs public provision of public goods. *Eur Rev Econ Hist* 19(4):454–468
- Mill JS (1848) Principles of political economy. Longmans, Green, London. Reprinted in Robson JM (ed) The collected works of John Stuart Mill, vol 3. University of Toronto Press, Toronto, 1965
- Pigou AC (1932) The economics of welfare, 4th edn. Macmillan, London
- Poder K (2010) The lighthouse in Estonia: the provision mechanism of “public goods”. In: Matti Raudjärvi (Ed.). *Discussions on Estonian economic policy*. Berliner Wissenschafts-Verlag, pp 323–347
- Samuelson PA (1964) Economics: an introductory analysis, 6th edn. McGraw-Hill, New York
- Samuelson PA, Nordhaus WD (2010) Economics. McGraw-Hill, Boston
- Sidgwick H (1901) The principles of political economy, 3rd edn. Macmillan, London
- Taylor J (2001) Private property, public interest, and the role of the State in nineteenth-century Britain: the case of the lighthouses. *Hist J* 44(3):749–771
- Van Zandt DE (1993) The lessons of the lighthouse: “government” or “private” provision of goods. *J Leg Stud* 22(1):47–72
- Varian HR (1993) Markets for public goods? *Crit Rev* 7(4):539–557
- Zerbe RO Jr, Mc Curdy HE (1999) The failure of market failure. *J Policy Anal Manage* 18(4):558–578

Limits of Contracts

Ann-Sophie Vandenberghe
Rotterdam Institute of Law and Economics
(RILE), Erasmus School of Law, Erasmus
University Rotterdam, Rotterdam, The
Netherlands

Synonyms

[Behavioral law and economics of contract law](#)

Definition

Limits of contracts refer to a number of exceptions in contract law to the rule that courts should fully enforce voluntary agreements between capable parties.

Introduction

Economic analysis and the rational actor model have dominated contract scholarship for at least a

generation. More recently, a group of behaviorists has challenged the ability of the rational choice model to account for contracting behavior. Numerous tests done by psychologists and experimental economics have shown that people often do not exhibit the kinds of reasoning ascribed to agents in rational choice models (Tversky and Kahneman 1974). Behavioral economics incorporates evidence of decision-making flaws that people exhibit to model consumer markets in which sophisticated firms interact with boundedly rational consumers. Behavioral law and economics uses existing scholarship in both cognitive psychology and behavioral economics to explain legal phenomena and to argue for legal reforms. For most of the legal scholars who apply behavioral approaches to contracts, evidence of cognitive biases provides at least a *prima facie* case for more paternalistic forms of legal intervention rather than strict reliance on freedom of contract. The premises of neoclassical law and economics push in the direction of freedom of contract: if parties are rational, they will enter contracts only when it is in their self-interest, and they will agree only to terms that make them better off; otherwise, they would not have voluntarily agreed to them (Posner 2003). Behavioral economics rejects the assumption that people are rational maximizers of their satisfactions in favor of assumptions of “bounded rationality,” “bounded willpower,” and “bounded self-interest” (Jolls et al. 1998). Bounded rationality refers to the fact that people have cognitive quirks that prevent them from processing information rationally, such as availability bias, overconfidence and overoptimism bias, the endowment effect, and status quo bias. The cognitive bias literature generally favors expanding the range of government regulation to address a wide variety of business practices that exploit the biases of consumers, or at least some consumers. Sunstein and Thaler (2003) advocate “libertarian paternalism,” an approach that preserves freedom of choice, but encourages both private and public institutions to steer people in directions that will promote their own welfare. The new paternalist claim is that deliberate structuring of decision contexts – such as assigning appropriate default options, providing cooling-off periods for

commitments, targeted disclosure, and so forth – can in principle enhance individuals’ welfare. This contribution represents the behavioral account as well as the neoclassical economic account of a number of contractual practices, clauses, and rules. It includes the penalty clause, pro-consumer default rules, the cooling-off period, automatic renewal provisions, choice manipulation, add-on and multidimensional pricing, and contracts with deferred costs.

Penalty Clauses

One of the earliest applications of behavioral insights to contract law was made by Melvin Eisenberg (1995) who attributed the courts’ reluctance to enforce penalty clauses to the limits of cognition. A penalty clause is a contractual provision that liquidates (fixes) damages in excess of the actual loss from contract breach. According to Eisenberg (1995), a special scrutiny of liquidated damages and penalty clauses is justified because such provisions are systematically more likely to be the product of limits of cognition. Parties at the bargaining stage are generally overoptimistic about their ability to perform and will sacrifice the detailed bargaining necessary to achieve an effective damage provision. The policy implication of this observation, according to Eisenberg, is that it is proper for courts to scrutinize these provisions more closely. Hillman (2000), however, argues that behavioral decision theory cannot resolve the puzzle of liquidated damages. Although some phenomena like overoptimism support the scrutiny of this provision, other cognitive heuristics and biases support strict enforcement of the provision. First, assuming that parties consider the default rule – here the award of expectation damages – as part of the status quo, contracting around the default and agreeing to liquidated damages suggest that the term must be very important for parties and that they bargained over the provision with care. Second, assuming that cognitively limited parties do not like ambiguity, they may prefer the safety of a liquidated damage provision over the uncertainty of expectation damages. Third, judges who exhibit

hindsight bias will overestimate the parties' abilities to calculate at the time of contracting the actual damages that would result from breach. Because judges will believe that the parties' remedial situation at the time of contracting was not ambiguous, judges will undervalue the importance the parties attach to the agreed-upon damage provision. For these behavioral reasons, courts should presume the enforceability of such provisions rather than making every effort to strike them out. Rachlinsky (2000) does not agree with these arguments. The status quo bias does not justify deference because the increased effort to bargain around the damage rule does not necessarily eliminate the effects of overoptimism. And although aversion to ambiguity justifies deference to liquidated damages, courts actually use this insight under the penalty doctrine by giving more deference to liquidated damages when damages are hard to calculate (and thus ambiguous). Eric Posner (2003) states that the behavioral account of the penalty doctrine cannot explain why the biases justify judicial scrutiny of liquidated damages but not other terms. If parties overlook low-probability events, then any contract term that makes obligations conditional on events that occur with a low probability could be defective on a behavioral account, but because they are not liquidated damages, actual courts do not subject them to scrutiny. The behavioral approach does not seem to explain existing contract law. It may also not provide a solid basis for normative recommendations for reforming contract law. Monumental floodgate problems would be created if courts were to police potentially all contract provisions to account for parties' irrationality in processing information. This would undermine contract law's goal of certainty and predictability (Hillman 2000).

The economic explanation for legal intervention rests on the argument that people tend to sign contracts without reading them (Goldberg 1974; Katz 1990). If contract drafters know that some parties will not read the document before signing, they can abuse that fact and incorporate clauses that benefit the drafter at the expense of the signers, with no corresponding welfare gains. Strict enforcement of contracts would imply a

duty to read and understand contracts before signing. It is however costly to read, evaluate, and understand contract terms. There are costs of consulting an expert and costs of time spent in reading and trying to understand the small print. Since rational parties balance the costs and benefits from reading and understanding contracts, it can be rational not to read when the costs are higher than the expected gain that a person stands to reap from reading and understanding contracts. This is likely to be the case for reading and understanding the contract terms' implications for rare events. Research in psychology and behavioral economics which suggests that most people are poor at estimating probabilities correctly can enrich the analysis at this point. To the extent that they systematically underestimate the probability of a rare event, they are unlikely to invest in understanding what the contract stipulates for rare events. The signing-without-reading problem can be sensibly addressed by a legal prohibition of contract terms which rational and informed parties would never accept (De Geest 2002). Rational and informed parties would never accept a contract term of which the costs to the debtor are higher than the benefits to the creditor. The fact that such a term is nevertheless adopted in a contract is implicit evidence that one of the parties did not read or understand the contract (Rea 1984). De Geest and Wuyts (2000) have shown that penalty clauses are in most cases irrational, providing in this way an economic ground not to enforce them.

Pro-consumer Default Rules

Pro-consumer default rules are a consumer protection technique commonly employed in American contract law, but also in European contract law. Sticky default options are a growing trend in consumer protection law (Bar-Gill and Ben-Shahar 2013). Parties can opt out, but the procedure for these opt-outs is more rigorous and costly. For example, the Common European Sales Law (CESL) rules on product conformity set high warranty standards which can be circumvented, but any agreement derogating from the requirements to the detriment of the

consumer “is valid only, if, at the time of the conclusion of the contract, the consumer knew of the specific conditions of the goods or the digital content . . .” (art. 99 (3) CESL). The technique is also used in labor law to protect employees. The behavioral justification for changing default rules – and sometimes making the default costly to change – is that people are subject to framing effects. This means their decisions tend to be sensitive to seemingly irrelevant aspects of how the choice situation is framed. Probably the best known type of framing effect is the endowment effect (Kahneman et al. 1990), which refers to people’s tendency to demand more compensation to give something up (their willingness to accept) than they would have paid to acquire the same thing (their willingness to pay). For rational agents, the default should not make a difference for choices (at least if transaction costs are low). But for less-rational agents, the default rule can matter. Employees, for example, might demand more compensation to eliminate a for cause termination clause than they would sacrifice to insert it. Sunstein and Thaler (2003) therefore suggest making “for cause” rather than “at will” the default employment termination rule. Yet behavioral theory does not provide policymakers with an answer to the question which default rule corresponds to the agent’s true preferences.

According to Verkerke (2015), legislators may choose pro-consumer default rules for another reason: such defaults may be thought to serve the purpose of better informing parties about the legal rules which apply to the transaction. People often lack basic information about the legal rules governing particular transactions in which they are routinely involved. Abundant empirical evidence reveals widespread ignorance about many aspects of civil law. When people do not know important legal characteristics of the things they buy, their willingness to pay may not accurately reflect their true valuation of those products and services. Information-forcing default rules suggest a possible solution to this problem of legal ignorance. Lawmakers could determine whether one party has a comparative advantage in obtaining and communicating information about the law governing the transaction. Then they

could select a default rule that disadvantages the better informed party knowing that the overwhelming majority of well-informed parties will opt out by drafting contract terms that meet certain standards for clarity. If so, a legal-information-forcing rule would force the comparatively better informed party either to reveal the relevant legal information or to accept a default rule that favors the less informed party. In order to escape the unfavorable default, the informed party must disclose information to her less well-informed contractual partner. Such a rule contrasts sharply with a conventional “majoritarian default” selected to mimic the terms that most parties would prefer for this type of transaction. However, if the purpose of these default rules is to convey legal information to all, or even many, unsophisticated parties, that objective is likely to be frustrated, according to Verkerke (2015). One salient fact about legal-information-forcing rules is that the targeted information is most often communicated in a standardized form contract such as a bill of sale, an employee handbook, a standard form insurance contract, a release of liability, or a rental agreement. Almost no one reads contracts carefully enough to digest the legal information that these default rules are designed to force, which may lead to the signing-without-reading problem. A legal requirement according to which opt-out is allowed only after the consumer expresses conscious, informed consent may be seen as procedural evidence (evidence about the circumstances of contracting) that there is no signing-without-reading problem (De Geest 2002). Ayres and Schwartz (2014) observe that consumers need not read to learn about prevailing contract terms. Instead, they form expectation about terms based on what they learn from store visits, read in the media, experience in their own transactions, and hear from friends. These beliefs may accurately correspond to the content of a particular contract. Or consumers may suffer from “term optimism” – a mistaken belief that a contract’s terms are more favorable than the provisions actually included in its text. Ayres and Schwartz (2014) propose an elaborate contracting regime designed to combat term optimism. Their system would induce mass-market sellers to survey consumers periodically to

determine if those consumers correctly understood the terms of the seller's agreement – a process they call “term substantiation.” Sellers then would be required to use a standardized warning box to disclose any terms that consumers mistakenly thought were more favorable. Terms that meet or exceed the median consumer's expectations would not have to be included in the warning box, to prevent overuse of the box. Behavioral support exists for targeted warnings instead of ever-expanding disclosures. People tend to read more when there is less to read (“less is more”) and tend to take more relevant criteria for decision-making into account when there is less choice, a phenomenon called the “choice paradox.”

Cooling-Off Periods

When cooling-off periods apply, contracts are not enforceable in the standard way, namely, from the moment the contract is concluded, but instead can freely be canceled within a certain time period after the conclusion of the contract. During the cooling-off period, consumers have the right to withdraw from the contract at no cost and for no reason, contrary to the general principle that contracts are binding (i.e., *pacta sunt servanda*). European law gives consumers the right to withdraw for a range of contracts for goods and services, such as doorstep selling transactions, time-share contracts, distance selling, and credit contracts (Eidenmüller 2011).

The behavioral support for cooling-off periods is the evidence that people make different decisions depending on whether they are in a “hot” or “cool” state (Loewenstein 2000). According to Sunstein and Thaler (2003), the essential rationale for cooling-off periods is that under the heat of the moment, consumers might make ill-considered or improvident decisions that they would not make if they were in a cool state. Both bounded rationality and bounded self-control are the underlying concerns. The legal provision of a cooling-off period allows the decision-maker to reconsider his decision once he is in a cooler mental state, but structures the decision context in other behaviorally relevant ways. The evaluation now takes place in

a state where the decision-maker already possesses the goods. The endowment effect refers to the tendency of people to value things more when they own them (Thaler 1980). Status quo bias and the pervasive drive for consistency may render the cooling-off period inoperative (Hoeppner 2014). To create an optimal cooling-off policy justified on the basis of hot-state bias, policymakers need to know the extent of any given bias, the rate at which the particular hot state dissipates, and the self-debiasing measures adopted by individuals and have to deal with the problem of interdependent biases and heterogeneity in the population. Rizzo and Whitman (2009) argue that policymakers do not have access to all the knowledge needed to implement welfare-improving paternalistic policies.

There is also a plausible economic basis for the right to withdraw (Rekaiti and Van den Bergh 2000; Ben-Shahar and Posner 2011). The economic rationale for the right to withdraw is based on the fact that ex ante information costs (before the contract is concluded) may be higher than the information costs ex post when the consumer is able to inspect the goods he has in his possession. Faced with high ex ante information costs and immediate binding force of the contract, a consumer may rationally decide not to buy a good. This effect is unfortunate from the standpoint of both consumers and firms. Some welfare-enhancing transactions do not take place due to high transaction costs. The right to withdraw then exists to minimize information costs and will increase the number of transactions compared to the status quo ante. Buyers are more likely to buy a good if they have the right to return it if they do not like it. According to the neoclassical economic approach, a case can be made for making the right to withdraw the default rule for modern market transactions characterized by high ex ante information costs. There is however no need for legally mandated withdrawal rights since market parties have incentives to offer them voluntarily. As a matter of fact, many stores have return policies even when they are not legally required to have them. A legal mandate would remove the signaling-of-quality function that return policies may have.

Automatic Renewal Clauses

An automatic renewal clause is a contractual provision according to which a contract is automatically renewed for a new term unless it is canceled. From a behavioral perspective, an auto-renewal clause is considered to be a provision for automatic (default) enrollment in subscriptions and contracts whereby consumers are allowed to opt out, as opposed to the practice of manual renewal which is a practice of not enrolling consumers until they opt in. The behavioral case for default enrollment is based on inertia or status quo bias: the psychological tendency of people to maintain current arrangements, whatever they might be (Samuelson and Zeckhauser 1988). For example, adherence to pension savings plans has been found to increase dramatically when employees are enrolled automatically (Madrian and Shea 2001). That automatic enrollment leads to more enrollment is something which firms know when they adopt automatic renewal provisions in their contracts. Should the use of automatic renewal provisions in contracts be restricted on the ground that they may exploit behavioral biases? Not necessarily, because it may also be regarded as a genuine paternalistic measure to help overcome behavioral biases. Suppose that consumers often fail due to status quo bias to enroll under the manual renewal opt-in system, but they would choose to enroll if they simply took the time to think carefully; then, by offering an automatic renewal service, firms are acting paternalistically. However, strict enforcement of an automatic renewal provision implies the danger of renewal without notice. Consumers are not always aware of the fact that and when the contract will renew automatically. Therefore, they may fail to opt out when renewal is not in their best interest which results in allocative inefficiency. Firms may speculate on this fact when they include an automatic renewal clause in the contract. Legislators have addressed this danger of renewal without notice by requiring conspicuous disclosure of automatic renewal clauses and by requiring notice of renewal to be provided by firms a number of days in advance.

Choice Manipulation

An important finding of behavioral research is that values and preferences are not fixed, but depend on endowment, context, or the way in which a choice is presented – a phenomenon designated as “framing.” The choices of a behavioral consumer can be manipulated. For example, due to anchoring bias, a particular choice might become much more attractive when placed next to a very unattractive one. It is well documented that many firms now hire consultants who advise them on how to better exploit behavioral biases of consumers. Such consultants may recommend, for instance, to use decoy pricing techniques. A decoy is a highly priced product that is not expected to be sold, but that only serves to make less-highly priced products look reasonable. Car dealers learn similar techniques to manipulate potential buyers. One such technique consists of asking a series of (irrelevant) questions to which customers are likely to answer “yes.” Then customers are asked whether they would like to buy options for their new car. Apparently, more customers answer affirmatively when this technique is used. Consumers also are more likely to make a choice on buying extra car options when they were asked before to make a choice on an irrelevant matter. Apparently, once the brain is set into a choice modus, it is likely to continue to follow a pattern of choice while neglecting the possibility of not choosing (e.g., to not buy any extra options). These methods to mislead consumers would lose their effect if firms would have to reveal the fact that they try to exploit behavioral biases. According to De Geest (2014), modern law on information duties is best explained by the least-cost-information-gatherer principle. According to this principle, sellers should reveal information on all aspects of which they are the least-cost information gatherer, provided that the information is material. Information is material when the cost to produce and disclose the information is lower than its value to the buyer. Information is valuable when it can influence the recipient’s decision to contract. Does this principle imply a duty to reveal the fact that one uses manipulative techniques? Clearly, the salesperson is the least-cost

information gatherer about her own use of manipulative techniques. But is this information material? Not in an ideal search process, where buyers can instantly see all products on the market, their features, their objectively tested quality, and their final prices. But in practice, search is often sequential or advice based. When search is sequential or advice based, much more information is material than price and quality, including information on the fact that manipulative techniques are used. De Geest (2014) concludes that sellers who use manipulative techniques should honestly reveal the fact that they use them. In the absence of such a duty, markets would be governed by the caveat emptor principle (“buyer beware”), which leads to allocative inefficiency, distributive distortions, wasteful precaution, and transaction avoidance.

Add-On and Multidimensional Pricing

In a wide range of plausible situations, firms systematically give the impression that the price is lower than it will in fact turn out to be. A common reason for this is that part of the service which many consumers will want is not included in the “headline” price used in marketing (Armstrong 2008). In many businesses, it is customary to advertise a base price for a product and to try to sell additional “add-ons” at high prices at the point of sale, such as extended warranties, shipping fees, and hotel phone and minibar charges in the hotel room. Add-on prices are not advertised, and they would be costly or difficult to learn before one arrives at the point of sale. For example, online, some auctioneers have tried to bury shipping costs deep in an item’s description so that casual bidders will overlook the fee. A fraction of consumers (the “sophisticates”) either observe or foresee the high price of the add-on product and can substitute away from it at little cost. They could decide in advance whether or not to buy the extended warranty, rather than have to make this decision under pressure from the sales assistant. Naive consumers do not consider that they may want the add-on product until they have purchased the main item. They fail to take

add-ons into account when comparing product prices across firms. Gabaix and Laibson (2006) show that competition will not induce firms to educate the public about the add-on market, even when unshrouding is free. The reason for this is a phenomenon which they call the “curse of debiasing.” Educating a consumer about competitors’ add-on schemes effectively teaches that consumer how to profitably exploit those schemes, thereby making it impossible for the educating firm to profitably attract the newly educated consumers. From a policy perspective, they argue that regulators might compel disclosure or could warn consumers to pay attention to shrouded costs. The duty to disclose information in a conspicuous way (instead of shrouding information) can be supported on economic grounds. The practice of shrouding information is a violation of the principle that the least-cost information gatherer should disclose information in the most efficient way. The practice of hiding information and untimely disclosure artificially increases search costs for all consumers, not only the unsophisticated ones.

Oren-Bar-Gill (2004, 2006, 2008) points out that consumers face difficulties in assessing information about the true product price and quality in case of multidimensional pricing and bundling. Examples are rebates, credit card pricing, bundling of printers with ink, and intertemporal bundling in subscription markets like health clubs. In these cases, consumers need additional information about use patterns in order to assess the price and quality (value) of the contracts correctly. For example, in order to assess how expensive the combined purchase of printer and ink really is, the consumer needs to know the price of the printer and the price of an ink cartridge, but also how often he will need to replace the ink cartridge, which depends in turn on the frequency of his use. In order to assess the price and value of health club subscription, consumers need to know the subscription price and the quality of the club services, but also how often they will attend the health club. Consumers who misperceive their future use will make use-pattern mistakes. For example, consumers may overestimate their likelihood of redeeming their rebate, underestimate

the likelihood of paying their credit card bills late, underestimate the amount of printing, or overestimate the number of times that they will visit the health club. To the extent that these mistakes are systematic, they may be due to cognitive biases, like the overconfidence and overoptimism bias. Bar-Gill argues that the importance of use-pattern mistakes requires more and better use-pattern disclosure. In particular, sellers should be required to provide average or individualized use-pattern information. For example, the disclosure apparatus of credit cards should include the amount that an average consumer pays in late fees and how much the individual has paid in late fees over the last year.

Contracts with Deferred Costs

Bar-Gill (2012) shows how a firm that faces a market of myopic consumers can exploit this flaw. Firms can do this by stacking benefits of a contract at the beginning of the contract term and leaving the costs for later. Deferred costs exploit the intense preference that some consumer may have for immediate gratification. This myopia leads them to underestimate whether and how often they will fall prey to the deferred fees that many consumer contracts impose. When consumers evaluate agreements with up-front benefits and deferred costs, they may be tempted by the benefits – such as low introductory interest rates or a subsidized cell phone – while also believing that they will not have to pay the later costs. Bar-Gill adheres to disclosure as the most desirable means to remedy consumer myopia. Badawi (2014) is skeptical that disclosure will be an effective guard against myopia. More extreme measures may be necessary to prevent the harm that the weakness of the will might cause, just like Ulysses has asked his men to strap him to the ship's mast to prevent him from succumbing to the sirens. Bar-Gill points out that firms have little incentive to provide myopia-limiting products. Should legislators take paternalistic measures upon request? Any legal restriction on contract terms needs to be done cautiously, according to Badawi (2014). There are potentially

welfare-enhancing reasons why firms might want to offer low introductory rates that increase substantially after a set period of time. A prohibition of the ability of firms to tantalize consumers with low introductory offers may harm the welfare of the rational, nonmyopic consumers. A behaviorist might see these marketing techniques as temptations to overconsume, while a pure rationalist might see them as attempts to induce an informed switch. How do we know what to do when theory can justify both intervention and the free-market status quo? According to Badawi, the lack of a clear answer to this question suggests that there is a pressing need for data that can discern whether rationality or myopia prevails. In the absence of such data, the decision to regulate or not will largely be a function of the strength of prior beliefs. In societies with a strong belief in the benefits of free markets, there will be more freedom of contract, but also more behavioral manipulation of consumers camouflaged as offering free choice (De Geest 2013).

References

- Armstrong M (2008) Interactions between competition and consumer policy. *Compet Policy Int* 4:96–147
- Ayres I, Schwartz A (2014) The no-reading problem in consumer contract law. *Stanford Law Rev* 66:545–610
- Badawi A (2014) Rationality's reach. *Mich Law Rev* 112:993–1014
- Bar-Gill O (2004) Seduction by plastic. *Northwest Univ Law Rev* 98:1373–1434
- Bar-Gill O (2006) Bundling and consumer misperception. *Univ Chicago Law Rev* 73:33–61
- Bar-Gill O (2008) The behavioural economics of consumer contracts. *Minnesota Law Rev* 92:749–802
- Bar-Gill O (2012) Seduction by contract: law, economics, and psychology in consumer markets. Oxford University Press, Oxford
- Bar-Gill O, Ben-Shahar O (2013) Regulatory techniques in consumer protection: a critique of European consumer contract law. *Common Mark Law Rev* 50:109–126
- Ben-Shahar O, Posner E (2011) The right to withdraw in contract law. *J Leg Stud* 40:115–148
- De Geest G (2002) The signing-without-reading problem: an analysis of the European directive on unfair contract terms. In: Schäfer H, Lwowski H (eds) *Konsequenzen Wirtschaftsrechtlicher Normen*. Gabler Verlag, Wiesbaden, pp 213–235
- De Geest G (2013) Irredeemable acts, rent seeking, and the limits of the legal system: a response to professor Raskolnikov. *Georgetown Law J Online* 103:23–28

- De Geest G (2014) The death of Caveat Emptor. University of Chicago Law School. Law and Economics Workshop. http://www.law.uchicago.edu/files/files/degeest_paper_0.pdf
- De Geest G, Wuyts F (2000) Penalty clauses and liquidated damages. In: Bouckaert B, De Geest G (eds) Encyclopedia of law and economics, volume III. The regulation of contracts. Edward Elgar, Cheltenham, <http://encyclo.findlaw.com/4610book.pdf>
- Eidenmüller H (2011) Why withdrawal rights? ERCL 1:1–24
- Eisenberg M (1995) The limits of cognitions and the limits of contracts. Stanford Law Rev 47:211–259
- Gabaix X, Laibson D (2006) Shrouded attributes, consumer myopia, and information suppression in competitive markets. Q J Econ 121:505–540
- Goldberg V (1974) Institutional change and the quasi-invisible hand. J Law Econ 17:461
- Hillman R (2000) The limits of behavioral decision theory in legal analysis: the case of liquidated damages. Cornell Law Rev 85:717–738
- Hoepfner S (2014) The unintended consequence of doorstep consumer protection: surprise, reciprocity, and consistency. Eur J Law Econ 38: 247–276
- Jolls C, Sunstein C, Thaler R (1998) A behavioural approach to law and economics. Stanford Law Rev 50:1471–1550
- Kahneman D, Knetsch J, Thaler R (1990) Experimental tests of the endowment effect and the coase theorem. J Polit Econ 98:1325–1348
- Katz A (1990) The strategic structure of offer and acceptance: game theory and the law of contract formation. Mich Law Rev 89:215
- Loewenstein G (2000) Emotions in economic theory and economic behavior. Am Econ Rev 90:426–432
- Madrian B, Shea D (2001) The power of suggestion: inertia in 401(k) participation and savings behaviour. Q J Econ 116:1149
- Posner E (2003) Economic analysis of contract law after three decades: success or failure? Yale Law Rev 112:829–880
- Rachlinsky J (2000) The “new” law and psychology: a reply to critics, skeptics, and cautious supporters. Cornell Law Rev 85:739–766
- Rea S (1984) Efficiency implications of penalties and liquidated damages. J Leg Stud 13:147–167
- Rekaiti P, Van den Bergh R (2000) Cooling-off periods in the consumer laws of the EC member states. A comparative law and economics approach. J Consum Policy 23:371–407
- Rizzo M, Whitman D (2009) The knowledge problem of new paternalism. BYU Law Rev 4:905–968
- Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. J Risk Uncertain 1:7–59
- Sunstein C, Thaler R (2003) Libertarian paternalism is not an oxymoron. Univ Chicago Law Rev 70: 1159–1202
- Thaler R (1980) Toward a positive theory of consumer choice. J Econ Behav Organ 1:39–60
- Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131
- Verkerke J (2015) Legal ignorance and information-forcing rules. William & Mary Law Rev 56:899–959

Liquidation Preference

► Absolute Priority

Litigation and Legal Expenses Insurance

Myriam Doriat-Duban

Department of Economics, BETA UMR 7522,
Université de Lorraine, Nancy, France

Abstract

Legal expenses insurance (LEI) covers all or a part of financial expenses occurring during a lawsuit, in exchange for the payment of a premium by the policyholder to the insurer. It makes access to justice less expensive by shifting the litigation costs from the litigant to the insurer. LEI covers many types of conflicts (labor litigations, consumer disputes, personal injuries, or medical malpractices), but some are generally excluded (divorces). LEI could have some significant effects on conflict resolution, not only for the litigants themselves but also for society. Thus, in the law and economics literature, three topics are generally studied: the consequences of LEI on conflict resolution and social welfare, the influence of the insurer in the litigation, and finally the access to justice by a comparative approach between LEI and legal aid and contingent fees.

Synonyms

Legal cost insurance; Legal insurance; Legal protection insurance

Definition

Legal expenses insurance (LEI) covers all or a part of financial expenses occurring during a lawsuit, in exchange for the payment of a premium by the policyholder to the insurer. There exist two types of LEI: on the one hand, before-the-event (BTE) which covers litigation costs that could be incurred in a future case, and on the other hand, after-the-event (ATE) which covers future legal expenses in a case that has already occurred (the economic literature only deals with BTE LEI, which is the most usual one). Insurers can propose a specific LEI policy or add this insurance to another one, for example, household insurance. This type of insurance covers many types of conflicts such as labor litigations, consumer disputes, personal injuries or medical malpractices, but some are generally excluded, such as divorces. In this respect, it does not constitute a perfect substitute for alternative financing for lawsuits (legal aid or contingent fees).

Introduction

Access to justice is costly for litigants (lawyers' fees, procedural costs, costs of appraisals, etc.). Some systems exist to make access to justice less expensive, among them legal expenses insurance that shifts the litigation costs from the litigant to an insurer. This system prevails in Europe whereas contingent fees, which are forbidden in most European countries, are the main alternative in the United States. Interest in LEI has risen with the increasing weight of legal aid in the budget of European States. Policymakers in many European countries (United Kingdom, France, Belgium, etc.) have tried to develop demand for LEI, which remains underdeveloped for several more or less relevant reasons: alternatives for access to justice (especially if legal aid is "relatively generous"), adverse selection if the insurer cannot identify the real risk of the policyholder, moral hazard if the insurance spurs the insured party to adopt a riskier behavior or to go to trial more frequently, positives externalities in terms of deterrence, and free rider problems for potential victims who have less

incentive to be insured (Faure and De Mot 2012). Policies in favor of LEI attract the interest of law and economics scholars because LEI could have some significant effects on conflict resolution, not only for the litigants themselves but also for society. The law and economics literature can be divided into three parts: (1) authors who have developed economic models to study only the consequences of LEI on conflict resolution and social welfare, (2) authors who have studied the influence of the insurer in the litigation, and (3) authors who have adopted a comparative approach between LEI and alternative financing for access to justice, mainly legal aid and contingent fees.

Impact of LEI on Conflict Resolution and Social Welfare

Most authors study the consequences of legal expenses insurance on conflict resolution. Among them, Kirstein (2000) demonstrates that for the plaintiff LEI has a positive effect on his/her strategic position by making his/her threat to sue more credible, especially in cases with a negative expected value. More precisely, by decreasing the policyholder's legal costs, some cases with negative expected value might become positive and this makes the plaintiff's threat to sue credible. For cases with positive expected value, the negotiation gap widens (the plaintiff requests a higher amount while the defendant offers a lower amount because LEI covers their legal costs) and the amount of the agreement increases (the plaintiff's threat to sue becomes more credible).

Heyes et al. (2004) extend Kirstein's paper by studying the insurance decision of a risk-adverse plaintiff and the effects of this insurance on settlement amounts, settlement probabilities, volume of accidents and trials. They show that the plaintiff chooses to be insured if and only if he/she plans to refuse the defendant's offer. Conversely, he/she chooses to be uninsured if he/she plans to accept the agreement. The effects of LEI on the plaintiff's bargaining position explain this result. Indeed, LEI reduces his/her legal costs and thus increases his/her incentive to go to trial and to reject the defendant's offer. This higher bargaining power

can have a negative impact on settlement so that the probability of settlement decreases and the number of trials rises. Thus, LEI appears to encourage access to justice instead of alternative dispute resolution. But LEI could have a positive effect if these higher incentives to go to court increase the defendant's level of care. Finally, the overall effect is unknown and depends on the plaintiff's risk aversion. Consequently, LEI can increase or decrease social welfare.

Van Velthoven and van Wijck (2001) confirm that the effect of LEI on social welfare is uncertain because the deterrence is improved but at the same time, the number of trials increases. For liability cases, they explain that LEI makes the strategic position of the plaintiff stronger. The effect of LEI comes into play not only *ex post*, during the conflict, but also *ex ante*, where potential injurers will take greater care if they anticipate that plaintiffs are insured against litigation costs. Thus, LEI could have a preventive role and hence a positive impact on social welfare, but only if the expected prevention gains offset the losses of welfare due to the higher number of trials.

The Influence of the Insurer in the Litigation

Visscher and Schepens (2010) analyze the multi-party relationship in a conflict where the litigant has taken out a LEI policy. They explain that the specificity of this insurance lies in the joint presence of several actors: plaintiff, defendant, insurer, and lawyers. Interactions between these actors create conflicts of interest due to agency relationships where, for example, the plaintiff delegates the defense of his/her interests to a lawyer or an insurer. There could also be a specific relationship between the lawyer and the insurer, in particular if the insurer employs the lawyer (especially in the stage of bargaining in the shadow of the law). Visscher and Schepens (2010), like Kirstein (2000) and Heyes et al. (2004), confirm that LEI increases the credibility of the threat of trials by reducing the plaintiff's legal costs. The worst situation in terms of arrangement is where both litigants are insured:

freed from their trial costs, both have more incentive to go before the judge. Nevertheless, the interests of the insurer have to be taken into account. The insurer's aim is to minimize the legal costs it covers. This is what is at stake in the agency relationship between the litigant and his/her insurer: the insurer may incite its uninformed client (on his/her rights or the actual compensation to which he/she is entitled) to accept an agreement, not always in his/her best interests but just to save legal costs. Visscher and Schepens add that the impact of this type of "disadvantageous" agreement for the insured on the reputation of the insurer may not constitute a safeguard against these practices.

Finally, Qiao (2013) takes into account the role of the insurer in the negotiations, studies the three-way relationship between the client, the lawyer, and the insurer, and determines the optimal LEI. He shows in particular that demand-side cost-sharing is necessary to design an optimal LEI. More precisely, the plaintiff is required to share the cost of the service received to mitigate his/her incentive to overuse the insurance.

Comparative Analysis Between LEI and Two Alternatives: Contingent Fees (CF) and Third-Party Litigation Funding (TBF)

Some other authors adopt a comparative and more normative approach. Their aim is to compare several systems whose aim is to allow access to justice and to determine which of them maximize social welfare.

In this perspective, to choose between two alternative systems, Baik and Kim (2007) compare the American practice of contingent fees with the European practice of legal expenses insurance. They introduce the role of the lawyer through the type of fees. They study the American system by considering that the plaintiff's lawyer works on a contingent-fee basis and the defendant's lawyer on an hourly fee basis. For the European practice with legal expenses insurance, the defendant may have to purchase a LEI policy and both lawyers work on an hourly fee basis. They show that the

plaintiff's expected payments are higher under a legal insurance system than under a contingent fee system but litigants' legal expenditures are also higher because the lawyers are likely to be more aggressive.

Friehe (2010) compares contingent fees and LEI, two systems which help liquidity-constrained litigants to finance their case. He generalizes the previous results by introducing the defendant's degree of fault. He finds that the performance of both regimes (in terms of expected plaintiff payoffs, plaintiff and defendant expenditures, total contest effort, incentives for delegation, and justice) is highly dependent on the defendant's fault.

Ancelet et al. (2012) extend the analysis of Heyes et al. (2004) by studying the consequences of a substitution between LEI and legal aid on conflict resolution. Their results differ significantly from those of Heyes et al. (2004) who study only LEI and show that LEI is taken out only by plaintiffs who anticipate going before the judge, regardless of their risk-aversion. The introduction of the choice between legal aid and LEI modifies the analysis. Indeed, the plaintiff can take out LEI even if he/she anticipates an arrangement or, conversely, prefers legal aid if he/she expects to go to trial. But they show also that the agreements accepted by plaintiffs if they are covered by LEI are superior to those required if they are beneficiaries of legal aid, regardless of their risk-aversion. Accordingly, under the authors' assumptions, legal aid seems more favorable to agreements than LEI, because it makes the plaintiffs less demanding (they accept a lower offer under legal aid than under LEI). The development of LEI could then appear desirable for litigants (who would obtain better arrangements) but not for the legislator if its aim is to reduce access to courts.

Moreover, in the same way that insurers want to encourage an out-of-court settlement in order to reduce their own costs, the lawyer who is poorly paid in a legal aid system may encourage agreement instead of judgment to avoid additive costs. Thus both legal aid and LEI may encourage their clients to settle the dispute even if it may be worthwhile for these clients to go to court.

Everything is then played out within the agency relationship between the lawyer and the beneficiary in the case of legal aid, and the insured and the insurer in the case of LEI.

In an empirical study in the Netherlands, van Velthoven and Klein Haarhuis (2011) examine the relationship between the decision to use the justice system, whether or not with LEI, and the income of individuals (divided into three classes: those who can benefit only from legal aid, those who are eligible for legal aid but can also take out LEI, those who can only take out LEI). They show that, all other things being equal, the likelihood of using a judge is 11% higher for litigants insured through LEI than for the others. There are two reasons for this. First, there appears to be a selection effect in that individuals who have already had a case brought before the courts are more inclined to take out LEI. Secondly, there is a moral hazard problem since insured individuals are more likely to go before the judge, because LEI covers their costs of litigation.

Finally, Faure and De Mot (2012) compare LEI and Third-Party Financing Litigation (TPLF) with the idea that LEI could be a barrier to the development of TPFL. The answer is negative, mainly because LEI is underused in the United States and many European countries. Their analysis shows that TPFL is not necessarily worse than LEI in terms of volume of litigation, quality of litigation, or timing of settlements. Moreover, in Europe there is hostility towards TPFL, because the trial is considered as a risk rather than a way to enforce the law.

Conclusion

The impact of LEI on litigation is complex because several effects interact. Some authors examine its impact on litigation in a positive light. Others adopt a more normative approach and compare the effects on litigation of LEI relative to the effects of other avenues of conflict financing (legal aid, contingent fees, TPLF). Empirical studies are still too scarce to highlight the theoretical results. Moreover, LEI ought to be more developed in Europe where it is still underused.

Cross-References

- [Legal Aid](#)
- [Third-Party Litigation Funding](#)

References

- Ancelot L, Doriat-Duban M, Lovat B (2012) Aide juridictionnelle et assurance de protection juridique : coexistence ou substitution dans l'accès au droit. *Rev Fr Econ* XXVII:115–148
- Baik KH, Kim I (2007) Contingent fees versus legal expenses insurance. *Int Rev Law Econ* 27:351–361
- Faure MG, De Mot JPB (2012) Comparing third party financing of litigation and legal expenses insurance. *J Law Econ Pol* 8(3):743–778
- Friehe T (2010) Contingent fees and legal expenses insurance: comparison for varying defendant fault. *Int Rev Law Econ* 30:283–290
- Heyes A, Rickman N, Tzavara D (2004) Legal expenses insurance, risk aversion and litigation. *Int Rev Law Econ* 24:107–119
- Kirstein R (2000) Risk neutrality and strategic insurance. *Geneva Pap Risk Insur* 25(2):251–261
- Qiao Y (2013) Legal effort and optimal expenses insurance. *Econ Model* 32:179–189
- Van Velthoven B, Klein Haarhuis C (2011) Legal aid and legal expenses insurance, complements or substitutes? The case of the Netherlands. *J Empir Leg Stud* 8(3):587–612
- Van Velthoven B, Van Wijck P (2001) Legal cost insurance and social welfare. *Econ Lett* 72:387–396
- Visscher L, Schepens T (2010) A law and economics approach to cost shifting, fee arrangements and legal expense insurance, chap. 2. In: Analysis M, Visscher L (eds) *New trends in financing civil litigation in Europe, a legal, empirical, and economic*. Edward Elgar, Cheltenham, pp 7–32

Litigation and Marital Property Rights

Antony W. Dnes
Northcentral University, Arizona, AZ, USA

Abstract

The article examines the function of marriage and associated incentive properties in relation to ancillary relief (often referred to as “settling up”) following divorce. Many countries have

experienced growing divorce rates, a decline in marriage, increased unmarried intimate cohabitation, and the delaying of marriage and childbirth to a later age. There has also been a recent increase in the pressure to extend marriage formalities to same-sex couples. All of these changes raise questions concerning the incentive structures attached to marriage and divorce. Major incentive issues arise whenever there is public-policy debate about changing the law of marriage and divorce with associated implications for litigation. It is vital to understand the economics underlying the debate since there is a danger that well-meant reform might lead to adverse unintended consequences.

Introduction

Over the past 50 years, many countries have experienced growing divorce rates, a decline in marriage, increased unmarried intimate cohabitation, and the delaying of marriage and childbirth to a later age. That commentary would once have been restricted to western societies, but the trends have begun to appear in developing economies: for example, the Chinese divorce rate increased by almost 13% in 2013 (*South China Morning Post*, March 11, 2015, “Heartbreaking news: China’s divorce rate jumps 13pc as more choose to untie the knot.”). In the USA, the divorce rate per thousand population has fallen in recent years, although as a result of marriage rates falling; the divorce rate per thousand marriages continues to rise (See data collected by the US Center for Disease Control at http://www.cdc.gov/nchs/nvss/marriage_divorce_tables.htm). These trends have caused concern, not least because of the unsettling nature of the apparent social instability. Additionally, and one might say curiously since heterosexuals are avoiding marriage, there has been a recent increase in the pressure to extend marriage formalities to same-sex couples. All of these changes raise questions concerning the incentive structures attached to marriage. Major incentive issues arise whenever there is public-policy debate about changing the law of marriage

and divorce with associated implications for litigation. It is vital to understand the economics underlying the debate because there is a danger that well-meant reform might lead to adverse unintended consequences.

Underlying the incentive structures in intimate partnerships, we find a potential for exploitation by at least one of the partners: a hold-up problem. This can arise if long-term promises induce detrimental reliance, such as one partner's giving up work to become a homemaker, in expectation of long-term benefits such as shared ownership of property and nonpecuniary benefits such as marital consortium. Under weak enforcement of promises, one partner could opportunistically make promises and then subsequently renege on them. The economic analysis of marriage and divorce has tended to focus on problems associated with the post-WWII easing of divorce law. Failure to enforce promises or compensate for breach by the promisor can create an adverse incentive structure identified as a "greener-grass" effect (Dnes 1998, 2011; Dnes and Rowthorn 2002). The effect is often associated with relatively wealthy men abandoning older wives. However, there is a comparable incentive for a dependent spouse to divorce if settlement payments, based on dependency, allow the serial collection of marital benefits without regard to the costs imposed on the other party: encouraging a "Black-Widow" effect (Dnes 1998). Two related issues concern contemporary unmarried cohabitation: (i) whether a lack of legal support for long-term relationship-specific "investments" results in low-quality commitment where more commitment could be beneficial and (ii) whether assimilating same-sex unions into marriage or marriage-like structures creates high-quality commitment signals, or whether these signals are irrelevant in such relationships (Dnes 2007).

Marriage and Opportunism

Becker's (1973, 1991) seminal work on the family is based on specialization and the division of labor within the household (Grossbard 2010) and does not really give a clear reason for the emergence of

a state-sanctioned standardized marriage contract. One could just as well cohabit with a grandparent and achieve economies of specialism. The theory is based on a neoclassical approach to rational decision-making (Dnes 2009). It is possible that the hostility toward Becker's work exhibited by some sociological writers might be reduced by moving to a context-dependent, ecological view of rationality (Smith 2008), in which individuals use heuristics to economize on decision-making capacity. Becker's work has led to more recent bargaining theories of the family, which, however, could still relate to cohabitation as much as marriage (Friedberg and Stern 2014).

Cohen (1987, 2002, 2011) pioneered a contractual view of marriage that is firmly anchored in institutional economics and does give a distinctive reason for marriage rather than mere cohabitation: the suppression of opportunistic behavior connected to asymmetries in the life cycles of men and women. Divorce can then be seen as analogous to breach of contract, although it should be noted that marriage predated the development of modern contract law and that fault-based divorce is not exactly the same as breach of contract. For example, no-fault divorce is consistent with legally sanctioning one spouse's desertion, a breach. Furthermore, fault encompasses criminal and tortuous behavior such as assault that is subject to separate legal penalties (Ellman and Lohr 1997) likely to add to deterrence of the behavior. A purely contractual starting point for ancillary relief in divorce litigation would be very modern, although contractual elements can be found in the case law (Cohen 1987, p. 270). A contractual approach is also capable of considerable sophistication, particularly where inherently economic issues like asset division are at stake, or when examining signaling aspects of marriage (Dnes and Rowthorn 2002; Probert and Miles 2009). Finally in this regard, one could base the enforcement of marital obligations on quasi contract, i.e., recognizing the absence of a formal contract and using *equity* rather than *law* as the approach, and still reach much the same results.

Cohen (1987) notes that spouses exchange unusual promises of support where the value of

the support is affected by the attitude accompanying it. In a traditional marriage, domestic services provided by the wife frequently occur early on in the marriage, permitting the husband to concentrate on employment. The wife typically makes nonrecoverable, i.e., sunk, investments in child-rearing and homebuilding early on in the marriage. The traditional male's support will typically grow in value over the longer term. If the opportunities of the parties change, one of them may develop an incentive to breach the "contract." Divorce usually imposes costs on both parties, but costs are asymmetrically distributed in dissolving traditional marriages: the husband typically finds it easier to remarry or repartner (Cohen 1987; Kreider and Rose 2006), whereas the wife has incurred many sunk costs at an earlier stage.

There are both pecuniary and nonpecuniary benefits to marriage. A spouse's willingness to commit is satisfying evidence of love. Marriage also gives a means of protecting long-term investments in marital property. Spouses may be regarded as unique capital inputs in the production of a new output, namely, "the family." In particular, children are important marital outputs. A further instrumental gain is insurance: parties mutually support each other through the ups and downs of life forsaking the freedom to seek new partners, which is rational if the net gains from marriage are sufficiently high (Posner 1992). However, the gains from marriage are surpluses that arise after sunk investments have been made, which can tempt the less-tied-down spouse into opportunistic behavior, specifically into attempts to appropriate the product of the sunk investments. One spouse's incentives to appropriate the gains from marriage may well operate similarly to the incentives arising within joint business ventures (Klein et al. 1978).

Marriage-specific benefits such as the prospect of losing association with children are "hostages" (Raub 2009) that may suppress opportunistic exit from the marriage in some cases. However, judicial approaches to obligations like long-term support can easily lead to too much or too little divorce. This observation brings in the idea of an optimal level of divorce, which might be encapsulated in a rule like "let them divorce when the

breaching party (the one who wants to leave, or who has committed a "marital offence") can compensate the victim of breach," but not otherwise. The underlying welfare perspective in the last sentence is the Kaldor-Hicks *utilitarian* approach of allowing a change when it will increase joint (i.e., social) welfare.

Marriage promises revolve around direct and instrumental benefits, bargaining influences (Lundberg and Pollak 1996), joint goods, marriage-specific investments, and incentives for opportunistic restructuring. Divorce law needs to suppress opportunism, or it may end up creating incentives for litigation and destabilizing marriage. Perceived instability in the marriage contract may then deter some from marrying: fewer marriages will occur than otherwise, and there may be less investment in activities such as child raising (Stevenson 2007).

The preservation of a very clear signal of commitment is particularly important in marriage, which may already be destabilized by inappropriate laws (Rowthorn 2002). Well-meaning legal reforms have tended to decouple obligations from fault in divorce cases, and in doing so have overlooked the importance of enforcing promises that support "investments" in the family. Rationally, it will be the economically weaker spouse who would alter behavior in the face of lower promissory enforcement, for example, increasing the time spent searching for a reliable spouse. Such a search-time effect has been observed empirically (Mechoulam 2006; Matouschek and Rasul 2008): introducing no-fault divorce settlements (as opposed to no-fault divorce rules per se) into US jurisdictions is statistically linked to increases in the age of first marriage. Some US states have responded to the perceived weakening of marital commitment with a "covenant marriage" movement, providing an option for alternative marriage rules making the exit rules tougher.

Divorce Litigation and Liability Rules

It is well known that breach of contract can be optimal, providing that the breaching party pays

compensation for the victim's lost expectation (Posner 2014; Parisi et al 2011). How far does this analogy carry over in a contractual view of marriage? Expectation damage, as the standard common-law remedy for breach of commercial contracts, has the characteristic of requiring the breaching party to pay compensation that places the victim of breach in the position they would have attained had the contract been completed (see Dnes 2005, p. 97). This requirement meets the Kaldor-Hicks welfare criterion since the gainer must gain more than the loser loses to be willing to make the change; the loser is as well off as before. The usual approach to a commercial contract breach also requires the victim to mitigate losses and does not compensate for avoidable losses. For the victim of marital breach, modeling ancillary relief on expectation damages would imply awarding the value of lost pecuniary support and lost nonpecuniary elements of promise such as consortium while making an allowance for the value of alternative occupations that may open up as a result of dissolution. The nonpecuniary elements do not present the inherent difficulties that might be anticipated since courts already make such calculations, for example, in tort cases involving the loss of a spouse.

Basing ancillary relief on court-imposed compensatory damages for lost expectancy applies just one kind of liability rule, in the terms used by Calabresi and Melamed (1972), and it is instructive to consider alternatives. In particular, the same welfare effects, in terms of deterring inefficient breach, could follow from awarding restitution damages aimed at disgorging the breaching party's gains back to the victim of breach. For simplicity, consider a case where the gains exceed the victim's lost expectation. Logically, the breaching party would still make the move if left with just enough of the gains to be slightly better off from divorce. The victim of breach could therefore be overcompensated, but, ignoring for the moment possible strategic considerations affecting the victim, the incentive to divorce for the breaching party would be just the same as under the expectancy liability rule. Therefore, a liability rule requiring compensation of at least the victim's lost expectancy and no more

than the breaching party's gains can support efficient breach.

Awarding anything less than expectancy damages risks making divorce too "cheap" for the breaching party, and, indeed, depriving the victim of the value of promises. Undercompensation is likely to result from approaches based on compensating detrimental "reliance," which would follow from limiting ancillary relief to compensating for lost opportunities such as career options forgone through entering the marriage. The expectancy must have exceeded the reliance, or the spouse would have remained single. Similarly, restitution of "contributions" or combinations of contributions and reliance would tend to undercompensate.

Another approach to settling up after divorce could be based on specific performance, which is generally not favored in relation to commercial contracts mainly because of difficulties of supervision. However, Parkman (1992, 2002) has suggested basing divorce law on a specific-performance requirement for spouses to remain married unless they mutually agree to divorce, whenever they could bargain at low cost. The spouse wishing to leave would have to offer the other at least expectation damages to obtain consent. This approach follows a property rule in the terminology of Calabresi and Melamed (1972) and (Ayres 2005). Parkman claims that an advantage of the specific performance requirement is the avoidance of coercive divorce, although coercion would in fact also be avoided by a (superefficient) court able to carry out precise liability rule calculations. The bargaining might extract the equivalent of restitution damages ("how much can you part with to leave?") rather than expectancy, but would still lead to an efficient result (gainer still gains something after compensating loser). The economics literature on bargaining in the presence of incentives to appropriate gains from trade opportunistically does support the need for the vulnerable party to have complete bargaining power (Rosenkranz and Schmitz 2007).

Bargaining may be impeded by indivisible elements such as the value to parents of contact with children, described by Zelder (1993, 2009) as a marital public good. During marriage, both

parents can enjoy the company of their children, and one parent's enjoyment does not in general reduce the value of the other parent's contact. From a welfare perspective, we should add up both parents' enjoyment (heuristically assuming directly measurable welfare) in valuing the marital public good. Once they divorce, the value of parental contact with the children becomes separable and contact times change. It is possible that the intact marriage has a positive value because of the public good, but a parent without court-awarded child custody may not be able to transfer assets to the other to prevent the divorce. There may be few other assets in the marriage, and the parent with custody retains private enjoyment of the marital public good. This reflection leads to the Zelder paradox: the more the family invested in children, products of love rather than market-valued labor, the more likely is inefficient dissolution. Indivisibilities are also problematic in the more general analysis of suppressing opportunistic bargaining (Rosenkranz and Schmitz 2007).

Consideration of the autonomous welfare of children is not commonly undertaken in the literature on divorce, notwithstanding the UN Convention on the Rights of the Child and national legislation such as the UK's Children Acts. An exception is Bowles and Garoupa (2003), which treats the welfare effects of the divorce on the children of divorcing parents as an externality. As might be expected, a marriage that is no longer efficient from the point of view of the parents' joint welfare may still generate net social benefits if there is enough harm to the children from divorce. That comparison does however beg a nontechnical question that should actually be addressed more frequently in economics: whose preferences are to count? If, as in the common-law world traditionally, minor children are not recognized as possessing mature mental capacity, then the externality cannot be said to arise.

In family law, children's preferences have not typically been *generally* recognized (Cretney 2005; Ellman and Ellman 2008; Ellman et al 2014). It is easy to be misled by the UN Convention, adopted by all countries except the USA and Somalia, which appears to require courts to give primary consideration to children's

welfare, but does this in a restricted sense falling far short of conferring autonomous enforceable rights. Similarly, the UK's Children Act 1989, which is good exemplar of national legislation, requires paramount consideration to be given to the welfare of children in matters relating to their *upbringing*. As Reece (1996) and Potter (2008) both point out, the primacy or paramountcy is required over a very narrow area and is to be sought by a court acting in the interests of the child, a somewhat indirect "voice." The UK Children and Families Act 2014 removed a residual power retained in the Matrimonial Causes Act 1973 under which the court could previously refuse a divorce decree if the court felt the harm to the family's children to be intolerably high. Family law tends not to regard children as fully autonomous "minds" and continues to acknowledge their general lack of capacity. Legal recognition of greater children's rights in divorce proceedings would require a showing, not only that there is an effect on children but also that we should recognize their capacity more.

Types of Intimate Cohabitation

Traditional Marriage

A traditionalist view of the marriage contract sees an exchange of lifetime support for the wife, in which she shares the standard of living of the *marriage*, for typically feminine domestic services such as housekeeping and child-rearing. Could the contract view easily include less traditional frameworks, such as modern marriages where the parties are more economically equal? The marriage contract appears to be a good example of a contract where performance of the parties is interdependent, and where consideration for performance remains executory (Parisi et al. 2011).

Consider first a lengthy traditional marriage that ends in divorce. The parties met early in life, and after some years the wife gave up work to have children and care for them. When she returned to work it was at a lower wage than previously. Then, after 20 years of marriage, the husband petitions for divorce on the grounds of

separation. Their assets have always been held jointly. On a contract approach to breach, the husband would be expected to share property and income to maintain the standard of living his ex-wife would have enjoyed for the *remainder* of the marriage (subject to any mitigation of loss that may be possible).

Expectation damages are identical to the minimum needed to buy the right to divorce under a system where divorce is only available by consent. The court would assess what that standard of living was and determine who had breached the contract by focusing on proximate causes. The fact that the divorced wife gave up work for a while, or now earns less than might have been the case without child-care responsibilities, is immaterial in finding expectation damages. Her own income would contribute to that expectation, as would her own share of the house and other assets and the combined decision-making of the spouses. The divorcing husband would be expected to contribute from his income and his share of the assets to provide that support for his ex-wife, regardless of the impact on his own lifestyle. Any requirement to maintain the standard of living of the children of the marriage could be dealt with separately by the court.

Under a contract approach, the courts would recreate the expected living standard of the wife by adjusting the property rights and income distribution of the parties at divorce. Fault would matter because the court would need to establish who the breaching party was, but this would not rule out calling some dissolutions “no-fault divorce,” which is simply a compensable breach. Fault would be irrelevant in a system of mutual consent, where both parties negotiate a settlement stating that neither was at fault; where bargaining would safeguard expectations. The classical-contracting approach preserves incentives for the formation of traditional families, if that were considered important.

Classical contracting is also consistent with the simultaneous existence of separate legal obligations for the maintenance of children. However, it would only be consistent with a literal interpretation of the clean-break principle favored in much recent family law if sufficient property rights can

be transferred to avoid the need for subsequent periodic payments. A classical-contract view would not be consistent with ultratraditional views emphasizing the sanctity of marriage, requiring specific performance, and creating *inalienable* rights.

Alternative approaches to expectation damages in ancillary relief tend to be based on partial views of social welfare. Feminist writers reject divorce rules that reinforce the dependency of women on men. Unfortunately, there is a real danger of lowering the welfare of women as a consequence of the push for independence. Of course, if one wanted to deter traditional marriage, undercompensating divorced traditional wives would do just that (Cohen 1987; Cohen and Wright 2011). Not all feminist writers in family law share the same views. Kay (1987) argues for measures to increase equality between males and females in their social roles. Others (Gilligan 1982) argue that men and women are different (women’s art, women’s ways of seeing, women’s writing, and so on). Moves in divorce law to compensate spouses for career sacrifices have been sympathetically received by these groups, but actually give less-than-expectation damages, such as reliance or restitution (Trebilcock 1993). Such moves have been influential in the case law and legislation of several countries (Ellman 2007).

Modern Marriage and Unmarried Cohabitation

Separating awards of ancillary relief in divorce litigation from the issue of breach of contract is highly likely to encourage opportunistic behavior. However, for some commentators (e.g., Ellman 2007), expectation damages in marriage contracts imply lifetime support, which conflicts with a modern emphasis on equality between spouses. Another concern is encouraging costly protracted arguments over the identification of breach, although this concern may be relevant only when the court system is run largely from public funds. Would a more contractual approach to marriage resolve or amplify these issues?

The social norms surrounding intimate relationships have clearly changed in favor of serial

marriages and, indeed, unmarried cohabitation (Almond 2006). A 2015 report shows that 40% of childbirths are now outside of marriage in the USA, split 59% to 41% in favor of unmarried cohabiting parents compared with noncohabiting mothers (US Sees Rise in Unmarried Births, *Wall Street Journal*, March 10, 2015 access at <http://www.wsj.com/articles/cohabiting-parents-at-record-high-1426010894>). Marriage rates are falling, cohabitation rates are rising, and divorce rates are rising in many countries, suggesting that the view of marriage as a lifelong commitment may not correspond with the wishes of the population at large. Legal liberalization of divorce allows people to change their minds as circumstances change and to modify marriage promises. Is a more flexible view of marriage useful?

The evolution of marital law, at least in common-law jurisdictions, shows marriage to have been heavily influenced by surrounding social norms, in a manner consistent with modern legal scholarship on “relational” contracts, which are shaped by a surrounding minisociety of norms (Macneil 1978; Williamson 1985 and Macaulay 1991; Ellman et al 2014). Brinig (2000) has developed the idea of wider governance of the family into an approach emphasizing marriage as a covenant with wider society, in which family ties do not end with events like divorce but become modified as an ongoing franchise. That is a view that allows some elements of promise to persist separately when other elements are rescinded.

One approach to flexible marital covenants might be to encourage the use of clearer marriage contracts with the possibility of enforceable modifications that could be substitutes for divorce (Jones 2006; Brown and Fister 2012). The literature on contract modifications is extremely pessimistic over the prospect of welfare gain from enforcing *mutually agreed* modifications (Jolls 1997; Dnes 1995; Miceli 2002). It is difficult to distinguish good-faith revisions from those resulting from opportunistic behavior. Consider the difficulty in marriage contracts in distinguishing between a good-faith modification reflecting changing conditions underlying the marriage and the case, with no underlying change, where a party threatens to make a spouse’s life

difficult unless new terms are agreed. Contract modifications are more likely to be welfare enhancing if, in the context of unforeseen events, (i) there was a risk and it is not clear who is the lowest-cost bearer of that risk, (ii) the events were judged of too low a value to be worth considering in the contract, or (iii) neither party could possibly have borne the risk (Dnes 1995, p. 232). Generally, the view that supporting all requested modifications is desirable is unsound: there may be undesirable long-term instability as a result. Fewer people will make contracts when they cannot be protected from opportunism.

The court can determine whether some change was foreseeable and whether the attendant risk would have been clearly allocated: for example, one’s spouse’s aging is not a reason for scooting off without making compensation. On the other hand, mutually tiring of each other would have been hard to allocate to *one* party. Generally, the focus of family litigation can be expected to remain the division of benefits and obligations on divorce. Thus, there is some interest in matching ideas of relational contracting with practical rules for courts (Scott and Scott 1998). Macneil (1978) has suggested that complex long-term contracts are best regarded holistically in terms of the promissory relation as it has developed over time. An original contract document (e.g., marriage vows) would not necessarily be of more importance in the resolution of disputes than later events or altered norms.

A relational contract may be of limited help in designing practical solutions to divorce issues unless it is possible to fashion legal support for the relational contracting process. It is a fascinating challenge to put the idea of contractual flexibility together with the persistent caution of this article over the dangers of opportunistic behavior. Many problems of dividing marital assets arise because social norms have changed but the individual marriage partners failed to match the emerging marital norm. One major issue concerns marriage-dependent older wives who are likely to have specialized in domestic activities but will now be expected to be more self-reliant upon divorce. A possible approach to settling litigation would use expectation damages to guard against

opportunism but allow the interpretation of expectation to be governed by differing “vintages” of social norms. Consideration could also be given to making prenuptial, and postnuptial, agreements between spouses legally binding, which they are not in all jurisdictions. Some tailoring of marriage might be allowed, with couples choosing between *several* alternative forms of marital contract (e.g., traditional, partnership, or embodying restitution damages on divorce). A more flexible system of marriage and divorce could operate around a statutory obligation to meet the needs of children.

Cohabitation as an Imperfect Marriage Substitute

It is still early days in relation to explaining the growth in unmarried intimate cohabitation that has occurred in many societies (Almond 2006). One line of inquiry emphasizes the lower risk of loss of lifetime welfare following out-of-wedlock pregnancy that takes effect after the 1950s as birth control improved (Akerlof et al 1996). This change may have caused dominant female behavior to switch to taking more risks in unmarried relationships, and life-cycle asymmetries may have also become less marked relative to men. The obvious increase in female labor-market participation since WWII would also tend to lower the relative benefit from being insured within a traditional marriage.

Cohabitation is treated quite distinctly from marriage across all countries. The distinction is maintained in proposals from the American Law Institute to increase the protection afforded to unmarried partners in quasi-divorce litigation. The traditional common-law position is revealed in *Marvin v. Marvin* (122 Cal. Rptr. 815 [App. 1981], which is a case in a community property state), which holds that property settlements after cohabitation depend on discernible contracts and explicit or implied trusts. The position is similar in England and other countries that derive their legal systems from English common law. Statutory reforms in Canada, Australia, and New Zealand have edged the common-law world toward the continental European position of extending

elements of family-law jurisdiction to the property settlements of unmarried cohabitants.

Is it appropriate to intervene in cohabitation arrangements that have been freely entered into by apparently rational adults (Probert and Miles 2009)? Questions can be raised about how rational or informed the parties are, and there is some evidence that many cohabitants have an erroneous view that cohabitation over a period of time leads to similar rights in law as those enjoyed by married couples (Brinig 2000). It is not clear that it is welfare enhancing as a general rule to intervene ex post in arrangements that resulted from unilateral mistake (Rasmussen and Rasmusen and Ayres 1993). If intervention were to protect relationship-specific investments, then the state would effectively be prohibiting unmarried cohabitation: it would simply be creating a further form of marriage. If the general problem is really ignorance, rather than intervene in arrangements, the state might limit its efforts to providing better information flows – as in a pilot scheme operated in 2007 in the UK. If men and women really are more equal in society, such that the life-cycle asymmetries discussed above are now much weakened, then it would be appropriate to worry less about imposing the protections of marriage upon cohabitants, who, after all, could have chosen to marry. Jones (2006) has gone so far as recommending the privatization of marriage – that is, recognition of individually tailored marriage contracts – given modern developments.

There may be firmer ground under concerns that adverse impacts on children follow from less stable relationships. Cohabitation typically dissolves more frequently than marriage. Suppose that cohabitation has grown because men can be held to marriage less easily by women, given all the social changes since the 1950s. Then choices are freely made, but subject, as ever, to constraints that alter the results of choice. It is legitimate to ask whether the characteristics of the resulting equilibrium are acceptable. Scholars such as Popenhoe (1996) and more general commentators such as Bartholomew (2006, p. 249) argue that some of the characteristics, particularly the results of growing fatherlessness for children, are not acceptable. Consideration of the external effects

of less stable relationships would lead to policies encouraging marriage rather than cohabitation: not really a basis for extending marriage-like ancillary relief rules to cohabitants.

Same-Sex Marriage

The litigation-related discussion surrounding marriage and cohabitation can be extended to same-sex relationships (Fandrey 2013). One important difference is that, whereas the heterosexual has always had a choice between marriage and cohabitation, the same-sex couple has not generally had this choice. Recently, many jurisdictions have extended marriage and divorce rights to same-sex couples. From an economic perspective, the recent extension raises the question whether same-sex marriage has the same functionality as heterosexual marriage, which seems unlikely. The question comes down to whether the spouses are subject to similar life-cycle asymmetries as those identified by Cohen (1987) for traditional couples. There must be some probability that domestic investments are made asymmetrically by one partner in some same-sex unions, just as in heterosexual union, which, if true, would suggest extending something like marriage rights to same-sex couples (Dnes 2007). However, same-sex relationships appear to be characterized by considerably greater equality compared with traditional marriage (Kurdek 2004; Weisshaar 2014). In fact, equality is positively related to relationship commitment for same-sex couples (Kurdek).

It may therefore be a reasonable objection that there are very few same-sex couples with life-cycle asymmetries comparable to those of heterosexual couples in traditional marriages, which are anyway in decline. In jurisprudence, an argument that few people are affected by a condition is not in itself dispositive of a principle. We might still wish to protect even a small number of possible victims of opportunistic behavior, and support welfare-enhancing institutions even if they just affected a few. However, standing alone, that consideration would appear to open up further

extensions of marriage, for example, allowing polygamists to marry.

A strong argument for caution over extending marriage rights to same-sex couples is given by Allen (2006, 2010), who notes the possible externality involved. The welfare of large sections of the population may be lowered by the existence of marriage forms of which they do not approve, and their welfare does count. Nonetheless, same-sex marriage may be becoming increasingly acceptable, which would weaken the effect of the externality as a limit on extending marital protection. From a welfare perspective, one can coherently argue that a loss of welfare across the population may outweigh the benefits of marital protection likely to affect very few same-sex spouses. Also, Allen's argument gives a plausible limit on extending marriage rights further. Even though some participants in polygamous relationships may be very vulnerable to opportunism from an economically dominant "spouse," there are so few polygamists that a widespread revulsion would indeed dominate as a welfare effect.

Conclusions

Sophisticated contract thinking suggests a case for an expectation-damages approach to divorce litigation encompassing ancillary relief (support obligations and asset division). This conclusion is founded on controlling the incentive for opportunistic behavior set up by the asymmetric life cycles of males and females in traditional marriages. Current divorce laws may encourage opportunistic behavior in both males and females that is predatory in nature.

The contractual view of marriage and divorce explored in this article is really a perspective that proceeds by useful analogy. Different vintages and varieties of marriage would need to be recognized in practice since few marriages are now of traditional type. The approach is also consistent with a separate system of liability for child support. Contract thinking also illuminates recent trends toward unmarried cohabitation and new forms of marriage.

References

- Akerlof GA, Yellen JL, Katz ML (1996) An analysis of out of wedlock childbearing in the United States. *Q J Econ* 111:277–317
- Allen DW (2006) An economic assessment of same-sex marriage laws. *Harv J Law Public Policy* 29:949–980
- Allen DW (2010) Who should be allowed into the marriage franchise? *Drake Law Rev* 58:1043–1075
- Almond B (2006) *The fragmenting family*. Clarendon, Oxford
- Ayres I (2005) *Optional law*. University of Chicago Press, Chicago
- Bartholomew J (2006) *The welfare state we're in*. Methuen, London
- Becker GS (1973) A theory of marriage: part 1. *J Polit Econ* 81:813–846
- Becker GS (1991) *A treatise on the family*. Harvard University Press, Cambridge, MA
- Bowles R, Garoupa N (2003) Household dissolution, child care and divorce law. *Int Rev Law Econ* 22:495–510
- Brinig MF (2000) *From contract to covenant*. Harvard University Press, Cambridge, MA
- Brown NM, Fister KS (2012) The intriguing potential of postnuptial contract modifications. *Hastings Womens Law J* 23:187–212
- Calabresi G, Melamed AD (1972) Property rules, liability rules and inalienability: one view of the cathedral. *Harv Law Rev* 85:1089–1128
- Cohen LR (1987) Marriage, divorce, and quasi-rents; or, 'i gave him the best years of my life'. *J Legal Stud* 16:267–303
- Cohen LR (2002) Marriage: the long-term contract. In: Dnes AW, Rowthorn R (eds) *The law and economics of marriage and divorce*. Cambridge University Press, Cambridge
- Cohen LR, Wright JD (2011) Introduction. In: *Research handbook on the economics of family law*. Edward Elgar, Cheltenham
- Cretney S (2005) *Family law in the twentieth century: a history*. Oxford University Press, Oxford
- Dnes AW (1995) The law and economics of contract modifications: the case of *Williams v. Roffey*. *Int Rev Law Econ* 15:225–240
- Dnes AW (1998) The division of marital assets. *J Law Soc* 25:336–364
- Dnes AW (2005) *Economics of law: property contracts and obligations*. Cengage, Mason
- Dnes AW (2007) Marriage, cohabitation and same-sex marriage. *Indep Rev* 12:85–99
- Dnes AW (2009) Rational decision making and intimate cohabitation. In: Probert R, Miles J (eds) *Modern approaches to family law*. Hart Publishers, Oxford
- Dnes AW (2011) Partnering and incentive structures. In: Cohen LR, Wright JD (eds) *Research handbook in law and economics series*. Edward Elgar, Cheltenham, pp 122–131
- Dnes AW, Rowthorn R (eds) (2002) *The law and economics of marriage and divorce*. Cambridge University Press, Cambridge
- Ellman IM (2007) Financial settlement on divorce: two steps forward, two to go. *Law Q Rev* 122:2–9
- Ellman IM, Lohr S (1997) Marriage as contract, opportunistic violence and other bad arguments for fault divorce. *Univ Ill Law Rev* 1997:719–772
- Ellman IM, O'Toole Ellman T (2008) The theory of child support. *Harv J Legis* 45:107–163
- Ellman IM, Mackay S, Miles J, Bryson C (2014) Child support judgments: comparing public policy to the public's policy. *Int J Law Policy Family* 28(3): 274–301
- Fandrey SB (2013) The goals of marriage and divorce in missouri: the state's interest in regulating marriage, privatizing dependency and allowing same-sex divorce. *St Louis Univ Public Law Rev* 32:447–486
- Friedberg L, Stern S (2014) Marriage, divorce and asymmetric information. *Int Econ Rev* 55:1155–1199
- Gilligan C (1982) *In a different voice: psychological theory and women's development*. Harvard University Press, Boston
- Grossbard S (2010) How "Chicagoan" are Gary Becker's models of marriage?. *J Hist Econom Thought* 32:377–395
- Jolls C (1997) Contracts as bilateral commitments: a new perspective on contract modification. *J Legal Stud* 26:203–237
- Jones CPA (2006) A marriage proposal: privatize it. *Indep Rev* 11:115–119
- Kay H (1987) Equality and difference: a perspective on no-fault divorce and its aftermath. *Univ Cincinnati Law Rev* 56:1–90
- Klein B, Crawford RG, Alchian AA (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J Law Econ* 21:297–326
- Kreider, Rose M (2006) Remarriage in the United States, presented at the annual meeting of the American Sociological Association, Montreal, August 10–14, 2006, <https://www.census.gov/hhes/socdemo/marriage/data/sipp/us-remarriage-poster.pdf>
- Kurdek LA (2004) Are gay and lesbian cohabiting couples really different from heterosexual married couples? *J Marriage Fam* 66:880–900
- Lundberg S, Pollak R (1996) Bargaining and distribution in marriage. *J Econ Perspect* 10:139–158
- Macaulay S (1991) Long-term continuing relations: The American experience regulating dealerships and franchises. In: C. Joerges (ed) *Franchising and the Law: Theoretical and Comparative Approaches in Europe and the United States*, Nomos Verlagsgesellschaft, Baden-Baden, pp 179–237
- Macneil IR (1978) Contracts: Adjustment of long-term economic relations under classical, neoclassical and relational contract law. *Northwestern University Law Review* 72:854–906

- Matouschek N, Rasul I (2008) The economics of the marriage contract: theories and evidence. *J Law Econ* 51:59–110
- Mechoulan S (2006) Divorce laws and the structure of the american family. *J Legal Stud* 35:143–174
- Miceli TJ (2002) Over a barrel: a note on contract modification, reliance, and bankruptcy. *Int Rev Law Econ* 22:161–173
- Parisi F, Luppi B, Fon V (2011) Optimal remedies for bilateral contracts. *J Legal Stud* 40:245–271
- Parkman A (1992) No-fault divorce: what went wrong? Westview Press, San Francisco
- Parkman Alan (2002) Mutual consent divorce. In: Dnes AW, Rowthorn R (eds) *The law and economics of marriage and divorce*. Cambridge University Press, Cambridge, pp 57–69
- Popenhoe D (1996) *Life without father*. Harvard University Press, Cambridge, MA
- Posner RA (1992) *Sex and reason*. Harvard University Press, Cambridge, MA
- Posner RA (2014) *Economic analysis of law*. Wolters Kluwer, New York
- Potter M (2008) The voice of the child: children's "rights" in family proceedings. *Family Law* 2:15–36
- Probert R, Miles J (eds) (2009) *Modern approaches to family law*. Hart Publishers, Oxford
- Rasmusen E, Ayres I (1993) Mutual and unilateral mistake in contract law, *Journal of Legal Studies*, 22:309–343
- Raub, Werner (2009) Commitments by hostage posting. *Rationality, Markets and Morals*, 207–225. Accessed at <http://www.mmm-journal.de/htdocs/volume0.html>
- Reece H (1996) The paramouncy principle: consensus or construct? *Curr Legal Prob* 49:267–304
- Rosenkranz S, Schmitz PW (2007) Can coasean bargaining justify pigouvian taxation? *Economica* 74:573–585
- Rowthorn, Robert (2002) Marriage as signal. In: Dnes AW, Rowthorn R (eds) *The law and economics of marriage and divorce*. Cambridge University Press, Cambridge
- Scott ES, Scott R (1998) Marriage as relational contract. *Virginia Law Rev* 84:1225
- Smith V (2008) *Rationality in economics: constructivist and ecological forms*. Cambridge University Press, Cambridge
- Stevenson B (2007) The impact of divorce laws on investment in marriage-specific capital. *J Labor Econ* 25: 75–94
- Trebilcock M (1993) *The limits of freedom of contract*. Cambridge University Press, Cambridge
- Weisshaar K (2014) Earnings equality and relationship stability for same-sex and heterosexual couples. *Soc Forces* 93:93–123
- Williamson OE (1985) *The economic institutions of capitalism*. Free Press, New York
- Zelder M (1993) Inefficient dissolutions as a consequence of public goods: the case of no-fault divorce. *J Legal Stud* 22:503–520
- Zelder M (2009) The essential economics of love. *Teoria* 29:133–150

Litigation Decision

Samantha Bielen^{1,2}, Wim Marneffe¹ and Wim Vereeck¹

¹Faculty of Applied Economics, Hasselt University, Diepenbeek, Belgium

²Faculty of Law, University of Antwerp, Antwerp, Belgium

Abstract

This entry provides an overview of the main economic models of settlement and litigation decisions. Starting from the basic models, as developed by Landes (*J Law Econ* 14:61–108, 1971), Posner (*J Leg Stud* (0047–2530) 2:399, 1973), and Gould (*J Leg Stud*, 279–300, 1973), we describe the evolution in literature toward the application of bargaining theory. Scholars, recognizing the existence of private information and strategic behavior, increasingly modeled the process of settlement negotiations.

Economic Modeling of the Decision to Sue or Settle

Literature concerning the economic analysis of settlement and litigation has been constantly evolving since the early 1970s of the past century. Prompted by the development of the first economic decision-making models by Landes (1971), Posner (1973), and Gould (1973), law and economics scholars have persistently researched the causes of litigation. Intuitively, one might expect that all disputes will be settled rather than litigated due to the desire to avoid substantial legal costs. Although the vast majority of disputes is resolved through settlement, a small proportion still ends up in court. By means of economic modeling, scholars have tried to identify the determinants of settlement failure.

Initially, scholars focused on the outcome of the decision-making process of rational disputants. Landes (1971) focused on criminal cases, Posner (1973) applied his framework to administrative proceedings, and Gould (1973) suggested

other applications such as labor-management disputes. The existence of a “settlement range” is at the heart of this research, which applies cooperative game theory to settlement and litigation decisions. According to this approach, a necessary condition for disputes to be settled out of court is a positive settlement range, which means the defendant’s expected loss from trial exceeds the plaintiff’s expected benefit from trial. However, this is not a sufficient condition since settlement negotiations can still break down because parties cannot agree on dividing the surplus.

Subsequently, Shavell (1982) described the two-stage nature of litigation. Before parties decide whether they will settle or proceed to trial, the plaintiff has to decide to file suit. According to Shavell (1982), a rational plaintiff will only file suit when his expected value of a trial is positive. Although the earlier articles assumed that the latter condition was always satisfied, more advanced models include strategic behavior. Even in case of a negative expected value of the lawsuit, plaintiffs may still extract a settlement offer from their opponent, due to informational asymmetries which may cause the defendant to wrongly believe that the plaintiff has a credible threat. The nature and existence of these “frivolous lawsuits” has been examined, for example, in Bebchuk (1987, 1996) and Katz (1990).

Initially, the analysis was restricted to the likelihood of disputes ending in court, rather than the successful settlement amounts. The main finding was that the critical component of settling is agreement on the likelihood of success. Although divergent estimates of the latter explained the occurrence of litigation in these models, the possible sources of divergence were not discussed. This was the purpose of the next bulk of papers, which abandoned the cooperative nature of the game. Law and economics scholars, such as Salant and Rest (1982), Ordover and Rubinstein (1983), and P’ng (1983), started including private information. While these first incomplete information models solely examined the determinants of the likelihood of settlement, Bebchuk (1984) expanded the analysis to the settlement amount.

The first asymmetric information models assumed one-sided information asymmetry and a

take-it-or-leave-it offer by the uninformed party (i.e., a screening model). Subsequently, the framework was expanded by assuming that the informed party makes the first offer (see, e.g., Reinganum and Wilde (1986)). Successively, the assumption of a take-it-or-leave-it offer was abandoned by Spier (1992), who allowed for a sequence of offers. Other authors, such as Schweizer (1989) and Daughety and Reinganum (1994), considered a two-sided information asymmetry, where both the plaintiff and the defendant are privately informed about a certain element of the game.

Most litigation models take a positivistic approach and focus on the circumstances in which settlements break down, the amount for which parties settle, how procedural rules influence the likelihood of settlement, etc. Alternatively, the normative approach of other scholars examines whether litigation *should* occur (Sanchirico 2007). At the heart of this literature lies a distinction between the social cost of litigation and the private incentive to file suit. Externalities occur, on the one hand, because the plaintiff does not take into account the defendant’s legal costs (negative externality), and, on the other hand, he does not consider the socially valuable precedents that the adjudication of his case creates (positive externality). The normative approach is considered, among others, in Posner (1973), Shavell (1981), Menell (1983), and Kaplow (1986).

The following section starts with the basic model of the decision-making process of (1) a plaintiff to file suit and (2) both plaintiff and defendant to settle or litigate. When a plaintiff files suit, disputants choose between settling the case out of court (at any time preceding a judgment) or to proceed to trial when negotiations break down. In the latter case, the dispute is resolved through a court verdict (Priest and Klein 1984).

Basic Decision-Making Model

This paragraph presents the basic economic decision-making model of litigation, as first

discussed by Landes (1971), Posner (1973), Gould (1973), and Shavell (1982). Section “[Decision to Sue](#)” elaborates on the decision to sue, while “[Decision to Settle](#)” focuses on the decision to settle.

Decision to Sue

Consider the example in which a pedestrian claims that injury was inflicted by an allegedly negligent car driver. First, the pedestrian has to decide whether or not to assert a claim. Following Shavell (1982), a rational plaintiff will decide to file suit based on the outcome of a recursively solved sequence game that weighs present costs (i.e., legal costs) against expected future benefits. When the expected value of the trial is positive (i.e., benefits outweigh costs), a rational plaintiff will file suit. For a risk neutral plaintiff, the expected value of the trial (EVT) is determined by the value of a judgment in his favor (J) discounted by the probability of winning (P) minus his legal costs (C_p):

$$EVT = (P * J) - C_p \quad (1)$$

This model assumes that the plaintiff can perfectly assess the value of the parameters included. Suppose that the plaintiff has a 60% chance of winning 20.000 euro, which equals the damages endured, such as health-care expenditures (most authors assume that in case of a verdict favoring the plaintiff, the award will be equal to his damages, which are known by the judge after trial). Furthermore, the plaintiff expects to pay 2.000 euro of legal costs (e.g., court fees and costs of legal advice). The plaintiff thus expects to gain 10.000 euro from going to court, i.e., his expected benefits (20.000 euro * 0.6) minus his expected costs (2.000 euro). Given that the expected value of trial is positive, the plaintiff has a credible threat and will therefore sue. After the plaintiff has filed suit, disputants will have to decide whether they will settle the case out of court or proceed to trial.

Decision to Settle

During the bargaining phase that elapses between the filing of the suit and the start of the trial, parties

can still decide to settle. This decision depends on the “threat values,” i.e., the plaintiff’s minimum settlement demand and the defendant’s maximum settlement offer.

The defendant’s highest settlement offer equals his expected loss from a trial minus his settlement costs (S_d). In turn, the expected loss consists of the value of an adverse judgment (J) discounted by the probability of losing (i.e., the plaintiff’s winning probability (P)) and his legal costs (C_d).

$$ELT = (P * J) + C_d \quad (2)$$

Logically, the plaintiff’s minimum settlement demand equals the expected value of a trial (see Eq. 1) plus his settlement costs (S_p). Consequently, a necessary condition for reaching settlement is $EVT + S_p < ELT - S_d$, or in other words:

$$(P * J) - C_p + S_p < (P * J) + C_d - S_d \quad (3)$$

Let us assume that settlement and litigation costs are, respectively, 500 euro and 2.000 euro for each party. Obviously, the settlement and litigation costs need not be identical for both parties, but we assume this for reasons of simplicity. The plaintiff’s minimum settlement demand will then be 10.500 euro. Faced with an expected loss of 14.000 euro at trial ((0,6 * 20.000 euro) + 2.000), the defendants’ maximum settlement offer is 13.500 euro. Consequently, parties can avoid a trial by settling for any amount between 10.500 and 13.500 euro, which is referred to as the settlement range. Logically, all the values comprised by the settlement range represent mutually beneficial outcomes.

The conditions for settlement (assuming perfect information) in Eq. 3 can be reformulated as follows:

$$S_p + S_d < C_p + C_d \quad (4)$$

Equation 4 shows that under perfect information (i.e., agreement of parties on the expected value of the judgment and the plaintiff’s likelihood of prevailing), a settlement will always be reached.

In other words, $S_p + S_d < C_p + C_d$ will always be satisfied since trials are undeniably costlier. Note that it is not necessary that parties correctly estimate the expected value of the judgment and the plaintiff's likelihood of prevailing. Rather, their estimations must be consistent. Moreover, many scholars make the assumption that settlement costs are nonexistent because they are negligible compared to litigation costs (Cooter and Rubinfeld 1989).

Divergence in Expectations

Both the decision to sue and subsequently to settle or litigate depends on litigants' expectations. The previous section demonstrated that rational and risk neutral parties will not proceed to trial since the underlying condition ($C < S$) will not be satisfied. Nonetheless, many conflicts are brought to trial. One of the reasons for the occurrence of trials stems from the assumption made in section "Basic Decision-Making Model" that litigants have equal expectations about the probability of winning (P) and the value of the judgment (J). This assumption is relaxed in the following section. We start with diverging estimates of *winning probabilities* and proceed with deviating assessments of the *value of the judgment*.

At this point, we solely focus on diverging estimates of P and J as a cause of litigation, without explaining discrepancies. These reasons for diffuse distribution of lawsuit valuations are discussed in section "Causes of Divergent Expectations."

Plaintiff's Likelihood of Prevailing

As discussed in section "Basic Decision-Making Model," settlements will always occur when parties agree on the plaintiff's probability of prevailing and the value of a judgment. However, in reality, the plaintiff's assessment of winning a trial (P_p) might and, most likely will, differ from the defendant's (P_d).

In case of divergent estimates of winning probabilities, the condition for settlement in Eq. 3 that assumed $P_p = P_d$ becomes

$$(P_p * J) - C_p + S_p < (P_d * J) + C_d - S_d \quad (5)$$

or

$$(P_p - P_d) * J < (C_p + C_d) - (S_d + S_p) \quad (6)$$

Two situations may occur. First, if the plaintiff's estimation of prevailing is smaller than the defendant's ($P_p < P_d$), the left-hand side of Eq. 6 necessarily becomes negative. In this case, the condition for settlement is always satisfied, since litigation costs ($C_p + C_d$) are higher than settlement costs ($S_d + S_p$).

If, however, the plaintiff's estimated chance of winning exceeds the opponent's ($P_p > P_d$), his likelihood of going to trial enhances (Landes 1971). Whether this will eventually lead to a trial depends on the value of the judgment and the legal costs. *Ceteris paribus*, a higher award by the court will narrow the settlement range and increase the likelihood of trial. Higher litigation costs and lower settlement costs, however, have the opposite effect. Assume that the plaintiff estimates the likelihood of a judgment in his favor at 80% (i.e., P_p), while the defendant estimates the likelihood of a judgment in the favor of the plaintiff at 30% (i.e., P_d). In this case, the plaintiff's minimum demand is 14,500 euro, while the defendant's maximum offer is 7,500 euro. Hence, parties will be unable to reach a settlement agreement.

Value of the Judgment

Not only winning probabilities (P) can diverge but also valuations of the judgment (J); for reasons, we will discuss in section "Causes of Divergent Expectations."

Ceteris paribus, divergent estimates of J result in the following condition for settlement:

$$(P * J_p) - C_p + S_p < (P * J_d) + C_d - S_d \quad (7)$$

or

$$(J_p - J_d) * P < (C_p + C_d) - (S_d + S_p) \quad (8)$$

Similar to diverging estimates of P , two scenarios are analyzed when parties have different estimates

of damages. First, if the plaintiff expects to gain less than the defendant expects to lose ($J_p < J_d$), the left-hand side of Eq. 8 becomes negative. It follows that the necessary condition for settlement will always be met, since litigation costs ($C_p + C_d$) are higher than settlement costs ($S_d + S_p$). In the opposite case ($J_p > J_d$), the plaintiff is relatively optimistic about the expected judgment and thus trial becomes a more favorable option. The occurrence of trial is dependent upon the likelihood of a plaintiff win and legal costs. *Ceteris paribus*, when the likelihood of a plaintiff win increases, the settlement range narrows and hence the likelihood of settlement decreases. Higher litigation costs and lower settlement costs have an opposite effect.

Causes of Divergent Expectations

The previous two paragraphs show that optimism ($J_p > J_d$ and $P_p > P_d$) can obstruct settlement. Naturally, the question remains *why* parties' estimations of trial outcomes may diverge, in other words, why optimism occurs. Recent models explicitly include the bargaining process. In contrast to the earlier models, they show that settlement negotiations may break down, even in case of a positive settlement range (e.g., because parties cannot agree on the division of the surplus). Three different models with perfect, imperfect, and asymmetric information are distinguished. Perfect information models (such as the basic model described in section "[Basic Decision-Making Model](#)") assume that there is no uncertainty concerning the value of damages, winning probabilities, legal costs, etc. As literature shows, deviations are possible even with perfect and identical information. This will be discussed in section "[Perfect and Imperfect Information](#)." Nonetheless, diverging assessments are usually considered a consequence of private information. Imperfect information models assume, on the one hand, that parties attach probabilities to uncertain parameters in the model. However, they have analogous insight into the distribution of these probabilities, indicating the symmetric character of uncertainty. On the other hand, asymmetric information models acknowledge that disputants assess probability distributions

based on private information, meaning that uncertainty applies to these distributions as well. Section "[Asymmetric Information](#)" elaborates on asymmetric information as a cause of optimism.

Perfect and Imperfect Information

With perfect and imperfect (or symmetrical) information, causes of divergence cannot be attributed to informational considerations. In perfect information models, there exists no uncertainty, while under imperfect information, some symmetrical uncertainty occurs about the distribution of parametric probabilities (i.e., information is incomplete). The imperfect information models assume, for instance, that neither party knows the amount a judge would award. It follows that parties will use a probability distribution function to assess the likelihood of different values of damages awarded. While some simplified models assume that J can take on two values ("high" or "low"), others consider a range of possible values. As mentioned above, the uncertainty is assumed to be symmetric, meaning that parties have the same estimate of *expected* damages because they have the same information about the judge (Daughety 2000).

Perfect and imperfect information models assume that parties' expectations can still differ due to, for instance, a self-serving estimation bias. Hence, the divergence is a consequence of a difference in opinion, *given identical information* (Lederman 1998). In this case, a trial becomes a more favorable option for the plaintiff who is relatively optimistic about either a judgment in his favor (P) or the value of the judgment (J) (Landes 1971).

The drivers of self-serving estimation bias are of a psychological rather than economic nature. In general, individuals overestimate positive outcomes attributed to personal factors (e.g., the share of credit for a collaborative task or their own driving capabilities). Accordingly, parties of a dispute are likely to estimate their winning probabilities and the value of the claim in a self-serving manner (Loewenstein et al. 1993). In their experimental study, Loewenstein et al. (1993) examine parties' divergent estimates of winning probabilities and damages, while controlling for

the information about the case (i.e., there is no private information). The authors assign the self-serving bias mainly to the concept of fairness: disputants systematically overestimate parameters in their favor because they are concerned with reaching a fair settlement, rather than predicting estimates based on actual facts. Hovenkamp (1991), among others, attributes the cognitive bias to the concept of endowment, which states that individuals are willing to pay less to obtain some entitlement compared to what they are willing to accept to forsake that same entitlement.

Self-serving estimation bias aside, other factors can explain divergent assessments in the absence of uncertainty. Examples are disputant's risk preferences or discount rates. Although most studies assume risk neutrality and identical discount rates for both plaintiff and defendant, relaxing these assumptions can cause divergence of estimates. The most important assumptions that crucially impact the outcome of the decision-making model are discussed in section "[Assumptions of the Decision-Making Model](#)."

Asymmetric Information

The absence of informational structures and bargaining processes led the early models to conclude that the occurrence of optimism is the most important reason of settlement failure. A later bulk of papers expanded the original models by including the bargaining process and accounting for asymmetrical information, which in turn led to a better understanding of settlement breakdown (Bebchuk 1984). Interestingly, many predictions based on these advanced models contradict the outcomes of earlier nonstrategic studies.

In asymmetric information models, the existence of private information causes the assessment of probabilities to differ among parties. Most models apply one-sided information asymmetry, which involves one uninformed party who attaches probabilities to different values of one unknown parameter. The other party, on the contrary, is perfectly informed. Most authors state that damages (J) are known to the plaintiff but not to the defendant, while liability (P) is known to the defendant but not to the plaintiff (Daughety 2000;

Salant and Rest 1982; Spier 1992). Furthermore, divergence in J does not only depend on the expected value of damages but also on private information about the judge, future performances of witnesses, etc. (Spier 1992). Last, some models apply asymmetric information to the aspect of legal costs (C) since the amount the opponent is prepared to spend is usually unknown.

Finally, private information does not only exist between defendants and plaintiffs but also between clients and their lawyers. This agency problem causes the attorney to maximize his own utility instead of his client's (Miller 1987; Hay 1996). For example, although a rapid settlement may be in the best interest of the client, an attorney may want to establish the reputation of a tough bargainer (Spier 1992). Furthermore, lawyers could be eager to maximize their own income by overestimating winning probabilities in order to persuade clients to go to court. The impact of fee structures plays a significant role in this agency problem (see, e.g., Emons 2000).

Assumptions of the Decision-Making Model

The goal of modeling litigation decisions is to determine under which conditions disputes are settled. Obviously, results are very dependent upon the underlying assumptions. In this section, we elaborate on the most important assumptions that can crucially affect the results of the analysis.

Bargaining Games

The first models, which were based on perfect or imperfect information, applied non-Bayesian game theory to analyze settlement decisions (Polinsky and Shavell 2007). In a cooperative game, parties explicitly or implicitly commit themselves *ex ante* to reach an efficient solution ("no money is left on the table") (Daughety 2000). Although these models generally focus on the decision to settle instead of the settlement amount, some authors state that the surplus, which is determined by the settlement range, is evenly divided. This bargaining solution is used in, for example, Gould (1973), Landes (1971), and Posner (1973).

A cooperative game setting suffices in the perfect information case. However, under private information, the game is characterized by strategic behavior, such as deception and bluffing, since it is in the interest of the better informed party to exaggerate its claim (Salant and Rest 1982). A later flow of papers therefore applied a noncooperative game theory approach to allow for strategic behavior and inefficient outcomes (Cooter et al. 1982; P'ng 1983; Bebchuk 1984). In these cases, settlement negotiations may break down even though the settlement range is positive. Instead of one single cooperative solution, multiple equilibriums can be reached.

While some scholars have used simultaneous bargaining games, others have suggested a sequential game, in which the second mover is aware of the opponent's first offer. Indeed, "the process of bargaining, however, reveals information. In particular, the uninformed disputant may try to deduce the true state of nature from the actions of his informed opponent who, in turn, may attempt to exploit his initial advantage by manipulating the flow of information" (Salant and Rest 1982). Therefore, bargaining games with multiple periods allow for a learning effect and acknowledge that the uncertain parameters are determined endogenously (Salant and Rest 1982).

Furthermore, these sequential bargaining models also differ in which party moves first. In a signaling game, the informed party makes the first offer, whereas the uninformed party moves first in a screening game. Most authors assume information asymmetry concerning the likelihood of success at trial. The defendant can assess this parameter more accurately than the plaintiff, because he holds more information about his alleged liability. Consequently, the plaintiff is the uninformed party. However, when there is uncertainty about the damages, the plaintiff is usually the better informed party.

Risk Preferences

Most models assume risk neutrality of disputants. Gould (1973), however, showed that under risk aversion, settlement will always occur if disputants agree on winning probabilities. In case one

party is risk averse and the other is a risk seeker, the solution depends on the relative curvature.

Value of the Judgment

In most litigation models, the stake of a trial, i.e., the amount awarded, is symmetrical. However, parties may anticipate different expected damages, even under perfect information, if, for example, they apply different discount rates. Furthermore, a certain judgment may have precedential value to one of the disputants (Posner 2014). Whenever there are long-term considerations (e.g., reputation), the outcome of trial does not solely consist of the monetary value of the damages (Daughety 2000).

Likelihood of Trial and Legal Costs

Basic litigation models often assumed that legal costs are fixed and exogenous. However, this assumption appears unrealistic since "[...] parties expend resources on legal research to produce arguments or favorable facts to a court or other decision-making body such as a jury or administrative agency" (Katz 1988). Furthermore, legal costs are more likely to be endogenous since parties' trial expenditures are dependent upon, among others, the stakes of the trial, their belief about winning the trial, their initial wealth, and their opponents' trial expenditures (Braeutigam et al. 1984; Choi and Sanchirico 2004; Posner 2014). In other words, the amount of money parties spend is itself the result of a noncooperative game in which they maximize the net benefits of trial (Polinsky and Shavell 2007). Analogously, estimates of winning probabilities are endogenous as well. As Gould (1973) described, winning probabilities are a function of the stakes of the case and the resources spent by each party.

Discount Rates and Court Delay

Court delay enters the settlement model since it obliges disputants to discount the expected damages. Basic models, only taking into account discounting by the plaintiff, assumed that delay decreased the value of the judgment and thus enhanced the likelihood of a settlement. However, the effect of delay on the value of J depends on the discount rate of each party. In case the defendant

invokes a higher discount rate than the plaintiff, the settlement range is impacted the most (Posner 1973). Furthermore, delay increases the plaintiff's winning probability as a consequence of diminishing quality of evidence (Vereeck and Mühl 2000). Consequently, the settlement range increases and litigation becomes less likely. Although delay increases the likelihood of settlement by reducing the stakes and the quality of evidence, an opposing effect exists because it increases uncertainty about the judgment (Posner 2014). In bargaining models, the impact of delay on the decision to settle is examined, as well as the timing of a possible settlement (Spier 1992).

Procedural Rules

Settlement models can also be used to analyze the effects of legal rules, such as trial cost allocation, on the likelihood of settlement. Under the "American rule," as applied in basic models, each party bears its own costs, whereas the loser pays all costs in a "British indemnity system" (Shavell 1982 (Shavell also discusses a system "favoring the plaintiff" and "favoring the defendant"); Bebchuk 1984; Braeutigam et al. 1984; Miller 1986; Cooter et al. 1982). "Offer-of-settlement" rules, on the contrary, allocate costs based on settlement offers: if a disputant refuses a settlement offer and is granted a lower award afterwards from judgment at trial, a compensation of the opponent could be imposed (Hay and Spier 1998; Daughety and Reinganum 1994). The effect of other procedural rules on the settlement decision has been examined as well. For example, the effect of discovery requirements and disclosure rules aimed at reducing information asymmetry (Bebchuk 1984; Cooter and Rubinfeld 1994; Hay 1994), rules on the burden of proof, and rules on alternative dispute resolution (Hay and Spier 1998).

Cross-References

- ▶ [Choice Under Risk and Uncertainty](#)
- ▶ [Good Faith and Game Theory](#)
- ▶ [Legal Disputes](#)
- ▶ [Signalling](#)

References

- Bebchuk LA (1984) Litigation and settlement under imperfect information. *RAND J Econ* 15:404–415
- Bebchuk LA (1987) Suing solely to extract a settlement offer. National Bureau of Economic Research, Cambridge, MA
- Bebchuk LA (1996) A new theory concerning the credibility and success of threats to sue. *J Leg Stud* 25:1–25
- Braeutigam R, Owen B, Panzar J (1984) An economic analysis of alternative fee shifting systems. *Law Contemp Probs* 47:173
- Choi A, Sanchirico CW (2004) Should plaintiffs win what defendants lose? Litigation stakes, litigation effort, and the benefits of decoupling. *J Leg Stud* 33:323–354
- Cooter RD, Rubinfeld DL (1989) Economic analysis of legal disputes and their resolution. *J Econ Lit* 27:32
- Cooter RD, Rubinfeld DL (1994) An economic model of legal discovery. *J Leg Stud* 23:435–463
- Cooter R, Marks S, Mnookin R (1982) Bargaining in the shadow of the law: a testable model of strategic behavior. *J Leg Stud* 11:225–251
- Daughety AF (2000) Settlement. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics*, vol I, The history and methodology of law and economics. Edward Elgar, Cheltenham
- Daughety AF, Reinganum JF (1994) Settlement negotiations with two-sided asymmetric information: model duality, information distribution, and efficiency. *Int Rev Law Econ* 14:283–298
- Emons W (2000) Expertise, contingent fees, and insufficient attorney effort. *Int Rev Law Econ* 20:21–33
- Gould JP (1973) The economics of legal conflicts. *J Leg Stud* 2:279–300
- Hay BL (1994) Civil discovery: its effects and optimal scope. *J Leg Stud* 23:481–515
- Hay BL (1996) Contingent fees and agency costs. *J Leg Stud* 25:503–533
- Hay B, Spier K (1998) Settlement of Litigation. In: Newman P (ed) *The New Palgrave Dictionary of Economics and the Law*, vol III. Stockton Press, New York
- Hovenkamp H (1991) Legal policy and the endowment effect. *J Leg Stud* 20:225–247
- Kaplow L (1986) The social versus the private incentive to bring suit in a costly legal system. *J Leg Stud* 15:371–385
- Katz A (1988) Judicial decisionmaking and litigation expenditure. *Int Rev Law Econ* 8:127–143
- Katz A (1990) The effect of frivolous lawsuits on the settlement of litigation. *Int Rev Law Econ* 10:3–27
- Landes WM (1971) An economic analysis of the courts. *J Law Econ* 14:61–108
- Lederman L (1998) Which cases go to trial: an empirical study of predictors of failure to settle. *Case W Res L Rev* 49:315
- Loewenstein G, Issacharoff S, Camerer C, Babcock L (1993) Self-serving assessments of fairness and pre-trial bargaining. *J Leg Stud* 22:135–159
- Menell PS (1983) A note on private versus social incentives to sue in a costly legal system. *J Leg Stud* 12:41–52

- Miller GP (1986) An economic analysis of Rule 68. *J Leg Stud* 15:93–125
- Miller GP (1987) Some agency problems in settlement. *J Leg Stud* 16:189–215
- Ordover JA, Rubinstein A (1983) On bargaining, settling, and litigation: a problem in multistage games with imperfect information. New York University, C.V. Starr Center for Applied Economics, Economic Research Report 83-07
- P'ng IP (1983) Strategic behavior in suit, settlement, and trial. *Bell J Econ* 14:539–550
- Polinsky AM, Shavell S (2007) *Handbook of law and economics*. Elsevier, Amsterdam
- Posner RA (1973) An economic approach to legal procedure and judicial administration. *J Leg Stud* (0047–2530) 2:399
- Posner RA (2014) *Economic analysis of law*. Wolters Kluwer Law and Business, New York
- Priest GL, Klein B (1984) The selection of disputes for litigation. *J Leg Stud* 13:1–55
- Reinganum JF, Wilde LL (1986) Settlement, litigation, and the allocation of litigation costs. *RAND J Econ* 17:557–566
- Salant SW, Rest G (1982) Litigation of questioned settlement claims: a bayesian nash-equilibrium approach. Rand Corp, Santa Monica
- Sanchirico CW (2007) *Economics of evidence, procedure and litigation: two volume set*. Elgar Northampton
- Schweizer U (1989) Litigation and settlement under two-sided incomplete information. *Rev Econ Stud* 56:163–177
- Shavell S (1981) The social versus the private incentive to bring suit in a costly legal system. National Bureau of Economic Research, Cambridge, MA
- Shavell S (1982) Suit, Settlement, and Trial: A Theoretical Analysis Under Alternative Methods for the Allocation of Legal Costs. *J Leg Stud* 11:59–60
- Spier KE (1992) The dynamics of pretrial negotiation. *Rev Econ Stud* 59:93–108
- Vereeck L, Mühl M (2000) An economic theory of court delay. *Eur J Law Econ* 10:243–268

Litigation Expenditures Under Alternative Liability Rules

Jef De Mot

Postdoctoral Researcher FWO, Faculty of Law,
University of Ghent, Belgium

Abstract

For a long time there was a widespread consensus in the literature that a comparative negligence standard imposes higher administrative

costs than a simple negligence standard and a contributory negligence standard. However, in a setting where the parties can choose their level of litigation expenditures and the litigation expenditures influence the outcome of the case, it can be shown that none of the negligence rules unambiguously leads to higher expenditures. Which rule creates larger expenditures strongly depends on the quality of the case (taking into account both the defendant's negligence and the plaintiff's negligence).

Introduction

Law and economics scholars have long debated the efficiency benefits of different negligence standards, mainly simple negligence, negligence with a defense of contributory negligence, and negligence with a defense of comparative negligence. Under a simple negligence rule, the injurer is liable if and only if his/her level of precaution is below the legal standard regardless of the precaution level exercised by the victim. Under a rule of (negligence with a defense of) contributory negligence, the negligent injurer escapes liability by proving that the victim's precaution fell short of the legal standard of care. Comparative negligence divides the cost of harm between the parties in proportion to the contribution of their negligence to the accident. A lively debate exists even to date as to whether comparative negligence or contributory negligence provides better incentives to adopt optimal care (see, e.g., Artigot i Golobardes and Pomar 2009; Bar-Gill and Ben-Shahar 2003). Whereas the efficiency-promoting aspects of various tort standards remain a controversial matter, for a long time there was a widespread consensus that a comparative negligence standard imposes higher administrative costs than a simple negligence standard (Landes and Posner 1987; White 1989) or a contributory negligence standard (Shavell 1987; Bar-Gill and Ben-Shahar 2001). White (1989), for instance, argues that a comparative negligence standard likely generates higher litigation and administrative costs than contributory negligence because courts must assess the degree of

negligence by both parties and not just whether the parties were negligent. Recently, De Mot (2013) has argued that this wisdom clearly applies to settings involving fixed litigation costs but does not hold in a more realistic setting where the parties can choose their level of litigation expenditures and the litigation expenditures influence the outcome of the case (recent contributions to the economic literature on litigation stress the importance of treating litigation expenditures as endogenously determined. See Sanchirico (2007)). In the next sections “Expenditures Towards Establishing Negligence on Behalf of the Defendant,” “Expenditures to Determine Negligence on Behalf of the Plaintiff,” and “Total Expenditures and the Merits of the Claim,” I compare the influence of the three negligence rules on expenditures in an intuitive, nonformal way. I first compare the expenditures towards establishing negligence of behalf of the defendant (section “Expenditures Towards Establishing Negligence on Behalf of the Defendant”) and the plaintiff (section “Expenditures to Determine Negligence on Behalf of the Plaintiff”), after which I examine the total expenditures and how these vary with the quality of the case (section “Total Expenditures and the Merits of the Claim”). Section “Conclusion” concludes.

Expenditures Towards Establishing Negligence on Behalf of the Defendant

The litigation investments in determining negligence on behalf of the defendant are the highest under simple negligence, the lowest under contributory negligence, and somewhere in between for comparative negligence. Legal expenditures are highest under simple negligence because, unlike under contributory and comparative negligence, the plaintiff does not run the risk of being held liable himself. By contrast, when contributory and comparative negligence standards apply, the possibility that the plaintiff will be held liable reduces the expected benefit of the expenditures devoted to establishing negligence on behalf of the defendant. Intuitively, the value of spending additional resources on determining negligence

by the defendant is lower under contributory and comparative because some (or *all* in the case of contributory negligence) of the damage is shifted to the plaintiff if the latter is held liable as well.

Next, the expected value of expenditures on determining the defendant’s negligence is lower under contributory negligence than comparative negligence. There are two reasons for this. First, (only) under contributory negligence, expenditures into establishing negligence on behalf of the defendant have absolutely no value when the plaintiff is also found negligent. Second, additional investments may convince the court to conclude that the defendant’s negligence is relatively large (or small). Under contributory negligence, the degree of negligence doesn’t matter for the ultimate division of the loss. Under comparative negligence, however, it matters a great deal, because the relative degree of negligence influences the division of the loss.

Expenditures to Determine Negligence on Behalf of the Plaintiff

The expenditures to determine negligence on behalf of the plaintiff are obviously lowest under simple negligence. Under this rule, expenditures towards establishing negligence on behalf of the plaintiff equal 0 – since a simple negligence standard does not inquire into the care level of the plaintiff. The expenditures can be either larger or smaller under comparative negligence than under contributory negligence. Because the expected value of expenditures to establish or rebut a plaintiff’s negligence is lower under comparative negligence than contributory negligence, the expenditures can be smaller under comparative negligence. That is because, if the plaintiff is held liable under contributory negligence, he or she bears the entire loss; whereas under comparative negligence, a plaintiff who is held negligent only bears part of the losses (if the defendant is also held negligent). On the contrary, the expenditures can also be larger under comparative negligence because under that rule additional investments can lead the court to conclude that the plaintiff’s fault is relatively small (or large).

Total Expenditures and the Merits of the Claim

In a model with fixed expenditures, legal expenditures will always be larger under contributory negligence than under simple negligence since there are more issues for the court to decide under the former rule. However, in a model where litigation expenditures are endogenous, the total costs may be either lower or higher under a contributory negligence standard. While under simple negligence there are no expenditures to determine negligence on behalf of the plaintiff, the expenditures to determine negligence on behalf of the defendant are larger under this rule than under contributory negligence. Similarly, total trial expenditures can either be higher or lower under comparative negligence than under simple negligence.

Finally, are total trial expenditures always lower under contributory negligence than under comparative negligence? Whether the total expenditures are smaller under a comparative negligence standard than under contributory negligence depends on the relative strength of the issues. The stronger (weaker) the plaintiff's claim (taking both the defendant's and the plaintiff's behavior into account), the more likely it is that the expenditures will be smaller (higher) under comparative negligence than under contributory negligence. The intuition is the following. First, when the actual degree of fault of the defendant increases, the importance of the outcome of the issue of the plaintiff's negligence increases under both rules but more so under contributory negligence than under comparative negligence (because of the all-or-nothing character of contributory negligence). Second, when the actual level of fault of the plaintiff decreases, the importance of the outcome of the issue of the defendant's negligence increases under both rules but more so under contributory negligence than under comparative negligence (due to the sharing element of comparative negligence). Third, the greater the difference between the true levels of fault of the parties, the smaller the value of

investing in trying to convince the court that one's level of fault was more modest. The effects of such investments are more heavily discounted when the difference between the levels of fault is relatively large. A simple example can illustrate this. Suppose we can represent the level of fault of the plaintiff by the number 0.2 and the defendant's level of fault by the number 0.8 and that the court uses the following formula to determine the plaintiff's share: $0.8/(0.8 + 0.2) = 0.8$. Now suppose the defendant can make an investment i that will reduce his level of fault (as ultimately perceived by the court) with 0.1 ($0.8 - 0.1 = 0.7$). Then the plaintiff's share would equal $0.7/(0.7 + 0.2) = 0.78$. Now we look at a case in which the levels of fault are more close: 0.5 for the plaintiff and 0.8 for the defendant. The plaintiff's share will be $0.8/(0.8 + 0.5) = 0.62$. Clearly, the investment i , which again reduces the level of fault of the defendant with 0.1, has a greater payoff for the defendant in such a closer case: $0.7/(0.7 + 0.5) = 0.58$. So here we get a reduction of 0.04, while in the previous case the reduction was only 0.02. Clearly, there will be more investment in the latter type of cases.

Conclusion

When litigants can choose their level of litigation expenditures and these expenditures influence the outcome of the case, none of the negligence rules unambiguously leads to higher expenditures. Which rule creates larger expenditures strongly depends on the quality of the case, taking into account both the defendant's negligence and the plaintiff's negligence.

Cross-References

- [Judicial Decision-Making](#)
- [Litigation Decision](#)
- [Rent Seeking](#)
- [Strict Liability Versus Negligence](#)
- [Tort Damages](#)

References

- Artigot i Golobardes M, Pomar FG (2009) Contributory and comparative negligence in the law and economics literature. In: Faure M (ed) Tort law and economics / In: De Geest G (ed) Encyclopedia of law and economics. Edward Elgar, Cheltenham, pp 46–79
- Bar-Gill O, Ben-Shahar O (2001) Does uncertainty call for comparative negligence? vol 346, Discussion paper series. Harvard Law School, John M. Olin Center for Law, Economics and Business, Cambridge, MA
- Bar-Gill O, Ben-Shahar O (2003) The uneasy case for comparative negligence. *Am Law Econ Rev* 5: 433–469
- De Mot J (2013) Comparative versus contributory negligence: a comparison of the litigation expenditures. *Int Rev Law Econ* 33:54–61
- Landes WM, Posner RA (1987) The economic structure of tort law. Harvard University Press, Cambridge, MA, p 330
- Sanchirico C (2007) The Economics of Evidence, Procedure and Litigation, Vol. 1. Edward Elgar, Northampton, MA
- Shavell S (1987) Economic analysis of accident law. Harvard University Press, Cambridge, MA, p 312
- White MJ (1989) An empirical test of the comparative and contributory negligence rules in accident law. *Rand J Econ* 20:308–330

Looting

- [Piracy, Modern Maritime](#)

M

Mafias

Guglielmo Barone¹ and Gaia Narciso²

¹Bank of Italy and RCEA, Bologna, Italy

²Trinity College Dublin, Dublin, Ireland

Abstract

Organized crime has far reaching economic, political, and social consequences. This entry presents the basic facts on the economics of mafias, with a special focus on the Sicilian case that may serve as a window into other types of criminal organizations. First, the entry provides a brief sketch of the theoretical modelling of the mafia. Then, it reviews the theoretical and empirical work testing the role of geography in the historical origins of the mafia. Finally, the entry explores the economic impact of the mafia in terms of missed opportunities of development, both in the short and in the long run. Different transmission channels are explored.

Definition

A secret and criminal organization which is engaged in a number of illicit activities such as racketeering, smuggling, trafficking in narcotics, and money laundering. It has a complex hierarchical organization, and its members are expected to follow a number of internal rules. The mafia

originated in Sicily in the second part of the nineteenth century and expanded to the United States through Italian emigration. Nowadays, mafia-type organizations are widespread in many countries, especially in Southern Italy, Russia and East Europe, Latin America, China, and Japan.

Introduction

Organized crime entails deep economic, political, and social consequences. Its presence is pervasive and threatens the functioning of democratic institutions (Allum and Siebert 2003; Bailey and Godson 2000; Fiorentini and Peltzman 1996). Due to its varied features and the lack of empirical data, very few empirical studies have, until recently, investigated organized crime and its impact on the economy. This emerging literature deals with different issues. Some papers focus on the theoretical framework (Dal Bò et al. 2006; Skaperdas 2001), while empirical studies investigate either the origins of mafia, stressing the role of natural resources and land value as key determinants, or its negative economic consequences. Overall, the mafia is found to have a negative impact on growth and GDP per capita; the underlying mechanism may work through lower government efficiency, lower foreign direct investment, distortions in the allocation of public funds, or unfair markets competition.

This entry presents the evidence on the economics of mafias, with a special focus on the

Sicilian case. The Sicilian Mafia is a complex phenomenon that acts, at times, within Italian institutions, at times against them. The Sicilian Mafia serves as a window into other types of criminal organizations, such as the Russian Mafia or the Japanese Yakuza (Maruko 2003), which are rooted in the political and socioeconomic spheres. First, the entry provides a brief sketch of the theoretical modeling of the mafia. Then, it reviews the theoretical and empirical work testing the role of geography in the origins of the mafia. Finally, the entry explores the economic impact of the mafia in terms of missed opportunities of development, both in the short and in the long run.

Theoretical Models of Mafia and the Analysis of Its Origins

On a theoretical ground, this entry adopts the widespread view according to which the mafia is treated as an industry that produces and sells a number of goods and services, such as private protection services, narcotics, or connections with politics (Gambetta 1993, 2000). Such a view is consistent with the historical origins of the mafia. Under this perspective, the supply of protection services plays a key role. According to a consolidated opinion, the mafia emerged in Sicily after (or around) the unification that took place in 1861. In 1876, Leopoldo Franchetti, a Tuscan intellectual, traveled to Sicily to conduct a private inquiry into the political and administrative conditions of the island. The report was published the following year and represents an original and detailed picture of the state of Sicily at that time (Franchetti 2011). The report is the first document of the issues related to the mafia and its permeation through the Sicilian society. The demand for private protection arose in Sicily for two interrelated motives. First, before Italian unification, a series of anti-feudal laws endorsed the opening up of the market for land, which led to an increase in the fragmentation of land property. Second, following the Italian unification in 1861, a weak protection of property rights and a vacuum of power from the recently constituted

State amplified landowners' need for protection against illicit expropriation. In such historical moments, the mafia emerged as an industry offering a number of services the State was not able to offer.

This theory has received robust empirical support. Some authors analyze the relationship between land fragmentation and mafia activity in the nineteenth century. From a theoretical viewpoint, in fact, it can be shown that the rise in the number of landowners (following the end of Feudalism in 1812) increased competition for protection, which ultimately led to an upsurge in mafiosi's profits. The data collected from the parliamentary survey conducted by Damiani in 1881 show that the historical presence of mafia in Sicily was indeed more likely to be found in towns where land was more fragmented (Bandiera 2003). The land fragmentation was not the only determinant of the upsurge in the mafia. The end of feudalism and the demise of the Bourbon Kingdom in the South of Italy were accompanied by a rapid increase in the demand of sulfur, of which Sicily became a major exporter between 1830 and 1850. Sulfur was used as an intermediate input in the industrial and chemical production, which was flourishing in France and England in the nineteenth century. Sulfur mines in Sicily were mainly superficial and did not need sophisticated extraction technology. By the end of the nineteenth century, over 80% of world sulfur production originated from Sicily. Consistently with the idea of mafia as the supply side of a market of protection and extortions, data on the Sicilian Mafia in the late nineteenth century (Cutrera 1900) shows that the intensity of mafia activity was higher in municipalities with sulfur mines, where the demand for private protection was higher Buonanno et al. (forthcoming). In a similar fashion, other authors have associated the origins of the mafia with the presence of citrus fruits, which were highly valuable (Dimico et al. 2012). Lemon trade between Sicily and the United States flourished: 34% of imported citrus fruits originated from Italy. This explanation of the origins of the Sicilian Mafia is indeed in line with that outlined in a parliamentary inquiry dated 1875, according to which "Where wages are low and

peasant life is less comfortable, [...], there are no symptoms of mafia [...]. By contrast, [...] where property is divided, where there is plenty of work for everyone, and the orange trees enrich land-owners and growers alike – these are the typical sites of mafia influence.”

Although the mafia historically emerged in the South of Italy, it gradually migrated to other Italian regions (and to the United States). The first main mechanism of diffusion of the mafia was the mass migration from Southern to Northern Italy, which took place between the 1950s and 1970s. It is estimated that four million people migrated from the South to the North, during the economic boom. The mass migration led to a change in the population composition of the receiving regions. As a result, mafia-type organizations were more likely to emerge in areas that were more migrant-abundant. The second mechanism of the expansion of the mafia in the North is due to the *confino* law: the imprisonment policy for mafiosi was based on the *confino*, a policy according to which mafia-related criminals were imprisoned in a different region from the one they originated from, in order to loosen the links with the local mafia. However, the *confino* had the perverse effect of spreading mafia activity in other Italian regions (Varese 2006, 2011).

Economic Consequences of Mafia: GDP Growth

Besides analyzing the origins of mafia, economists also examined its consequences in terms of economic growth, together with the potential underlying mechanisms. This is the second main strand of the economic literature on organized crime. Assessing the economic impact of mafia activity suffers from two main issues. The first issue is concerned with the lack of data. Only recently, new data on criminal activities have been made available. However, even where available, data on crime often suffer from measurement error. For example, the number of crimes could be underreported, in particular in relation to specific categories. Second, and more severely, analyzing

the impact of mafia activity on the economy implies knowing how the economy would have been in the absence of mafia activity. However, a country or a region is either mafia-ridden or not (this is known as the fundamental problem of causal inference); therefore, it is very difficult to have a credible counterfactual. A convincing way to tackle this issue is to adopt the synthetic control method, a methodology that has been recently proposed to statistically examine comparative case studies. This methodology has originally been introduced to estimate the effect of the Basque conflict on GDP per capita (Abadie and Gardeazabal 2003). Later, it has also been applied to estimate the impact of the mafia in two Southern Italian regions – Apulia and Basilicata – which experienced a surge in mafia activity starting over the last few decades. Until the 1970s, these two Southern regions had witnessed little or no mafia activity in their territory. Starting from the 1970s, mafia activity and its connected violence rapidly increased in these two regions. A counterfactual is constructed using the information related to other regions where the mafia has been absent throughout the period considered. Mafia activity emerged in Apulia and Basilicata following three relevant episodes. First, the profitable activity of tobacco smuggling led to a ferocious conflict among different criminal groups. Second, the earthquake that struck the area between Campania, Basilicata, and Apulia in 1980 was followed by a flow of public funding for the reconstruction of the area. The increased availability of public funding led to an increase in mafia activity, attracted by the opportunities of grabbing a portion of these funds. Finally, the imprisonment policy for mafiosi was based on the *confino*, the policy according to which mafia-related criminals were dislocated in a different region from the one they originated from, in order to loosen their links with the local mafia. Apulia had the highest number of criminals according to the *confino* policy, which led to the spread of mafia activity in Apulia. The impact of mafia organizations on economic growth appears to be very significant. According to the synthetic control estimates, mafias are responsible for a 16% loss in GDP per capita over a 30-year period Pinotti (forthcoming).

Economic Consequences of Mafia: Misallocation of Public Funds

Such a huge impact captures the reduced-form causal effect from the mafia to GDP. Researchers have also devoted their efforts to highlight the underlying transmission mechanisms. One relevant channel consists of the misallocation of public funds (Barone and Narciso 2015). The case study regards the Italian Law 488/92, which has been the main policy used by the central government to promote growth in the southern Italian regions, by offering a subsidy to businesses investing in underdeveloped areas. Even though the law governing funding allocation had very detailed provisions to distribute subsidies, aimed at reducing the risk of fraud, many investigative reports state that the mafia managed to circumvent these criteria. A number of accounting and financial mechanisms have been used to divert funds, such as the creation of made-up firms, i.e., firms set up with the only scope of applying for public subsidies. Moreover, the mafia was able to corrupt public officials involved in the allocation of funds. By combining information on the spatial distribution of Law 488/92 funds with that of mafia activity, the study investigates whether mafia presence is able to influence public funds' allocation. The endogeneity of the link between the mafia and public funding must be addressed, in order to provide a causal interpretation to their relationship. Endogeneity may arise due to measurement error, omitted variables, and reverse causality. Instrumental variable identification strategy is a proper way to deal with the endogeneity issue. Focusing on Sicily, it is possible to construct a credible instrumental variable that is conceptually based on the historical origins of the Sicilian Mafia. As stated above, around the Italian unification in 1861, the demand for private protection arose in Sicily for two main motives. First, before the country's unification, a series of anti-feudal laws led to the opening up of the market for land, which contributed to the division of land property. Second, the vacuum of power following Italian unification and lack of protection of property rights amplified

landowners' need for protection against expropriation. The Sicilian Mafia arose as an industry offering private protection in this specific historical juncture. Indeed, in line with the literature on the origins of the mafia presented above, the mafia was more likely to appear in areas where the land was more valuable. Consequently, historical and geographical measures of land productivity can be used as instrumental variables for current mafia activity. In particular, the set of instruments includes rainfall shocks in the nineteenth century and geographical features (e.g., altitude and slope). This empirical strategy suggests that mafia presence has a positive effect on the likelihood of obtaining funding and the amount of public transfers: according to the existing estimates, mafia presence raises the probability of receiving funding by 64% and increases the amount of public funds to businesses by more than one standard deviation. The results are robust to different econometric specifications and to the use of alternative measures of the mafia. The mafia has a positive causal effect on public subsidies, but how should this result be interpreted? For example, is the positive relationship between mafia presence and public transfers due to a more generous public spending towards mafia-ridden areas? A falsification test suggests that the answer is negative: these areas display a lower level of expenditure on culture and education, in comparison to those where the mafia is absent. Were the State more generous towards disadvantaged areas, presumably it would have spent on other budget categories, such as education or culture. Moreover, there is empirical evidence on the mechanism through which the mafia can determine the allocation of public resources: the mafia is used to make connections with local entrepreneurship, and its presence raises the number of corruption episodes in the public administration sector. The impact of the mafia can also be disentangled from that stemming from a more general criminal environment. In the end, by diverting public subsidies assigned to poorer areas, the mafia hampers growth, investment, and, ultimately, development. From this perspective, this finding provides a relevant contribution

to the debate on the desirability and the design of public subsidies to firms.

Economic Consequences of Mafia: Other Channels

The impact of mafia organizations has been found to also affect the bank credit market. Using data on bank-firm relationships, it has been shown that crime has a negative effect on access to credit in Italy. Borrowers in high-crime areas pay higher interest rates and pledge more collateral. These results are found to be driven, in particular, by mafia-related crime, extortion, and fraud. By distorting loan conditions to firms, mafia organizations indirectly negatively affect investment and growth in the long run.

The effect of mafia on the economy can also work through the public sector. Organized crime usually operates as a pressure group that uses both bribes and the threat of punishment to influence policy. Consistently, on the empirical side, one can observe that criminal activity before elections is correlated with lower human capital of elected politicians and an increased probability that these politicians will be later involved in scandals. However, a strict legal institutional framework can contrast this bad influence. In the Italian case, when the central government imposed the dissolution of municipal government because of mafia infiltration, the quality of local politicians (proxied by their average education level) significantly improved Daniele Geys ([forthcoming](#)).

Finally, organized crime hinders fair market competition because mafia-related firms, which are usually run to launder money, can operate under the break-even point, thus forcing legal competitors out of the market; mafia also represents a deterrent for foreign investors, so depressing foreign direct investment.

References

- Abadie A, Gardeazabal J (2003) The economic costs of conflict: a case study of the Basque country. *Am Econ Rev* 93:113–132

- Allum F, Siebert R (2003) *Organized crime and the challenge to democracy*. Routledge, Abingdon
- Bailey J, Godson R (2000) *Organized crime and democratic governability: Mexico and the U.S.-Mexican borderlands*. University of Pittsburgh Press, Pittsburgh
- Bandiera O (2003) Land reform, the market for protection, and the origins of Sicilian mafia: theory and evidence. *J Law Econ Organ* 19:218–244
- Barone G, Narciso G (2015) Organized crime and business subsidies: Where does the money go?. *Journal of Urban Economics* 86:98–110
- Bonaccorsi di Patti E (2009) Weak institutions and credit availability: the impact of crime on bank loans. *Bank of Italy occasional papers*, no. 52
- Buonanno P, Pazzona M (2014) Migrating mafias. *Reg Sci Urban Econ Elsevier* 44(C):75–81
- Buonanno P, Durante R, Prarolo G, Vanin P (forthcoming) Poor institutions, rich mines: resource curse and the origins of the sicilian mafia. *Econ J*
- Cutrer A (1900) *La Mafia e i Mafiosi*. A Reber, Palermo
- Dal Bó E, Dal Bó P, Di Tella R (2006) Plata o pomo: bribe and punishment in a theory of political influence. *Am Polit Sci Rev* 100:41–53
- Daniele G, Geys B (forthcoming) Organized crime, institutions and political quality: empirical evidence from Italian municipalities. *Econ J*
- Dimico A, Isopi A, Olsson O (2012) Origins of the Sicilian mafia: the market for lemons. *Working papers in Economics of the University of Gothenburg*, no. 532
- Fiorentini G, Peltzman S (1996) *The economics of organised crime*. Cambridge University Press, Cambridge, UK
- Franchetti L (2011) *Condizioni politiche e amministrative della Sicilia*. Donzelli Editore, Roma
- Gambetta D (1993) *The Sicilian mafia: the business of private protection*. Harvard University Press, Cambridge, MA
- Gambetta D (2000) Trust: making and breaking cooperative relations, 49–72. Basil Blackwell, New York
- Maruko E (2003) Mediated democracy: Yakuza and Japanese political leadership. In: Felia A, Renate S (eds) *Organized crime and the challenge to democracy*. Routledge, London
- Pinotti P (forthcoming), *The economic consequences of organized crime: evidence from Southern Italy*. *Econ J*
- Skaperdas S (2001) The political economy of organized crime: providing protection when the state does not. *Econ Gov* 2:173–202
- Varese F (2006) How Mafias migrate: the case of the 'Ndrangheta in northern Italy. *Law Soc Rev* 40:411–444
- Varese F (2011) *Mafias on the move: how organized crime conquers new territories*. Princeton University Press, Princeton

Majority Control

- [Concentrated Ownership](#)

Manne, Henry

Enrico Colombatto

Università di Torino, Turin, Italy

IREF (Institut de Recherches Économiques et Fiscales), Lyon, France

Abstract

This entry illustrates the very prominent role Henry Manne played in the Law and Economics tradition. Manne made seminal contributions in two key areas: the dynamics of corporate control, and the ethics and efficiency of insider trading. He showed that the market for corporate control is an efficient way of protecting shareholders' interests and restraining abuse by the managers. From a normative standpoint, he emphasized that no specific regulation is required in these areas. The same normative conclusions also apply to insider trading, which is the quickest way of circulating information and avoiding bubbles. This entry concludes by summing up Manne's contributions in education.

Biography

Henry Manne (1928-2015) deserves a very prominent place among the founding fathers of Law and Economics, along with Guido Calabresi and Ronald Coase. In particular, since the late 1950s, Manne applied economic reasoning to investigate policy issues in two key areas that had previously been considered as the exclusive object of legal analysis: the dynamics of corporate control, and the ethics and efficiency of insider trading. In both cases, and despite considerable initial skepticism, Manne's contributions radically changed the way the economic and legal professions have regarded these topics. Furthermore, and typical of his vision, he used the law-and-economics approach to show that regulating the life of the corporation is unjustified and possibly harmful.

Manne (1962) and Manne (1965) are the path-breaking articles that opened new research

agendas in the economics of the modern corporation. In his 1962 contribution, Manne took on the traditional argument according to which small shareholders have little incentive to monitor the managers' behavior and actively partake in the life of companies. In particular, the traditional argument held that regulation is required in order to force the managers to disclose the information shareholders need, to enhance shareholders' participation, and to restrain the managers' potential abusive behavior. By contrast, Manne argued that the market for corporate control is effective in protecting the shareholders' interests, regardless of how much time and efforts they devote to monitoring the managers. If the managers misbehave, the value of the company declines, outsiders will be interested in buying the shares, gaining control and replacing the inefficient executives. In other words, in Manne's view, the market for mergers and takeovers ensures that share prices cannot drop for long because of managerial slack, as long as outside buyers are allowed to intervene and buy shares to obtain control. Indeed, the threat of a takeover is already enough to prevent the shareholders from being injured: if the incumbent managers fear the consequences of a hostile takeover, they are likely to react by improving their performance and meeting the incumbent shareholders' expectations. If so, performance improves and share prices recover. Once again, shareholders are protected, and share prices stay close to the company's true value. Therefore, Manne concludes, under both circumstances there is no need for regulation/legislation: Competition is indeed effective in keeping the managers under pressure and shielding small investors from abuse. By contrast, regulation frequently ends up defending the managers and justifying redistributive activities that have nothing to do with the purpose of the modern corporation.

Manne (1965) develops his earlier insights by moving from analyzing the relationships among shareholders and managers, to studying the economics of mergers. A takeover takes place when company A buys shares from company B's shareholders. It is an operation that prevents B from going bankrupt and that can be

completed in three different ways: proxy fights, direct purchase and mergers. As mentioned, Manne focuses on mergers, the buyers' preferred course of action, since they do not need to collect cash to finance the operation and they can skirt taxation. Not surprisingly, in these cases, corporate statutes play a crucial role in determining the outcome. For example, mergers usually require qualified-majority voting and can hardly succeed if B's incumbent management owns a relatively large portion of B's shares. Thus, one should not be surprised if company A and company B managers end up colluding. Under such circumstances, the merger would then take place with the consent of the management. This is still a desirable outcome: the top executives would be forced to exchange information, which promotes efficiency. As a result, according to Manne the frequent concerns raised by the antitrust authorities are misplaced: mergers are the result of a healthy, free-market, welfare-enhancing environment, rather than a threat to competition or to shareholders' interests.

Innovative and Original Aspects

Insider trading was the object of Manne (1966), a book that was actually his J.S.D. dissertation thesis at Yale and ignited a debate still lively today. In contrast with the views that dawned at the beginning of the twentieth century and became dominant since the mid-1930s, Manne made two key arguments. First, insider trading is a form of compensation for the managers and employees at large. It is cheaper than other forms of remuneration, and provides sets of incentives that contribute to bringing together the interests of the managers, the entrepreneurs and the owners. Second, insider trading is efficient, since it represents the quickest vehicle to circulate information, and reduces the risk of bubbles. In other words, insiders have accurate information about the company where they work, they ensure that this information is immediately available to the ordinary shareholders, and perform better than the regulators. Hence, legislation designed to curb insider trading is ineffective, harmful, and ethically

questionable, since it punishes an alleged crime in the absence of victims.

Impact and Legacy

Manne's work left a lasting legacy in two other areas: teaching and the economics of higher education. Manne was not only a great scholar, but also a determined and successful intellectual entrepreneur. As detailed in Manne (1993), while serving as a Professor at the University of Rochester, he started planning a new type of law-school curriculum, which emphasized the need for specialization, but at the same time required a cross-disciplinary approach to the selected area of specialization. While still in Rochester, in 1971 Manne launched an intensive summer course in economics designed to suit the needs of law professors. In 1974, Manne moved to the Miami Law School, where the program became the Law and Economics Center at the University of Miami, and offered economics courses to federal judges, too. Later, the Center moved to Emory and then in 1986 to George Mason University, where Manne was heading the Law school and eventually put in practice the innovative curriculum he had drafted in Rochester. These programs were an impressive success: they contributed to the professional life of generations of lawyers, law professors and judges, and made a real difference in the way the legal profession regarded economic issues, in the classroom and in court.

Manne (1973) is perhaps his best-known contribution to the understanding of the nature and dynamics of modern universities. Although focused on the American experience, much of Manne's arguments also apply to the academic environment prevailing in Europe. In particular, Manne drew attention to the consequences of an academic context characterized by extensive funding by governments and governmental agencies, foundations and companies, and in which the role played by the students' fees is modest. This setting has ensured that modern universities are less and less answerable to their clients (students). Put differently, Manne claimed that public funding has ensured that tenured professors are

all but unaccountable, and tend to pursue their own interests (mainstream research and consultancy), rather than meet the educational needs of their students, transfer knowledge, and help them develop critical abilities and creative thinking. Moreover, by making students and public opinion trust the virtues of educational establishments as heavily subsidized, not-for-profit organizations, faculties have succeeded in transforming modern universities into stable systems where intellectual entrepreneurship remains marginalized. According to Manne, in order to ensure stability, hiring procedures frequently reward scholarship accomplishment (mainstream research agendas, which do not necessarily imply scholarly value), and favor candidates who guarantee loyalty to the incumbents and feature nonmarket attitudes. The upshot is an ongoing decline in the quality of education, with very few exceptions, on both sides of the Atlantic.

Cross-References

- ▶ [Asymmetric Information in Litigation](#)
- ▶ [Economic Analysis of Law](#)
- ▶ [Efficient Market](#)
- ▶ [Governance](#)
- ▶ [Higher Education](#)
- ▶ [Law and Economics](#)
- ▶ [Law and Economics, History Of](#)
- ▶ [Merger Control](#)

References

- Manne H (1962) The 'higher criticism' of the modern corporation. *Columbia Law Review*, 3, March, pp 399–449
- Manne H (1965) Mergers and the market for capital control. *J Polit Econ* 73(April):110–120
- Manne H (1966) Insider trading and the stock market. Free Press, New York
- Manne H (1973) The political economy of modern universities. In: Burleigh AH (ed) *Education in a free society*. Liberty Fund, Indianapolis, pp 165–205
- Manne H (1993) *An intellectual history of the School of Law George Mason University*. the Law and Economics Center, Arlington

Further Reading

- Colombatto E, Cass R. (guest editors) (forthcoming) Symposium in honour of Henry Manne, *Eur J Law Econ*
- McChesney FS (ed) (2009) *The collected works of Henry G. Manne*, 3 volumes. Liberty Fund, Indianapolis

Market Definition

Javier Elizalde

Department of Economics, University of Navarra, Pamplona, Navarra, Spain

Abstract

The goal of this essay is to explain the role attributed to the analysis of market definition in the guidelines which rule in the United States and the European Union and to revise some of the empirical tests proposed by researchers in the fields of competition Economics and Law along with some of the critiques they received regarding their applicability for antitrust purposes.

Definition

Market definition is the analysis of determining the products which compete with each other and the geographic area where that competition takes place. The attention is paid, on one hand, to the concept of substitution among potential competitive products and, on the other, to the existence of joint market power by the firms which operate in the market especially for antitrust purposes.

Introduction

Market definition consists on the delineation of the market boundaries in the product and geographic dimensions from a competitive perspective. The interest for the researchers in the field of Law and Economics has led to the formulation of empirical tests of market definition, and much discussion has arisen mainly regarding their

compatibility with the market definition test which rules in the US antitrust legislation for merger control, known as the “hypothetical monopolist” (or SSNIP) test. The analysis of market definition has been performed under two alternative (not necessarily incompatible) purposes. First, in the tradition of Marshall (1920), the market comprises all the goods which are substitutes so they should exhibit identical prices with differences due to transportation costs. This approach is most plausible when the goal of the analysis is to find the geographic market for a homogeneous good. The second approach is to identify the products and the geographic area such that the firms who serve them may jointly enjoy substantial market power without relevant competitive constraints from other products or areas regardless of whether the other alternatives can be considered substitutes from the consumer perspective.

Market Definition in Merger Control

The market power approach has become predominant for researchers in the field of Law and Economics since the publication of the Horizontal Merger Guidelines in 1982 by the US Department of Justice (see Department of Justice and Federal Trade Commission (2010) for the latest revision). They introduced the hypothetical monopolist test of market definition. The rationale for this test is that the relevant market should include all those products and the geographic area such that, if they were all owned by a single firm (the hypothetical monopolist), the latter would enjoy some market power, being able to profitably exercise a “small but significant and non-transitory increase in price” (the initials SSNIP are used when referring for a price increase with that feature). This means that products other than those forming the relevant market do not impose a significant competitive constraint. The text mentioning the hypothetical monopolist test has been modified in the revisions of the horizontal merger guidelines of 1984, 1992, and 2010, but the role attributed to market definition as a preliminary and screening process in the investigations of competitive effects of mergers prior to the calculation of market shares

remains. Market definition is a previous step to the assessment of market power which is the main concern of antitrust authorities when investigating mergers.

In the European Union, there is no test which must be used for market definition. The concept of a relevant market was established in the 1997 “European Commission Notice on the definition of the relevant market for the purposes of Community competition law” and relies on the concept of consumer substitution “[...] by reason of the products’ characteristics, their prices and their intended use.” The definition of a relevant market ruling in the European Union includes both demand-side and supply-side substitutes when substitution takes place “quickly and easily.” The Notice mentions a version of the hypothetical monopolist test as a suggested method and enumerates a series of econometric tests, which are actually detailed in the following paragraphs of this essay.

Empirical Tests of Market Definition

Regarding the empirical literature on market definition, the earliest works were mainly intended to identify the geographic market for a homogeneous good. An approach which was dominant for geographic market definition before the attention was led to prices was the Elzinga-Hogarty test (after Elzinga and Hogarty 1972, 1973) which was a test of shipment data. It used aggregate inflows and outflows of consumers to determine market boundaries. Geographic market boundaries were expanded until both flows were below a cutoff level. This test was used for analysis of hospital mergers in the United States in the 1980s and 1990s, and it received many critiques especially for the limitation of the test to account for the heterogeneity of patients to travel for medical care.

Both in the fashion of the Marshallian concept of a market of equal prices and on the emphasis on the ability to increase prices stated in the US guidelines, the tests of market definition became primarily tests of price data or both prices and quantities when data were available.

There is a vast stream of literature on market definition using tests of time series of prices. Stigler and Sherwin (1985) proposed a test based on price *correlations*. Under the prior that the differences in prices of products of the same market are due to transport costs, their price movements must follow the same pattern so the time series of those prices must be correlated.

Horowitz (1981) suggested that the Marshallian prediction of equality of the prices only occurs in equilibrium. When shocks happen, there are adjustment lags before returning to equilibrium. Horowitz proposed a test called the *speed of adjustment* test as it is focused on estimating the speed at which the short-term difference in prices between two areas returns to the long-term difference. A persistent divergence between the short-term difference and the long-term difference would mean that the two products or areas are in different markets.

Another test called the *causality* test includes products in the same geographic market if the long-run difference between price series is independent of exogenous influences not related to costs. The test, proposed by Uri and Rifkin (1985) and Uri et al. (1985) and based on the concept of Granger causality, defends that the price in one area of a geographic market could be predicted with information on prices in the other area.

Slade (1986) proposed an *exogeneity* test according to which, if a disturbance in one area spills into another area, the exogeneity of price formation is rejected and both areas are in the same market.

A *stationarity* test of market definition was performed by Forni (2004) analyzing the long-run price ratio between two areas rather than a series of short-run price differences which were used in previous tests. Using this test, if the long-run proportional relationship between the prices in two areas is not stationary, then the areas can be said to be in different markets.

On the grounds of market definition related to substitution between products for the consumer, much work has relied on elasticities of demand when data on both prices and quantities has been available. The own-price elasticity of demand provides information on the existence of

substitutes for a good, and the cross-price elasticity of demand between two goods provides information on whether those two goods are substitutes, which would be the case if it takes a positive value. A noteworthy approach is the analysis of critical elasticity and critical loss, following the works by Johnson (1989) and Harris and Simons (1989). The main focus of this analysis is the own-price elasticity of demand of the hypothetical monopolist. The critical elasticity is the maximum elasticity of demand a hypothetical profit-maximizing monopolist could face at pre-merger prices to be able to profitably increase its price by a SSNIP. If the elasticity faced at pre-merger prices is higher than the critical elasticity, the hypothetical monopolist would not raise profits with that increase in price so this would lead to market aggregation. The critical loss is the maximum reduction in output a hypothetical monopolist can tolerate in order for the increase in price to be profitable. If the actual loss is higher than the critical loss, the price increase is not profitable, so the relevant market should be expanded.

An alternative approach is based on the own-price elasticity of the residual demand of the hypothetical monopolist. Baker and Bresnahan (1988) used this methodology to estimate the degree of market power in an industry with differentiated products (the beer industry in the United States). Scheffman and Spiller (1987) applied it to the geographic market definition for a homogeneous good (unleaded gasoline in the Eastern United States). Once the value of the own-price elasticity of the hypothetical monopolist's residual demand is estimated, the effect of the price increase on the monopolist's profit is computed (see also Kamerschen (1994), Ekelund et al. (1999), and Cardona et al. (2009) for other works under this approach).

The recent emergence of detailed retail data on prices, quantities, and other variables (often through scanner systems) has favored the estimation of each firm's individual demand, and, through the estimation of both own- and cross-price elasticities of demand, the likely effects of mergers can be simulated. Some works use this methodology for the definition of the relevant market for cars (Brenkers and Verboven 2006,

analyzing competition at both the manufacture and retail levels), computer servers (Ivaldi and Lörincz 2011), and movie theaters (Elizalde 2013, which analyzes both demand-side and supply-side substitution).

There is a vast literature, such as Werden (1990, 1998, 2003), Werden and Froeb (1993), O'Brien and Wickelgren (2003), and Hosken and Taylor (2004) among many others, which criticizes most of the tests enumerated above by showing their incompatibility with the hypothetical monopolist test of the US guidelines and their invalidity for predicting the likely effects of mergers, as some of the tests are intended to the identification of substitutes or are based on past evidence with limited capability to predict future events among other reasons.

Coate and Fischer (2008) provide a review of the tests employed by the enforcement agencies in the United States along with comments about the works which proposed them and the critiques they received.

Controversy Regarding Market Definition

Market definition itself, especially with the role attributed in the US antitrust legislation, is strongly criticized in some grounds of Law and Economics which are very much skeptical about the competition authorities' interest on market concentration and market power, as the immediate usefulness of market definition is the calculation of market shares. Coinciding with the 2010 revision of the US Horizontal Merger Guidelines, an alternative to market definition, based on the upward pricing pressure (UPP) resulting in a merger, was proposed by Farrell and Shapiro (2010) and critically discussed by Carlton (2010), whereas Kaplow (2010, 2011) defends the elimination of market definition as conclusions may be misleading.

Cross-References

- Demand
- Economic Analysis of Law

- Empirical Analysis
- European Community Law
- Law and Economics
- Merger Control

References

- Baker JB, Bresnahan TF (1988) Estimating the residual demand curve facing a single firm. *Int J Ind Organ* 6(3): 283–300
- Brenkers, R. and F. Verboven (2006): Market definition with differentiated products – lessons from the car market. In: Choi J.P. (ed) *Recent developments in antitrust: theory and evidence*. MIT Press
- Cardona M, Schwarz A, Burcin Yurtoglu B, Zulehner C (2009) Demand estimation and market definition for broadband internet services. *J Regul Econ* 35:70–95
- Carlton DW (2010) Revising the horizontal merger guidelines. *J Compet Law Econ* 6(3):619–652
- Coate MB, Fischer JH (2008) A practical guide to the hypothetical monopolist test for market definition. *J Compet Law Econ* 4(4):1031–1063
- Department of Justice and Federal Trade Commission (2010). Horizontal merger guidelines. Available at <https://www.ftc.gov/sites/default/files/attachments/merger-review/100819hmg.pdf>
- Ekelund R, Ford G, Jackson J (1999) Is radio advertising a distinct local market? An empirical analysis. *Rev Ind Organ* 14(3):239–256
- Elizalde J (2013) Market definition with differentiated products. A spatial competition application. *Eur J Law Econ* 36(3):471–521
- Elzinga K, Hogarty T (1972) The demand for beer. *Rev Econ Stat* 54(2):195–198
- Elzinga K, Hogarty T (1973) The problem of geographic market delineation in Antimerger suits. *Antitrust Bull* 18:45–81
- Farrell, J. and C. Shapiro (2010): Antitrust evaluation of horizontal mergers: an economic alternative to market definition. *J Theor Econ*, 10(1), Article 9
- Forni M (2004) Using stationarity tests in antitrust market definition. *Am Law Econ Rev* 6(2):441–463
- Harris BC, Simons JJ (1989) Focusing market definition: how much substitution is enough. *Res Law Econ* 12: 207–226
- Horowitz I (1981) Market definition in antitrust analysis: A regression-based approach. *South Econ J* 48(1):1–16
- Hosken D, Taylor CT (2004) Discussion of 'using stationarity tests in antitrust market definition'. *Am Law Econ Rev* 6(2):465–475
- Ivaldi M, Lörincz S (2011) Implementing relevant market tests in antitrust policy: application to computer servers. *Rev Law Econ* 7(1):29–71
- Johnson FI (1989) Market definition under the merger guidelines: critical elasticities. *Res Law Econ* 12: 235–246

- Kamerschen DR (1994) Testing for antitrust market definition under the Federal Government guidelines. *J Leg Econ* 4(1):1–10
- Kaplow L (2010) Why (Ever) define markets? *Harv Law Rev* 124:437–517
- Kaplow L (2011) Market definition and the merger guidelines. *Rev Ind Organ* 39:107–125
- Marshall A (1920) *Principles of economics*, book 5. Macmillan, London
- O'Brien DP, Wickelgren AL (2003) A critical analysis of critical loss analysis. *Antitrust Law J* 71(1):161–184
- Scheffman DT, Spiller PT (1987) Geographic market definition under the U.S. Department of Justice merger guidelines. *J Law Econ* 30:123–147
- Slade ME (1986) Exogeneity tests of market boundaries applied to petroleum products. *J Ind Econ* 34(3):291–302
- Stigler GJ, Sherwin RA (1985) The geographic extent of the market. *J Law Econ* 28(3):555–586
- Uri ND, Rifkin EJ (1985) Geographic markets, causality and railroad deregulation. *Rev Econ Stat* 67(3):422–428
- Uri ND, Howell J, Rifkin EJ (1985) On defining geographic markets. *Appl Econ* 17(6):959–977
- Werden GJ (1990) The limited relevance of patient migration data in market delineation for hospital merger cases. *J Health Econ* 8(4):363–376
- Werden GJ (1998) Demand elasticities in antitrust analysis. *Antitrust Law J* 66(2):363–414
- Werden GJ (2003) The 1982 merger guidelines and the ascent of the hypothetical monopolist test. *Antitrust Law J* 71(1):253–275
- Werden GJ, Froeb LM (1993) Correlation, causality and all that jazz: the inherent shortcomings of price tests for antitrust market delineation. *Rev Ind Organ* 8(3):329–353

Market Failure: Analysis

José Luis Gómez-Barroso
Dpto. Economía Aplicada e Historia Económica,
UNED (Universidad Nacional de Educación a
Distancia), Madrid, Spain

Abstract

Given that there is no agreement on the procedure by which economic efficiency should be measured, a closed catalogue of market failures cannot be talked about. There is however a reasonable agreement in economic literature on the identification of up to a total of five reasons for the existence of market failures:

public goods, externalities, imperfect competition, information failures, and incomplete markets. There are three other situations that some authors also include in the list of market failures: merit goods, an unbalanced macroeconomic situation, and economic situations that assault criteria of equity.

Definition

Market failure is any situation in which the autonomous action of the market does not lead to an economically efficient outcome.

Introduction

Drawing the border between public and private is and has been a constant concern throughout the history of human thought. The economy is one of the fields in which such a distinction is vital. In this discipline, whatever space there is for what is public is derived from the answer to a basic question that every economist has faced: does the joint action of individual agents in pursuit of self-interest (utility or “happiness” in the case of individuals; benefit in the case of entrepreneurs) cause an acceptable result from the point of view of the common interest? When the answer is no, public intervention to correct such situations could be admissible. The unacceptability of the result of private activity derives from the use of equity or efficiency criteria. In this second case, that is, when the market does not lead to an efficient outcome of its own accord, it is said that we are in the presence of market failure.

Given that there is no agreement on the procedure by which economic efficiency should be measured (not even on the actual definition of the concept of efficiency), a closed catalogue of market failures cannot be talked about. Consider that some school of economic thought even denies their existence. Nor is there consensus on what should be done (or even whether anything should be done) to correct market failures.

By dint of being an entry in an encyclopedia proceeds to adopt a nonrestrictive approach, and

the following section describes all the circumstances in which the existence of a market failure has been reasonably argued. There is insistence on the fact that many currents of economic thought would reduce the list and would furthermore qualify the extension that should be given to each type of market failure. In this respect, the definitions given are the most repeated and accepted, though for each market failure there is an extensive bibliography that would enable each of the concepts to be specified, formalized, or criticized.

Types of Market Failures

There is a reasonable agreement in economic literature on the identification of up to a total of five reasons for the existence of market failures. They are not mutually exclusive or even independent. For the purposes of classification, they can be regrouped into two blocks:

- Those linked to the characteristics of the activity or of the good in itself: public goods and externalities
- Those related to the market situation: imperfect competition, from which information failures can become independent, and incomplete markets

There are three other situations that some authors also include in the list of market failures, though this is not the norm in literature:

- Merit goods
- An unbalanced macroeconomic situation (existence of unemployment, waste of resources)
- Economic situations that assault criteria of equity

Intrinsic Characteristics of the Good or Activity

Public Goods

There are two characteristics that define a public good: it is non-excludable and consumption is non-rivalrous. In other words, no one can be

deprived of enjoying the good and its consumption by an individual does not exhaust it or even affect the utility that others can extract from its consumption. The most typical examples that are mentioned in literature are national defense or coastal lighthouses.

“Pure” public goods that rigorously meet these two restrictions are few and far between. A large part of those considered public goods possess these attributes under certain conditions and could even be treated as private in different contexts. Moreover, the situations in which it is possible to talk about public goods are sometimes circumstantial. In a free wireless Internet area, consumption is non-rivalrous, as long as there are no agglomerations, but if the number of connections is excessive, “consumption” of the good by other people reduces utility and there is competition for the resources. The term impure public goods is usually used when there is a congestion problem.

With some goods, even though their underlying structure functions *naturally* as a public good, uses that break the public space in private spheres, where access is conditioned, can be developed. If a beach becomes private, a toll is set up on a motorway, or a television transmission is codified, it is evident that exclusion is possible. These situations are described as club goods.

In the case of public goods, the non-efficient outcome originates in the market not being interested in offering these goods, since in a situation of impossible exclusion income would depend on the will to contribute to their financing by those who enjoy them: how can the “right to see” a fireworks display be charged? And if exclusion can be and is actually chosen, non-efficiencies would also be generated, considering that the marginal cost of the enjoyment experienced by an additional person is zero (non-rivalrous consumption). The second reason why a non-efficient situation can be created is the overexploitation of goods that would otherwise be public. Excessive use does not only end non-rivalry in consumption, but can lead, in the extreme, to the disappearance or exhaustion of resources that are common property, such as fishing grounds or aquifers.

Nowadays, the concept of public goods has been extended from a local scale to a global scale with the introduction of the concept of global public goods. World peace, financial stability, and the eradication of epidemics would be global public goods because, once achieved, their benefits, which no one could be left out of, would be geographically unlimited. What is true is that some of these concepts, such as peace, are more desirable political objectives than goods that could be supplied by the market and, therefore, it is not strictly accurate to talk about market failures.

Externalities

There is an externality when a concrete activity influences other activities or individuals that do not directly participate in the first activity. Externalities can be positive, if this spillover is beneficial, or negative, if they cause harm. The most typical examples collected in literature are, respectively, fruit trees pollinated by bees from a nearby apiarist and the contamination of a river.

In the presence of externalities, a social cost should be added to the internal cost reflected in a company's bookkeeping. The opening of several drinking bars in a specific neighborhood may favor certain businesses in the area (car parks, takeaway restaurants), but may have a negative effect on local residents being able to sleep. Not considering this social cost (both positive and negative) will lead to the production of an amount of the good that is greater or less than what is socially desired, resulting in a non-efficient allocation of resources. This could be resolved if the parties involved negotiated (and, therefore, the externalities "become internalized"), but even if it were possible to locate all the potential beneficiaries or injured parties, it would be difficult to reach agreements, and, moreover, were agreements to be reached, the transaction costs could exceed the benefits generated by the elimination of undesired external effects.

Nowadays, so-called network externalities associated with the growth in the number of users who use a service are gaining importance. There are direct and indirect network externalities. The first arise from the fact that each new

subscriber benefits from access to the group of preexisting users, but at the same time assumes a new possibility for communication (real or potential) for this customer base already connected. The second arise from an increase in the quality or quantity of available services, a catalogue that grows with the number of users.

Market Situation

Failure of Competition

The prevailing opinion in economics is that perfect competition is the market structure that leads to efficient outcomes. The problem is that markets with perfect competition are a fiction, since the conditions required to receive this description are impossible to achieve in practice. There are several types of obstacles to perfect competition: differentiated (not uniform) products, lack of information, producers of an influential size that use their power to hinder the actions of their rivals, need for prior investment or other entry barriers. Therefore, in practice almost all markets have imperfect competition. Or the other way round, it could be interpreted that a market failure would be the outcome of almost all markets.

As the catalogue of possibilities is extremely extensive and covers any situation between perfect competition and a monopoly, the problem lies in deciding when the failure of competition is considered sufficiently significant to assume that it generates non-efficiencies. Furthermore, markets are dynamic and situations change. It is in this respect important to determine what constitutes a "reasonable" waiting period before assessing whether or not obstacles to developing "sufficient" competition are disappearing.

Natural monopolies are included in this category of competition failures, even though on some occasions they are mentioned as independent market failures. It is important to stress that it is not market structure per se but the result to which this structure can lead that generates the market failure: if only one bicycle shop existed in a town with a population of 5,000 (in which "there is no space" for two shops), this would not be a problem; the problem would be that its owner would

take advantage of his condition as the only supplier to set abusive prices or conditions.

Incomplete Markets

In this case, the problem is not that market structure is far from perfect competition. It is simply that no one provides the service to certain users. Properly speaking (in terms of efficiency), it is only possible to talk of incomplete markets when there is a demand not met by producers, and the cost of satisfying this demand is lower than what those seeking the products would be willing to pay.

Information Failures

Strictly speaking, information failures are another of the causes that contribute to imperfect competition. The point of considering that it is an autonomous market failure comes from its special importance and from the fact that it also affects the demand side, unlike other “imperfections” linked to the offer.

If the information is incomplete or scant, the producer could not use the most suitable factors or reach all potential customers; in turn, the consumer could not choose the product or supplier that most suits him. One particular case is that of the “experience goods,” in which it is necessary to have had a prior experience before fully appreciating their value: the consumer’s lack of knowledge can reduce potential demand.

When information is asymmetric, the parties have different knowledge of a specific fact and the best informed party may use this advantage for his benefit. Asymmetries can lead to different situations of non-efficiency. Some of them have been formalized as specific categories. Adverse selection occurs when the majority of a market is made up by goods/customers/producers that the other party would not choose having all the information (insurance is taken out by those that are more likely to need it, something the company does not know; in a used items market, most present hidden defects). A moral hazard problem may exist when someone bears the potential negative consequences of a certain action performed by someone else, an action that is not observable for the first (and therefore, for instance, individuals

can assume greater risks in their decisions or even be tempted to cheat).

Other Possible Public Goods

Merit Goods

Merit goods are those whose consumption the State judges to be “positive” and, therefore, considers it appropriate to encourage it. Examples are education or using a seat belt in cars. Their opposites are demerit goods, such as the consumption of drugs. In merit (or demerit) goods, therefore, public opinion differs from the private assessment. There is an interest attributable to the community as a whole that does not result from the “mere” addition of individual interests. Whoever judge efficiency by assessing common or social interest in this way do then consider that we are facing a market failure.

It is not often, however, that merit goods appear in this category. For many authors, the foundation of the argument presented in the above paragraph has nothing to do with the ability or inability of the market to supply these goods efficiently. Others, even recognizing that we are facing cases of “consumer myopia” (in which consumers would not be able to assess self-interest; merit goods would become exceptions to the premise that it is the consumers themselves who are better placed to maximize their welfare according to their current income), place them in a different category of possible justification of public intervention. Finally, others consider that we are facing a case of mere presence of externalities.

Macroeconomic Situation

The fact that macroeconomic indicators are not performing well seems to indicate that the market (understood as an abstract entity formed by the union of all specific markets of goods and services) is not operating correctly, that is, it is not achieving an efficient outcome. Specifically, in the presence of a high unemployment rate, it would be logical to think that the economy could achieve a better (more efficient) outcome if part of the now wasted resources was used.

In order that this failure *of the* market be caused, failures *in some* markets or also in the

structures framing the development of economic activity (and which, therefore, affect all markets) should be produced. This feature is the reason why there is a view (not widely shared) that advocates for the existence of a market failure.

Equity

Income distribution criteria are usually considered an additional cause among the reasons that could justify State activity. Some authors, however, believe that in situations in which wealth distribution is very unequal, it is also necessary to talk about market failure. Their argument is that people with scant purchasing power can barely “communicate” with the market to make it aware of their needs, since only those who can pay the prices of the goods and services offered by companies manage this. Other authors consider that achieving an equitable society (or at least the reduction of poverty) would enter into an extensive definition of public good.

Cross-References

- [Anticommons, Tragedy of the](#)
- [Economic Efficiency](#)
- [Efficient Market](#)
- [Externalities](#)
- [Government Failure](#)
- [Market Failure: History](#)
- [Public Choice: The Virginia School](#)
- [Public Goods](#)

Market Failure: History

José Luis Gómez-Barroso
Dpto. Economía Aplicada e Historia Económica,
UNED (Universidad Nacional de Educación a
Distancia), Madrid, Spain

Abstract

The existence of market failures is linked to any opinion on the role reserved for the State in

the economy. In practice, this means that the notion of market failure (though not necessarily the term) can be traced back throughout all contributions made to economic science. This entry reviews the historical evolution of the opinion on market efficiency and, consequently, of the opinion on the existence of market failures.

Definition

Market failure is any situation in which the autonomous action of the market does not lead to an economically efficient outcome.

Introduction

The goal of this entry is to obtain some knowledge, albeit meager, of the different conceptual positions with regard to market failures. To that end, it reviews the historical evolution of the opinion on market efficiency and, consequently, of the opinion on the existence of market failures. It is obvious to say that, as in many branches of knowledge and specifically in economics, those that are considered landmarks are based to a large extent on previous developments and the mere description of the most significant contributions necessarily leaves other valuable works to one side.

Historical Evolution of Economic Thought on Market Failures

Economic literature usually awards Bator (1958) the merit of having coined (or of at least having used it for the first time in a publication) the term “market failure,” defined as *the failure of a more or less idealized system of price-market institutions to sustain “desirable” activities or to estop [sic] “undesirable” activities*.

However, the existence of market failures appears to be explicitly or implicitly linked to any opinion on the role reserved for the State in the economy. This means, in practice, that one

way or another, the notion of market failure (though not necessarily the term) can be traced back throughout all contributions made to economic science. And although the work of Adam Smith, and his “invisible hand,” would seem to mark, in the opinion of many, a starting point in the construction of a system capable of suitably assessing the question of efficiency in markets, what is true is that reasonings in one or another sense, with varying degrees of rigor, can be found in any chapter of the history of economic thought.

It can be generically stated that all schools prior to Smith had doubts about whether private activity (the markets) managed to reconcile individual and social welfare. The ancient Greek thinkers, who did not conceive of the economy as an autonomous discipline, saw in its fellow citizens’ pursuit of wealth a danger for the harmony of social order. Consequently, both Aristotle and Plato invoked strict control of economic activity by the State. In *Politics*, Aristotle proposed a superintendency whose main functions were to see that everyone involved in transactions was honest and holds to their agreements and contracts and that orderliness was maintained; some authors have even suggested that Aristotle’s idea was that these supervisors could set prices (Mayhew 1993).

With all the considerations derived from some very different context and motivation, these ideas were taken up again in the Middle Ages by scholastics to reach some similar conclusions: the sinning nature of man means that, against the will of God, his love for himself comes before his love for others and, therefore, the result of unregulated individual activity (ergo the market) does not agree with divine dictates. The intervention of the authority to ensure the harmony of socioeconomic order is, once more, the corollary of this reasoning.

For mercantilists, in the sixteenth and seventeenth centuries, it was not the precepts of justice, be it divine or not, but national interest that was discredited when self-interest was pursued. And given that the first representation of this national interest is the accumulation of precious metals, commercial activity is the sector where control of private initiative should be directed at in the first place.

As a reaction to mercantilism, in the eighteenth century, the physiocrats postulated the abolition of obstacles to trade. Even though their thinking often appears to be associated with the phrase *laissez-faire, laissez-passer*, the pursuit of self-interest, derived from an excessive demand for manufactured and luxury goods, continued to be part of the problem that had to be resolved (Medema 2009). In fact, François Quesnay, one of the greatest representatives of this school, advocated control of the markets, especially of agrarian markets, given their special transcendence due to agriculture being the source of the *produit net*.

What distinguishes Adam Smith from his predecessors is the assessment of the result that the sum of individual actions produces: even if individuals do not consider in their actions anything similar to the common good, but rather the strictly personal, the most beneficial outcome for society is derived from the sum of all these actions. It is important to stress the idea of “sum,” since Smith in *Wealth of Nations* gives some 60 examples in which the pursuit of self-interest causes harmful consequences for social good (Kennedy 2009). Moreover, the idea of an invisible hand infallibly guiding the markets is more a modern interpretation of his thinking than the foundation of his work, in which he only uses the expression incidentally and in the religious and cultural context of his time (among many others Harrison 2011; Kennedy 2009). Whatever interpretation is given to the metaphor of the invisible hand, what Smith is clear about is that public interference could not improve the result of private activity. Again, here it is possible to make a (important) qualification, since the three functions assigned to the sovereign in the system of “natural freedom” described in Book IV of *Wealth of Nations* are defense, justice, and “the duty of erecting and maintaining certain public works and certain public institutions, which it can never be for the interest of any individual, or small number of individuals, to erect and maintain,” an exception that has even led to calling him “cautious interventionist” (Reisman 1998).

In spite of all the detailed statements that can be made on what Smith wrote, it is at least

unambiguous that the market should be much less guided by governments or religions in his system of natural freedom than in all contributions that had been made up to then. His vision was shared, and refined, by the classical economists of the nineteenth century. Without being too ardent in their defense of *laissez-faire*, as they are often represented, for them the pursuit of self-interest, duly channeled through the activity of the markets, produces results that, on most occasions, are better than any other that could be obtained by government policy. The assessment of those results and, therefore, the conclusion they reach incorporate ideas developed by Jeremy Bentham and his utilitarianism ethic.

By the middle of the century, some nuances on firmly held concepts began to be introduced. The first to do so was John Stuart Mill, who in his “harm principle” puts as a limit to individual activity the cases in which said activity negatively affected the interests of others, which in modern terminology would be called the presence of negative externalities. As with so many theorists, his thinking is more complex than what is often presented: the circumstances in which public intervention could be admissible include situations in which individuals are not able to judge the result of their own actions (e.g., with regard to education). The above notwithstanding, Mill shares his reservations regarding public intervention with classical economists (there is a clear rule for not interfering, but none for interfering), though these reservations come more from his misgivings regarding the competence of governors than from some solid theoretical principles. Henry Sidgwick extended even further the catalogue of situations in which the principle of *laissez-faire* did not maximize common welfare. Examples are the overexploitation of natural resources, the occasions on which companies do not offer sufficient quantities of goods or services because they cannot recoup their investment or the cases in which there is not enough information on the effects of a certain product or action. For Sidgwick, a direct need for public intervention did not, however, come about in these situations. His answer to the question follows the rules of utilitarianism to their extreme and is, therefore,

much more pragmatic than that of Mill: the cost of potential intervention (that includes aspects that already concerned many of his predecessors, such as corruption and the possibility that certain groups are intentionally favored) should also be valued and then confronted with the potential benefit obtained.

At the dawn of a new century, economists from the Cambridge School applied the tools of marginalism to the problem of market limits. Alfred Marshall, by means of the consumer surplus calculation, identified situations in which public activity could increase welfare. Arthur Cecil Pigou, comparing social and private net marginal product, gave much more concrete form to what we today call the theory of externalities. His book *The Economics of Welfare* (Pigou 1920) became an obligatory support or criticism in any subsequent contribution to the debate. The role reserved for the State was, according to Caldari and Masini (2011) and despite what is usually argued, more important for Marshall (for whom the market should be substituted in questions of social relevance) than for Pigou (for whom it should merely be complemented). All this in theoretical terms, since in practice (normative economics against positive economics) both authors, especially Marshall, did not openly commit to intervention given the limitations and inefficiencies that both pointed out in the political processes (Backhouse and Medema 2012).

In the following decades, these misgivings disappeared to the extent that economists who extended the work of Pigou not only kept enhancing and shaping the theory of externalities but made progress in the mathematical demonstration of the benefit generated by public intervention (therefore necessary) in these situations (Meade 1952; Scitovsky 1954; Buchanan and Stubblebine 1962). Also, by the middle of the century, further progress in the categorization of public goods was made; the work of Samuelson (1954), where the characteristics of “collective consumption” goods are described, and that of Buchanan (1965), on impure public goods, merit to be highlighted.

But at the same time as this current was developing, other economists were challenging their

conclusions: the critical work of Coase marks an inflection point accompanied in time by the theory of Government failures.

Though sketched in other previous works, it was in *The Problem of Social Cost* (Coase 1960) that Ronald Coase structured his arguments. For him, when an activity is restricted due to the presence of negative externalities, there is also a cost associated with the restriction of such activity that is not taken into account in the calculations of social and private welfare. Which of the two “damages” is permitted is a question of assigning property rights. That assignation, though, does not necessarily lead to an efficient outcome. Ideally, it is not public intervention, but rather negotiation (the market) that could lead to an optimal (efficient) situation if the rights were well defined and there were no transaction costs. When this is not fulfilled, other options are possible. Regulation is one of them. However, if the costs associated with regulation exceed those it aims to prevent, it would leave not doing anything as the only solution; thus, the importance of having an appropriate institutional structure.

It was also in the second half of the century that a theory that did not strictly refer to market failures, but which should compulsorily be mentioned, was formulated: it is the parallel theory of Government failures. The precedents of the school of public or collective choice, as they are known, can be traced in the work of the Italian school of *scienza delle finanze* and of Knut Wicksell, who had incorporated public decision processes into theoretical analyses, thereby turning them into a factor that also determines what should be the nature and extension of the functions of the State (see Medema 2009), and further on in the work of Kenneth Arrow (most especially in Arrow 1951). With this foundation, some economists from the universities of Virginia and Chicago in the early 1960s developed a theory whose maxim was that individuals who choose between alternatives are guided by an actual rationality, both when they do so in the public function as when they choose it for themselves: they always tend to maximize self-interest.

Further on in the second half of the twentieth century, some developments came about in the

theory of market failures, very linked overall to the exploration of the consequences of the asymmetries of information, which includes concepts such as adverse selection (the famous market for “lemons” described in Akerlof 1970), the moral hazard, or the lock-in effects. There has also been abundant literature theoretically or empirically resisting these advances: a dozen studies are compiled in Cowen and Crampton (2002).

At this point, it is apparent that some important schools have been left out of this historical summary. Though they have not directly referred to market failures or even to self-interest, their concept of the role of the State enables us to infer what their position is in this regard. Here we consider three that are basic for understanding modern economic thought: Marxism, the Austrian School, and Keynesianism.

For Marxism, the market always generates undesired results and, therefore, does not consider perfect markets (in which there is also capitalist exploitation and economic crises) as a desirable or reasonable end. Therefore, market failures are an irrelevant argument or, from another perspective, all markets are pure failure.

The spontaneous order advocated by the Austrian School generates a more efficient allocation of society’s resources than that which any design can achieve. There are, therefore, no market failures, or rather it would be impossible to know whether the market is failing. To do so it would be necessary to carry out an impossible assessment, since they deny the neoclassical concept of efficiency that is substituted by the non-hindrance of the actions of individuals. In this respect, any market failures would come from public action.

Finally, Keynesianism places the emphasis on the stickiness of prices and (especially) of wages in the short run, a circumstance that makes markets with no intervention generally not able to generate efficient outcomes.

Cross-References

- [Coase and Property Rights](#)
- [Market Failure: Analysis](#)

- [Mercantilism](#)
- [Smith, Adam](#)

References

- Akerlof GA (1970) The market for lemons: quality uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Arrow KJ (1951) Social choice and individual values. Wiley, New York
- Backhouse RE, Medema SG (2012) Economists and the analysis of government failure: fallacies in the Chicago and Virginia interpretations of Cambridge welfare economics. *Camb J Econ* 36(4):981–994
- Bator FM (1958) The anatomy of market failure. *Q J Econ* 72(3):351–379
- Buchanan JM (1965) An economic theory of clubs. *Economica* 32(125):1–14
- Buchanan JM, Stubblebine WC (1962) Externality. *Economica* 29(116):371–384
- Caldari K, Masini F (2011) Pigouvian versus Marshallian tax: market failure, public intervention and the problem of externalities. *Eur J Hist Econ Thought* 18(5):715–732
- Coase RH (1960) The problem of social cost. *J Law Econ* 3(October):1–44
- Cowen T, Crampton E (eds) (2002) Market failure or success – the new debate. Edward Elgar, Cheltenham/Norhampton
- Harrison P (2011) Adam Smith and the history of the invisible hand. *J Hist Ideas* 72(1):29–49
- Kennedy G (2009) Adam Smith and the invisible hand: from metaphor to myth. *Econ J Watch* 6(2):239–263
- Mayhew R (1993) Aristotle on property. *Rev Metaphys* 46(4):803–831
- Meade JE (1952) External economies and diseconomies in a competitive situation. *Econ J* 62(245):54–67
- Medema SG (2009) The hesitant hand. Taming self-interest in the history of economic ideas. Princeton University Press, Princeton/Oxford
- Pigou AC (1920) The economics of welfare. Macmillan and Co., London
- Reisman DA (1998) Adam Smith on market and state. *J Inst Theor Econ* 154(2):357–383
- Samuelson PA (1954) The pure theory of public expenditure. *Rev Econ Stat* 36(4):387–389
- Scitovsky T (1954) Two concepts of external economies. *J Polit Econ* 62(2):143–151

Marking

- [Labeling](#)

Mass Media

- [Media](#)

Mass Tort Litigation

- [Class Action and Aggregate Litigations](#)

Mass Tort Litigation: Asbestos

Michelle J. White
University of California, San Diego, La Jolla,
California, USA

Abstract

Litigation over harm due to asbestos exposure is the largest mass tort in U.S. history. This article explores why the asbestos mass tort grew so large and argues that a large set of factors, rather than a single cause, are needed to explain the size of the asbestos mass tort. Among these factors are the seriousness of asbestos diseases and the fact that harm due to asbestos exposure was concealed by producers of asbestos products, changes in the law and legal procedures that favored plaintiffs, the large number of both potential plaintiffs and potential defendants, and plaintiffs' lawyers high profit from finding and representing asbestos claimants. I also argue that because of the unique circumstances explaining asbestos litigation, future mass torts are unlikely to be as large. The article also discusses research on asbestos litigation and explores various approaches—successful and unsuccessful—that have been proposed to reduce the number of asbestos claims.

Definition

Mass Tort: A mass tort involves numerous plaintiffs filing civil lawsuits against one or a few

corporate defendants in state or federal court. The plaintiffs allege that they were harmed by exposure to products produced by the defendants. Lawsuits may or may not be grouped in a class action. Law firms representing plaintiffs in mass torts often use advertising to locate and recruit plaintiffs.

Asbestos litigation is the largest mass tort in US history. As of 2002, 730,000 people had filed lawsuits against more than 8,400 defendants, and the cost of resolving claims was estimated at \$70 billion. The number of claims increased fourfold in the 1990s, and, in 2000 alone, 12 large companies reported that 520,000 new claims were filed against them. Because individual plaintiffs typically sue many defendants, estimates of the total number of asbestos claims range as high as 10 million. As of 2003, 73 corporations had gone bankrupt due to asbestos liabilities (Carroll et al. 2005). Asbestos litigation has been extremely profitable for lawyers, since 57% of spending goes to lawyers' fees (Carroll et al. 2003). Two studies in 2001 predicted that asbestos litigation in the USA would eventually cost \$200 billion (Angelina and Biggs 2001; Bhagavatula et al. 2001).

Asbestos was once considered to be a "miracle mineral" for its effectiveness as insulation and in preventing the spread of fires. It was used in ships, buildings, and consumer products, including wall-board, roofing, flooring, pipes, automotive brakes, hair dryers, children's toys, clothing, paper, and gardening products. Asbestos was used to coat the steel girders of skyscrapers such as the World Trade Center in New York, to insulate furnaces, and to make theater curtains fire resistant so that backstage fires would not spread to the seating area. Because asbestos had so many uses, estimates of the number of people who were exposed to it range from 27 to 100 million (Biggs et al. 2001).

But asbestos crumbles into microscopic fibers that become airborne and embed themselves in the lungs, causing a variety of diseases. Mesothelioma is cancer of the pleural lining around the chest and abdomen and is quickly fatal. Asbestosis is scarring of the lungs that reduces breathing capacity; it can range from non-disabling to fatal.

These two are "signature diseases" that are uniquely associated with asbestos exposure. Other asbestos diseases include lung cancer, gastrointestinal cancer, and pleural plaque, which is non-disabling thickening of the pleural lining. These latter conditions can be caused either by asbestos exposure or by other factors, such as smoking. Most asbestos diseases have a long latency period, so that they do not develop until 20–40 years after exposure. Individuals' likelihood of developing asbestos disease is low, but increases as the length and intensity of exposure rise (Carroll et al. 2003).

Asbestos exposure was recognized to be harmful as early as the 1920s and safe substitutes for many of its uses were developed in the 1930s. But it nonetheless became widely used – US consumption of asbestos grew from 100,000 metric tons in 1932 to 750,000 in 1994 (Castleman 1996, p. 788). Since then, asbestos use has fallen nearly to zero, but new cases of asbestos disease continue to occur because of the long latency period.

One question concerning asbestos is why government regulation did not prevent it from becoming so widely used. The British government began in the early 1930s to regulate workplace safety in the asbestos industry and provide workers' compensation to those disabled by asbestos exposure. In the USA, many states set up workers' compensation programs around the same time. However workers' compensation programs were oriented toward providing compensation for immediate workplace injuries, while asbestos exposure caused diseases that developed many years later and were not initially connected with asbestos exposure. Because statutes of limitation were short, most workers no longer qualified for compensation at the time they developed asbestos disease.

Workers' compensation systems also protected asbestos producers from liability for harm to their workers, since these systems were workers' exclusive remedy against their employers for workplace-related harm. Thus injured asbestos workers did not qualify for workers' compensation and also were barred from suing their employers for damage. And because employers were not liable for asbestos-related harm to their

workers, they had little incentive to improve workplace safety.

Workplace and product safety regulation also failed to protect workers who were exposed to asbestos. Occupational safety programs started in many US states in the 1950s and 1960s, but rules were often voluntary and poorly enforced. Some regulations actually increased workers' exposure to asbestos, such as building code regulations that required ventilation systems to be lined with asbestos insulation. As the insulation aged, it crumbled into microscopic fibers and fans blew the fibers through the workplace, where workers breathed them. Federal regulatory agencies such as the Occupational Health and Safety Administration (OSHA) and the Consumer Product Safety Commission (CPSC) came along in the 1970s and began to regulate asbestos exposure. But for many years, OSHA's workplace standards for preventing asbestos exposure were not tight enough to prevent workers from developing asbestos disease. Similarly, the CPSC's standards for limiting asbestos in consumer products in the 1970s and 1980s were mainly voluntary. Overall, state and Federal efforts to limit exposure to asbestos in the USA largely failed until the 1990s. This failure of regulation meant that many asbestos workers and product users suffered injuries due to asbestos exposure. This failure of regulation was not unique: other countries were no more successful in preventing asbestos exposure and they were generally slower than the USA to remove asbestos products from the market (White 2004; Wikipedia 2014).

In the next sections, I consider various factors that explain why asbestos litigation in the USA grew so large. I also review research on asbestos litigation and discuss various solutions – successful and unsuccessful – that have been proposed to resolve asbestos litigation.

Why Asbestos Litigation Grew

A combination of factors, rather than a single factor, was responsible for the growth of asbestos litigation. Because workers' compensation systems are workers' exclusive legal remedy against

their employers for on-the-job injuries, asbestos producers in the USA were not liable when their workers developed asbestos-related diseases. But asbestos producers were not shielded from liability to users of their products, and asbestos litigation therefore developed based on product liability law. The first successful trial of a lawsuit for damage due to asbestos exposure occurred in 1973 and involved an insulation worker who sued one of the large manufacturers of asbestos insulation (*Borel v. Fibreboard*, 443 F.2nd 1076 [5th Cir. 1973]). During the ensuing decade, 25,000 additional lawsuits were filed against asbestos product manufacturers and the number of lawsuits continued to grow. Because asbestos lawsuits were brought under products liability law rather than workers' compensation, plaintiffs could receive both compensatory and punitive damage awards. Damage awards could be in the millions of dollars, especially when juries awarded punitive damages (Berenson 2003).

One factor that favored asbestos plaintiffs was a change in products liability law in the 1960s that made producers strictly liable for harm to users of their products; previously, they were liable only if they were found to be negligent. The strict liability doctrine made producers liable as long as their products were “unreasonably dangerous,” or users were not adequately warned of the danger. The change from negligence to strict liability made it easier for plaintiffs to win asbestos lawsuits, both because asbestos products were extremely dangerous and because they rarely contained warnings.

Another factor that favored plaintiffs in asbestos litigation is that a number of plaintiffs' law firms specialized in handling asbestos claims. These law firms invested in developing evidence against asbestos producers that could be used in all of their lawsuits. The need for law firms to invest in developing evidence kept the number of entrants small, so that the asbestos litigation “industry” remained concentrated, with the ten top law firms representing 50–75% of asbestos claims filed. The high concentration meant that profits were high (Carroll et al. 2003).

In developing a strong legal case against asbestos manufacturers, plaintiffs' lawyers were aided

by the fact that several independent epidemiological studies were published in the 1960s that demonstrated strong links between asbestos exposure and asbestos disease. Asbestos plaintiffs' lawyers also developed evidence that producers conducted research on the health effects of asbestos exposure starting in the 1930s and found that exposure was harmful. But producers kept the results secret and did not warn workers or product users of the danger. This evidence of a cover-up of the dangers of asbestos exposure strengthened plaintiffs' claims in subsequent asbestos trials (Carroll et al. 2003).

The evidence suggesting a cover-up of the dangers of asbestos caused juries to frequently award punitive damages as well as compensatory damages in trials involving asbestos claims. One-sixth of all damage awards in asbestos lawsuits include punitive damages – a high proportion compared to other types of litigation. Unlike compensatory damages, punitive damage awards are often not covered by defendants' products liability insurance. The high damage awards and lack of insurance coverage made defendants eager to settle rather than litigate asbestos claims. But when claims frequently settle, they are very profitable for plaintiffs' lawyers to file, since lawyers' costs occur mainly at trial. This meant that plaintiffs' lawyers had an incentive to locate and file as many claims as possible.

Asbestos plaintiffs also benefit from the fact that they can sue many defendants. Typical asbestos plaintiffs sue 25 or more defendants, including producers of all of the asbestos products that they might have been exposed to while working or engaging in other activities. In a number of states, joint and several liability applies, so that each defendant found liable for damages is liable for the full amount of the damage award. Joint and several liability makes damage awards more valuable, since plaintiffs can collect up to the full amount of the award from any defendant(s) or their insurers. Thus even if some defendants pay little or nothing, damage awards can be collected from other defendants.

Another advantage that plaintiffs have in asbestos litigation is that their lawyers choose the most favorable court in which to file

lawsuits – a phenomenon known as “forum-shopping.” Plaintiffs' lawyers handling asbestos claims have a choice between filing in Federal versus state courts, and, if the latter, they can choose a state that has pro-plaintiff laws and legal procedures. Particular states are often favored because they do not require judges to approve lawyers' fees when claims are settled (this means legal fees can be higher), because they use joint and several liability and/or because they do not limit the size of punitive damage awards.

Within a particular state, plaintiffs' lawyers also choose a favorable location in which to file claims. Many asbestos claims are filed in out-of-the-way county courts where plaintiffs' lawyers have a relationship with local judges. These judges can help plaintiffs' lawyers by reducing defendants' ability to conduct pretrial discovery, scheduling trials at short notice so that defendants' lawyers have difficulty getting to the court in time, directing juries to consider awarding punitive damages, and pressuring defendants to settle. In return, plaintiffs' lawyers contribute to judges' reelection campaigns and benefit the local region by bringing in economic activity that raises demand for local hotels and restaurants. Favored locations for asbestos litigation in the past have included Madison, Illinois, Kanawha, West Virginia, and Jefferson County, Mississippi, as well as larger cities such as Philadelphia, Houston, and San Francisco – the latter because they are home to large shipyards and many former sailors who were exposed to asbestos while serving on navy ships.

Judges also developed new legal doctrines and legal procedures that favored plaintiffs and therefore encouraged plaintiffs' lawyers to file claims. One important change was a decision that greatly increased insurers' liability to asbestos claimants by legally reclassifying products liability insurance policies as premises insurance policies. While products' liability policies have a coverage limit that limits insurers' total liability under the policy to a fixed dollar figure, premises policies apply the coverage limit to each occurrence – where each individual asbestos claim is interpreted as an occurrence. Other legal changes

expanded insurers' liability for claims made after the time period when their policies were in effect. These changes greatly increased insurers' liability for asbestos damage by reviving old insurance policies that had already paid out their coverage limits (Epstein 1984; Anderson 1987).

Another legal change was that judges allowed multiple asbestos lawsuits to be litigated together, thus creating informal class actions. Asbestos plaintiffs' lawyers initially tried to have all asbestos claims certified as a class action in Federal court and settled all at once, but the Supreme Court overruled two settlements of class actions involving asbestos claims (*Amchen Products v. Windsor*, 117 S.Ct. 2231 (1997) and *Ortiz v. Fibreboard Corp.*, 119 S.Ct. 2295 (1999)). After these two decisions, plaintiffs' lawyers shifted to filing most asbestos claims in state courts. Judges in these courts allowed groups of lawsuits to be consolidated for either the pretrial or the trial stages of litigation, or both, using a procedure known as mass joinder. These consolidations often combined multiple claims by out-of-state plaintiffs with a small number of claims by in-state plaintiffs. The total number of claims consolidated ranged from a few to up to 9,600. Judges would hold a single trial before a single jury for all claims, with the jury sometimes making separate decisions for each plaintiff and sometimes making a single decision for all plaintiffs (Carroll et al. 2005). Combining multiple lawsuits for litigation benefits plaintiffs by making the trial outcomes more positively correlated. This makes going to trial more risky for defendants, because losing many cases at once could exhaust their insurance coverage and force them to file for bankruptcy. The more claims that are combined, the more bargaining power plaintiffs' lawyers have. Thus when large numbers of asbestos claims are consolidated for trial, defendants are likely to settle even claims that are legally weak.

Another legal change that benefitted plaintiffs in asbestos lawsuits is the use of bifurcated or reverse bifurcated trials. In a bifurcated trial, evidence concerning liability is presented first and the jury decides separately on each defendant's liability. Then the trial is suspended while

plaintiffs and defendants who have been found liable bargain over a settlement. If they fail to settle, the trial is resumed at a later date – sometimes with the same jury – for the damages portion of the trial. In a reverse bifurcated trial, the format is the same, but damages are tried in the first stage and liability in the second stage. Bifurcation saves on trial time relative to holding a unitary trial if the parties settle after the first stage. The parties are also more likely to settle at the end of the first stage than before the trial starts, since they have some of the information that the trial will generate.

Reverse bifurcation was developed specifically for asbestos trials and is particularly thought to benefit plaintiffs. This is because plaintiffs often have severe damage from their asbestos exposure – making damage awards high. In contrast, plaintiffs' claims are often weak on the liability side, because they cannot show that they were exposed to particular defendants' asbestos products. So using reverse rather than straight bifurcation strengthens plaintiffs' bargaining power in settlement negotiations, because the information generated by the first stage of the trial is very favorable to plaintiffs and raises their bargaining power in settlement negotiations.

Bouquet trials are another procedural innovation developed for asbestos litigation. In a bouquet trial, a small group of asbestos plaintiffs is selected for trial from a larger group of consolidated claims. The trial group includes plaintiffs with severe asbestos disease and plaintiffs with no impairment. The idea of the bouquet trial is that the outcomes at trial for the various types of plaintiffs will be used as a template for settling the remaining claims in the larger group. Using a bouquet trial allows larger numbers of claims to be consolidated, since a bouquet trial can be held even when the full consolidated group of claims is too large to hold a single trial. One well-known example is a trial of 12 asbestos claims in Mississippi that were selected from a larger group of 1,738 asbestos claims. At the bouquet trial, the jury awarded plaintiffs damage of \$4 million each. The prospect of the jury assessing similarly high damage awards for the remaining plaintiffs caused defendants to settle

all the remaining claims on very favorable terms (Parloff 2002).

Another factor that allowed the asbestos mass tort to grow so large is the large number of potential plaintiffs. As discussed above, the widespread use of asbestos meant that millions of people were exposed. Typical plaintiffs include ex-sailors who were exposed to asbestos on ships during World War II, workers who install insulation, workers in shipyards and steel mills, and textile workers who were exposed to airborne asbestos fibers in factories. Plaintiffs' lawyers search for new plaintiffs by extensive advertising and by conducting mass screenings. A frequent procedure was to bring a van equipped with an X-ray machine to a factory and take chest X-rays of all the factory workers. Any found to have scarring or thickening of the lungs or the pleural lining would be signed up as asbestos plaintiffs. Doctors often read hundreds of X-rays per day and found that nearly all of them had asbestos-related damage. More recently, plaintiffs' law firms have shifted to television advertisements to recruit plaintiffs whose exposure to asbestos may be non-work-related.

Another issue that has allowed asbestos litigation to become so large is that claims are valuable even when plaintiffs have no impairment from their asbestos exposure or had no asbestos exposure so that their claims are downright fraudulent. Because asbestos lawsuits are mainly settled rather than tried, non-impaired and fraudulent claims are valuable because they increase the size of consolidations and raise plaintiffs' lawyers bargaining power with defendants. Settlements cover both fraudulent and valid claims. Estimates suggest that as few as 10% of plaintiffs with asbestos claims have asbestos-related cancers – a widely used measure of disabling asbestos disease (Carroll et al. 2003). Legal standards that allowed non-impaired plaintiffs to collect damages are an important feature of asbestos litigation.

The asbestos mass tort also involves many types of defendants. In the first stage of the litigation, defendants were the major producers of asbestos insulation. These companies eventually went bankrupt. In the second stage, these defendants were replaced by producers of asbestos-

containing products, retailers that sold these products, and firms that operated workplaces containing asbestos. Examples include the automobile companies, sued because car brakes contained asbestos; Sears Roebuck, sued because its stores sold asbestos-containing products; 3M Corporation, sued because it made dust masks that didn't protect users from asbestos exposure if they used the masks improperly; and Crown Cork and Seal, sued because it briefly owned a company that included a division which produced asbestos-containing insulation. Crown Cork quickly sold the division that produced insulation, but nonetheless it eventually paid out \$700 million in asbestos settlements and damage awards. Both small and large firms were sued, since even small defendants have insurance. Each time new defendants were added to the litigation, previous plaintiffs filed new claims against them. Because there were so many plaintiffs and so many potential defendants, the asbestos mass tort continued to grow.

Finally, bankruptcy also played a role in encouraging asbestos litigation. Many of the large firms that produced asbestos insulation and asbestos-containing products went bankrupt due to their asbestos liabilities – the first was the Johns-Manville Corporation in 1982. When asbestos-producing firms go bankrupt, present and future damage claims against them are assigned to a trust which receives some or all of the reorganized firms' equity and uses the funds to pay compensation to asbestos victims. Congress adopted legislation defining these trusts in 1994 and required that they follow the general outlines of the Manville Trust that was set up following the Johns-Manville bankruptcy. Trusts first estimate the number and severity of future asbestos claims against them and then determine what level of compensation payments they can pay such that their funds will cover both present and future claims. Trusts payments vary with the severity of the claimant's asbestos disease and the length of exposure to asbestos. The trusts do not require that claimants show impairment from their asbestos exposure and they use quite loose standards for demonstrating exposure to the bankrupt firm's asbestos products. This was done in order to

reduce transactions costs and increase the fraction of damage payments that went to claimants rather than lawyers. However the loose standards for receiving compensation caused the number of claims to increase, causing many of the trusts to cut their compensation payments. On average, claimants with no asbestos-related impairment receive a total of around \$8,000 in compensation from all of the trusts, while claimants with moderate impairment receive around \$19,000. Compensation trusts have paid out a total of around \$17 billion to asbestos claimants (Scarcella et al. 2013).

The bankruptcy trusts encourage asbestos litigation in two ways. First, when corporations go bankrupt, their damage payments fall drastically. This encourages plaintiffs' lawyers to find new asbestos defendants to substitute for those that have gone bankrupt. The bankruptcies thus have contributed to bringing in many new corporations as defendants whose involvement with asbestos is increasingly remote. Second, although the trusts' compensation payments are relatively small, representing trust claimants is nonetheless profitable for plaintiffs' lawyers if they represent large numbers of claims. The trusts therefore encourage plaintiffs' lawyers to continue recruiting large numbers of non-impaired claimants, since the loose compensation rules allow these claimants to receive payments from many or all of the trusts.

Overall, a combination of factors is needed to explain why asbestos litigation grew so large.

Research on Asbestos Litigation

In White (2006), I examined why judges adopt the procedural innovations used in asbestos trials and the effect of both forum-shopping and procedural innovations on trial outcomes. The procedural innovations, discussed above, are consolidation of multiple lawsuits for trial, bifurcation and reverse bifurcation, and bouquet trials.

Why do judges adopt these innovations for asbestos trials? Judges in favored jurisdictions for asbestos litigation have crowded dockets. Because it would be impossible to hold individual trials for all cases, judges favor procedures that

encourage the parties to settle and therefore reduce trial time. Consolidating claims for trial is a method of reducing trial time, because only one jury must be selected and some of the evidence can be presented only once for all plaintiffs. Consolidation also increases the probability of settlement, because trial outcomes become more positively correlated and defendants therefore find it riskier to go to trial. Bifurcating trials reduces trial time relative to holding a unitary trial, because the information generated in the first phase of trial increases the probability of settlement when the parties bargain after the first phase. Finally, bouquet trials save trial time by allowing larger numbers of asbestos claims to be consolidated – if a trial is needed, then a bouquet trial can be held when the number of claims in the consolidation would otherwise make it too large for a single trial.

The study uses a dataset consisting of all asbestos lawsuits that were tried in court to a verdict on liability or damages or both between 1987 and 2003. Each observation consists of a trial of a single plaintiff's asbestos claim against all defendants. There were around 5,200 observations in the dataset, implying that less than 1% of asbestos plaintiffs' claims go to trials.

The data include the plaintiffs' alleged disease, the trial venue, the trial outcome, whether the claim was consolidated for trial and the number of claims in the consolidated group, whether the trial was bifurcated or reverse bifurcated, whether a bouquet trial was used, and the number of defendants that each plaintiff sued.

Half of all claims had individual trials, while the rest were consolidated with at least one other claim for trial. Approximately one-fifth of trials were bifurcated or reverse bifurcated and 4% were bouquet trials. Use of the procedural innovations was geographically concentrated: bifurcated trials were frequently used in Manhattan and Philadelphia, while bouquet trials mainly occurred in Mississippi. Sixty-four percent of plaintiffs were awarded compensatory damages and the average compensatory damage award (contingent on defendants being found liable) was \$1.3 million in 2003 dollars; 20% of plaintiffs were awarded punitive damages and the average punitive

damage award (contingent on defendants being found liable for both compensatory and punitive damages) was \$1.8 million. Plaintiffs' expected return from going to trial was \$1.1 million for the entire sample, with those having mesothelioma receiving around \$3 million more.

To examine the effect of consolidating claims for trial on the correlation of the trial outcomes, I computed a correlation coefficient for all trials involving two plaintiffs and compared the result with the correlation coefficient for single-plaintiff trials when plaintiffs were randomly assigned in pairs. I also followed the same procedure for three- and five-claim consolidations. The results show that the correlation coefficient of expected total damages ranges from 0.84 to 0.92 in the actual groups, compared to only 0.01–0.04 in the randomly assigned groups. The results were similar if only liability or only damages are considered. These results suggest that consolidating claims for trial makes trial outcomes much more positively correlated and supports the hypothesis that going to trial in a consolidation is much more risky for defendants.

To examine the effect of forum-shopping and the procedural innovations on trial outcomes, I estimated probit regressions explaining whether plaintiffs were awarded compensatory damages and whether they were awarded punitive damages conditional on receiving compensatory damages. I also estimated Tobit regressions explaining the amount of compensatory and punitive damages, with damages set equal to zero when the plaintiff loses. Forum-shopping was found to be extremely favorable to plaintiffs, with plaintiffs' probability of receiving compensatory damages increasing by up to 30 percentage points in the most favorable jurisdictions relative to the most commonly used jurisdiction. Also plaintiffs' probability of being awarded punitive damages rose by up to 91 percentage points in the most favorable jurisdiction relative to the most commonly used jurisdiction.

Use of the procedural innovations also increased plaintiffs' expected return from going to trial. Having a bifurcated trial raised plaintiffs' probability of being awarded compensatory damages by 27 percentage points and raised compensatory damage awards by \$924,000. Having

a bifurcated trial also increased plaintiffs' expected return from going to trial by \$650,000. But bifurcated trials did not significantly increase plaintiffs' probability of winning punitive damages or the size of the punitive damage award. Having a bouquet trial raised plaintiffs' probability of being awarded punitive damages and caused both compensatory and punitive damage awards to be higher. Plaintiffs' expected return from going to trial increased by \$1.2 million when a bouquet trial was held. Having a small consolidated trial consisting of 2–5 plaintiffs' claims increased plaintiffs' probability of winning both compensatory and punitive damages, but was associated with lower compensatory damage awards. Surprisingly, having a larger consolidated trial of six or more plaintiffs did not significantly change plaintiffs' returns from going to trial.

Overall the results suggest that the return to plaintiffs and their lawyers from filing asbestos claims is greatly increased by forum-shopping and by plaintiffs' lawyers picking jurisdictions where judges use the procedural innovations. Although the research did not address the issue of how forum-shopping and procedural innovations affect the size of asbestos settlements, the standard economic model of settlements suggests that they mirror trial outcomes and are higher in courts where plaintiffs' expected returns from going to trial are higher (Mnookin and Kornhauser 1979). Thus forum-shopping and procedural innovations are also likely to raise the amount that defendants pay to settle asbestos claims.

Methods of Resolving Asbestos Litigation: Hypothetical and Actual

In this section, I consider solutions for resolving asbestos litigation – including both proposed solutions that were never adopted and actual solutions that were.

One proposed solution in the 1990s was to certify a class action of all asbestos claimants. In a class action, all asbestos claims are combined in a single lawsuit and all are resolved at once, usually by a settlement. Both present and future

asbestos claims are resolved. Individual plaintiffs would be bound by the outcome of the class action and would not have had the right to opt out. The Federal courts certified two class actions of asbestos claimants, but – as discussed above – the US Supreme Court rejected both class certifications in 1997 and 1999, on the grounds that asbestos claimants were too diverse to be combined into a single class.

This was followed by another proposed solution for asbestos litigation: a Federal government-administered compensation scheme for asbestos victims. The proposed bill was the Fairness Asbestos Injury Resolution or “FAIR” Act of 2005, S. 852. It was based on previous federally administered programs, one that compensated miners who developed black lung disease and one that compensated children harmed by childhood vaccines. Compensation of up to \$140 billion would have been financed by levies on asbestos producers and insurers. Asbestos victims would lose their right to file lawsuits, but would instead receive compensation from the trust. Claimants who had mesothelioma or cancer would receive the highest awards of \$1.1 million and those with less disabling diseases would receive \$25,000 or more. Non-impaired claimants would receive medical monitoring, but no compensation (Stengel 2006). However the FAIR Act was not enacted (Barnes 2011).

While both the class action settlement and the compensation scheme for asbestos claims failed, courts began in the early 2000s to adopt new procedural innovations that reduced the amount of asbestos litigation. One such device was the “inactive docket” which put claims by non-impaired asbestos plaintiffs on an inactive basis, preserving their right to sue in the future, but preventing their claims from proceeding in the legal system until they become impaired from their asbestos exposure. Inactive dockets solve the problem that plaintiffs must file claims quickly after discovering their asbestos-related harm in order to satisfy statutes of limitations. But because most asbestos claims are classified as inactive, asbestos litigation now consists mainly of plaintiffs who have severe asbestos-related diseases.

As a result of the use of inactive dockets, plaintiffs’ lawyers can no longer litigate large groups of claims consisting mainly of non-impaired plaintiffs, and they therefore have less bargaining power to force defendants to settle. The fraction of asbestos damage awards going to non-impaired plaintiffs has fallen from around 50% in 1997–1999 to less than 5% in 2013 (Scarcella et al. 2013). This change has greatly reduced plaintiffs’ lawyers’ return from recruiting non-impaired asbestos claimants. It has been so successful in reducing the volume of asbestos litigation that an observer is led to wonder why judges did not adopt it much earlier.

Another recent development is that some states that were centers for asbestos litigation have adopted legal reforms to discourage the filing of asbestos claims in the state. An important change in several states was to bar judges from consolidating out-of-state with in-state asbestos claims for litigation. As a result, out-of-state claims could no longer be litigated in the state and therefore plaintiffs’ lawyers could no longer put together large consolidations. Among states that previously allowed large consolidations of asbestos claims, West Virginia, Mississippi, and Illinois all made changes along these lines in the early 2000s. Several other states changed their legal rules to explicitly disallow large consolidations of asbestos claims, although they generally still allow out-of-state claims to be consolidated with in-state asbestos claims. Another change is that New York, Texas, and several other states substituted proportional liability for joint and several liability to asbestos claimants, so that individual defendants are no longer liable for plaintiffs’ entire damage award. This shields non-bankrupt defendants from being held liable for bankrupt defendants’ share of plaintiffs’ damage (Hanlon and Geise 2007). The result of these changes in state law is that most asbestos litigation now involves a much smaller number of claims by plaintiffs with serious asbestos-related diseases and these claims are litigated individually or in small groups.

Finally, judges have become more likely to dismiss fraudulent claims, rather than pressure defendants to settle them. This approach was used recently to resolve a different mass tort:

claims for damage due to silica exposure. Silica litigation was a spinoff from asbestos litigation and took a similar form. Plaintiffs allege harm from inhaling airborne silica crystals that can lead to scarring of the lung lining, silicosis, or lung cancer. Because of the similarity between asbestos disease and silica disease, plaintiffs' lawyers recruit silica claimants using the same mass screenings with chest X-rays that they use to recruit asbestos claimants. In fact, plaintiffs' lawyers often file both silica claims and asbestos claims on behalf of the same individuals, using the same chest X-rays; this is despite the fact that it is rare for individuals to have been exposed to both silica and asbestos. However the judge who presided over the silica litigation dismissed nearly all of the claims on the grounds that they were fraudulent and threatened to bring criminal charges against the doctors who read the plaintiffs' X-rays. This effectively ended the silica mass tort, leaving only a small number of lawsuits by plaintiffs with severe silica-related disease. The publicity given to the silica litigation has probably made judges more likely to dismiss asbestos claims as well (Behrens and Goldberg 2005/2006).

Future Directions

Because asbestos litigation has been so lucrative, plaintiffs' lawyers have searched widely for other defective products that could serve as the basis for new mass torts, using the techniques they developed for asbestos litigation. Among potential future mass torts are litigation involving harm due to exposure to lead paint, harm due to guns, and claims of obesity due to consumption of fast food (White 2004). However none of these spinoff mass torts have been successful in court.

But the asbestos mass tort itself continues to mutate into new forms that keep it alive. One recent development is lawsuits filed by family members of asbestos workers who claim second-hand exposure to asbestos from relatives' clothing. Family members, unlike workers themselves, are not barred by workers' compensation from suing their relatives' employers. Thus they can

both sue their relatives' employers and the producers of asbestos products that their relatives were exposed to. Another new development in asbestos litigation is claims by lung cancer victims against asbestos producers and the asbestos bankruptcy trusts. Most lung cancer is caused by smoking, but plaintiffs with lung cancer nonetheless claim that their cancer was caused by exposure to asbestos. These claims qualify for compensation from the asbestos bankruptcy trusts, and, because lung cancer is a serious disease, their lawsuits against non-bankrupt defendants are not placed on the inactive docket (Nocera 2013). And since there are 200,000 new lung cancer cases each year compared to only 2–3,000 new mesothelioma cases, lung cancer claims present a valuable opportunity for lawyers to continue the asbestos mass tort.

References

- Anderson DR (1987) Financing asbestos claims: coverage issues, Manville's bankruptcy and the claims facility. *J Risk Insur* 54(3):429–451
- Angelina M, Biggs J (2001) Sizing up asbestos exposure. *Mealey's Litigation Report: Asbestos* 16:32–38
- Barnes J (2011) Dust-up: asbestos litigation and the failure of commonsense policy reform. Georgetown University Press, Washington, DC
- Behrens MA, Goldberg P (2005/2006) The asbestos litigation crisis: the tide appears to be turning. *Conn Insur Law J* 12:477
- Berenson A (2003) 2 Large verdicts in new asbestos cases. *New York Times*, 1 Apr 2003
- Bhagavatula R, Moody R, Russ J (2001) Asbestos: a moving target. *AM Best Rev* 102:85–90
- Biggs JL et al (2001) Overview of asbestos: issues and trends. Report prepared by the American Academy of Actuaries Mass Torts Work Group. Available at www.actuary.org/pdf/casualty/mono_dec01asbestos.pdf
- Carroll S, Hensler D, Abrahamse A, Gross J, White M, Scott Ashwood J, Sloss E (2003) Asbestos litigation costs and compensation. DRR-3280-ICJ. RAND Corporation, Santa Monica
- Carroll S, Hensler D, Gross J, Sloss EM, Schonlau M, Abrahamse A, Scott Ashwood J (2005) Asbestos litigation. DRR-3280-ICJ. RAND Corporation, Santa Monica
- Castleman BI (1996) Asbestos: medical and legal aspects, 4th edn. Aspen Law & Business, Englewood Cliffs
- Epstein RA (1984) The legal and insurance dynamics of mass tort litigation. *J Legal Stud* XIII:475–506
- Hanlon PM, Geise ER (2007) Asbestos reform – past and future. *Mealey's Litigation Report: Asbestos* 22:5, 4 Apr 2007

- Mnookin R, Kornhauser L (1979) Bargaining in the shadow of the law: the case of divorce. *Yale Law J* 88(5):950–997
- Nocera J (2013) The asbestos scam. *New York Times*, 2 Dec 2013
- Parloff R (2002) The \$200 billion miscarriage of justice. *Fortune*, 4 May 2002
- Scarcella MC, Kelso PR, Cagnoli J (2013) Asbestos litigation, attorney advertising and bankruptcy trusts: the economic incentives behind the new recruitment of lung cancer cases. *Mealey Asbest Bankrupt Rep* 13:4
- Stengel JL (2006) The asbestos end-game. *NYU Annu Surv Am Law* 62:223
- White MJ (2004) Asbestos and the future of mass torts. *J Econ Perspect* 18:2
- White MJ (2006) Asbestos litigation: procedural innovations and forum-shopping. *J Legal Stud* 35(2): 365–398
- Wikipedia (2014) Asbestos and the law. http://en.wikipedia.org/wiki/Asbestos_and_the_law, viewed 21 Apr 2014

Measure of Environmental Regulation

Vittoria Colombo
Università degli Studi di Torino, Turin, Italy

Definition

Environmental law and economics has been largely focused on the effect that different environmental rules can have on economic outcomes, while only few detected the difficulty of analysis of such subject, mainly due to the measurement of environmental stringency. There is no homogeneous way to measure environmental regulation, due to its peculiarities; environmental rules are often designed based on elements other than just the legal principle, such as geographical characteristics, scientific data, the level of pollution and the industry sector involved, and elements that must be taken into account in order of not incurring in methodological mistakes. This essay examines the main approaches and indicators of environmental regulation, considering the advantages and disadvantages of any of them. It also analyzes the main risks of a wrong

methodological approach to the measurement of environmental rules.

Environmental regulation is an atypical form of regulation, which needs to be assessed in its entirety and complexity. Its design requires scientific and technical information that are pivotal for reaching its objectives. Environmental rules are difficult to assess, since decisions on how to design a rule do not only depend on legal elements but also on scientific data, risk assessment, and scientific criteria.

Environmental Stringency

Currently the most used method of measuring environmental regulation is by looking at costs imposed by it, namely, at the *environmental stringency*, which in its broad and shared definition is the “cost of polluting by firms across different sectors and policy instruments,” where “a higher value represents a more stringent policy.” The use of environmental stringency allows to ascertain environmental regulation as a whole, and at the same time, it is sufficiently general to allow scholars to use different methods and datasets. The main conceptual difficulties in analyzing environmental stringency can be summarized in seven broad problems: multidimensionality, simultaneity and identification, trade-off between *de jure* and *de facto* situations, industrial composition, sampling, grandfathering, and lack of data (Brunel and Levinson 2013).

Multidimensionality

Environmental regulation is typically *multi-dimensional* (Wing-Hung et al. 2011): it deals with different media, such as soil, water, and air, as well as different pollutants that affect those media (dioxin, chemicals, carbon dioxide, etc.). Moreover, it has different recipients: firms, consumers, households, and even public administration itself. Finally, it can set standards or limits that vary on the basis of several factors, which can change depending on the environmental quality of territory, on the technology

used by firms, or on the sector in which firms operate (Kozluk 2014a). On the one hand, there could be many environmental regulations applying to the same industry, aiming at protecting different media (water, land, human health, etc.); on the other hand, policy-makers' decisions are also relevant, with the same rules applied in different ways depending on several internal or external factors that have a link with environmental standards. The level of emissions, the location of the activity, the technology applied in a certain plant, or the sector in which the firm operates can have a decisive weight on the decision not only on the application of a certain rule but also on the intensity of the instrument itself. The intensity of the application of instruments depends on factors that are linked with the environment, and the decision to forbid a certain production method in a region could depend on specific geographical characteristics of the territory.

Multidimensionality is the reason why case studies on a specific industry sector are limited, because they are not representative of other categories not covered by that specific regulation (Sauter 2014).

De Jure and De Facto Applications

The *de jure* stringency derives from environmental legislation applied in a certain legal system *ex ante*, while the *de facto* one corresponds to the real application of the *de jure* instruments. This assessment is made by looking at courts' decisions, namely, at the difference between regulation and its enforcement. *De facto* application of the rules may also involve other informal elements that are not easily detectable as judicial decisions: enforcement of rules does not only occur in courts but can take different shades, i.e., it could consist in a fine given by the administrative local authority but also in some checks made by the public authority in the concerned sites. For some environmental regulations, soft mechanisms are used, i.e., flexibility mechanisms or the provision of consultations with the authority in charge of enforcement.

Simultaneity and Identification

Simultaneity is the difficulty of identifying the direction of causality between the rule and its effect. Countries may introduce stricter environmental regulations in order to address concerns on high-polluting industries. However, many pollution-intensive industries have consistent political power and can lobby decision-makers to enact lax environmental regulation. This makes difficult to assess the causal link between a certain environmental policy and its results (Brunel and Levinson 2013). Similar to the problem of simultaneity is the broader problem of identification, which in this case occurs when there is a problem in identifying whether a certain result is due to an environmental rule or to other regulatory instruments, i.e., legislation on employment, or to certain market characteristics (Malatu 2008).

Some scholars conducted natural experiments to avoid such problem, taking into account relevant changes in regulation forced by external factors (as a Supreme Court decision or international treaties' implementation), and comparing the *ante-* and *post-*scenarios (McConnell and Schwab 1990; Henderson 1996), others looked at instrumental variables correlated with the regulatory stringency (Xing and Kolstad 2002; Levinson and Taylor 2008) but uncorrelated with the measure of economic activity.

Industrial Composition

Industrial composition of countries may be due to endogenous sources that are not measurable, such as the geographic characteristics or natural resources in the country, i.e., carbon or natural gas. Industrial composition, if not considered, can cause methodological problems.

Sampling and Grandfathering

Sampling deals with the causal link assessment but related to the sample of industries subject to a given environmental policy. An industry's market

share in a country can be itself the result of a determined environmental policy, i.e., high-polluting industries can be present more in a country than in another, because of that country's environmental policy toward a certain industry sector. This problem has to be taken into account when assessing the effect of policy stringency in a sector.

Also grandfathering, namely, the provision according to which new rules apply only to new situations, while the old rules continue to apply to existing situations, could create problems. In the field of environmental law, stricter rules may apply to new plants or pollution sources, while lax ones are applied to old sources of pollutions that, most probably, are also the most polluting ones. This problem concerns more consumers than firms: whereas the former ones are not bound to buy new products – consumers are not obliged to buy a new car and can continue to drive the old one, which is much more polluting than new models – the latter ones are often shifting to new technologies because they are obliged by law to do so (Heutel 2010).

Lack of Data

Lack of data must not be intended in absolute terms, but as lack of data giving source to a robust outcome. This factor is linked with multi-dimensionality, identification, simultaneity, and sampling; often the available data consider just a small part of environmental legislation while not dealing with the assessment of the quality and quantity of all environmental regulation applied in a determined country or region.

Approaches in Measuring Stringency

Although the definition of environmental stringency is widely accepted, it is not the same for what it concerns its *measure*.

The main challenge of measuring environmental stringency is the delineation of a conceptual framework that considers all the principal layers of

environmental law and that includes them in the analysis (Sauter 2014). The distinction between market- and nonmarket-based instruments allows to assess the impacts that these two regulatory approaches may have. However, this distinction alone can be misleading and too reductive, because it does not consider other elements such as flexibility and stability of regimes (Johnstone et al. 2010). De Serres (De Serres et al. 2010) distinguishes four categories of market-based instruments: (a) environmental taxes, (b) pollution trading systems, (c) deposit-refund systems, and (d) subsidies for incentivizing environmental friendly activities. Nonmarket-based instruments include the general category of (a) command and control regulations, (b) technology-support policies, and (c) voluntary approaches.

The main approaches for measuring environmental regulation will be presented in the following.

The first approach looks at changes in a single regulation or rule. The second looks at perceptions that firm directors, public officials, and, more generally, people directly touched by environmental regulation have on its stringency. The third considers shadow prices for environmental inputs or expenditures related to environmental rules. The fourth looks at the variation of environmental performances to deduct a more or less stringent environmental legislation. Finally, composite measures aim at avoiding multi-dimensionality, by including all the elements in the same indicator.

Change in a Single Regulation

The advantage of looking at the effect of a change in a single regulation is the level of certainty it can bring to the analysis; the risk of multi-dimensionality is highly reduced, as well as the one of simultaneity and identification. In addition, this kind of analysis is more feasible than the general one, because of the availability of data. However, the limits are relevant. By looking at the single regulation, results will be applicable only to it and not to other sectors (Smarzynska and Wei 2001). The reason is exactly multidimensionality that, when eliminated though a circumscribed

analysis, limits the extension of the same results to other cases.

Analysis Based on Perceptions

The majority of datasets on environmental stringency are based on perceptions by businessman and firms' directors or civil servants. The advantage of such approach is the availability of data and the reliability of the analysis, which can be circumscribed by asking the same questions on a limited number of regulations or media. However, the premises are not always exempted from critics. When data are used for a cross-country analysis, perceptions are difficult to compare, since the requisites for an objective perception of situations by individuals are the knowledge of the other terms of comparison, which in such case are assumed (Becker and Henderson (2000). Interviewees should be aware of the strictness of environmental regulation in other countries in order to give an objective opinion on the one of their country. Moreover, according to some authors, perception-based surveys involve some sampling problems, given by the fact that "the sample of respondents may actually be a result of environmental policies" (Becker and Henderson (2000), Botta and Kozluk (2014), p. 11).

In addition, complexity of environmental regulation makes sometimes difficult to disentangle pollution abatement costs from other costs, and sometimes for firms, it could also be beneficial to indicate wrong estimates. Finally, also the perception of the strictness of a determined regulation could be influenced by external factors and therefore be unreliable (Gallaher et al. 2006). One of such cases is when reporting is affected by business cycles: in periods of crisis, environmental regulation may be considered stricter than in other times (Brunel and Levinson 2013).

Shadow Prices and Performance Indicator Approach

Shadow prices are the costs of polluting that can be calculated from the firm's production function.

The production function becomes the starting point for the calculation of policy stringency; emissions would be considered as an element to

calculate the environmental stringency of regulation. The advantage of such approach is that, by looking directly at emissions, it considers the de facto situation and therefore assesses the real policy stringency and not the de jure one.

Performance indicators, as for the shadow price, look at the amount of emissions actually produced or at the pollution intensity in a specific media. However, contrary to the former approach, this one directly considers the amount of pollution produced as sign of environmental strictness.

Three main performance indicators have been used by the literature so far: emissions, energy consumption, and gasoline. The problem is that it cannot exclude multidimensionality and identification issues (Botta and Kozluk 2014). Most of the indicators quantify the problem to be addressed and not the stringency of environmental regulation. Identification problems are consistent, since the amount of emissions can be due to factors other than environmental regulations, which can be associated with the particular geographical structure, to the effect of other policies, and to other external factors (Albrizio et al. 2014). Also simultaneity can occur, since the amount of emissions could influence regulation itself.

Composite Measures

Composite measures seek to introduce multidimensionality into the analysis, trying not to eliminate the inevitable complexity of environmental regulation, but including it into an index. By doing this, composite measures are more complete, because they take into account all the factors of multidimensionality (Smarzynska and Wei 2001).

However, risks are very high: given the complexity of environmental legislation, the index must be adequately constructed and fine-tuned in order avoid methodological shortcomings, which in this case can be less evident and therefore more dangerous. Another disadvantage of indexes is that they are representatives of the only de jure situation. Such shortcoming is not negligible, since the enforcement and application of environmental regulation are crucial elements of environmental stringency (Sauter 2014, p. 8).

Choice of Indicators for Measuring Environmental Stringency

This section presents the indicators and proxies adopted by the literature for measuring stringency, namely, pollution abatement costs, the quantity of polluting emissions produced, and, finally, indexes, and then combines different measures in order to cover all possible elements of environmental stringency.

Pollution Abatement Costs

Environmental stringency has often been defined by using the proxy of the private sector's pollution compliance costs (Keller and Levinson 2002; Levinson 1999), namely, costs that firms face to comply with environmental regulation (Carraro et al. 2010). Compliance costs can take the form of environmental taxes, liability rules, emissions limits, and technological standards. They can also include permitting costs, regulatory delays, threat of lawsuits, and the redesign of a process or product (Levinson and Taylor 2008).

Some authors considered all types of pollution abatement costs and expenditures while others only a particular media (Brunnermeier and Cohen (2003).

The principal and most used method to calculate pollution abatement costs is based on surveys collected at regional and industry level about firms' costs on abating pollution. One of the most used and complete datasets is the PACE survey (Pollution Abatement Costs and Expenditures) which has been conducted for the first time on 1994 in the United States on annual data from 1973 and then repeated with few modifications on 1999 and 2005. It collects data on pollution abatement capital expenditures and operating costs mainly in the manufacturing industry of the United States. Other countries started collecting data on environmental expenditures. Above them are Canada, with its SEPE (*Survey of Environmental Protection Expenditures*), and the European Union with the Eurostat *Questionnaire on Environmental Protection Expenditure and Revenues* (EPE). Also the OECD has drafted a dataset on environmental protection expenditure and revenue.

Public Entities' Pollution Abatement Expenditures

The costs of the environmental effort made by the state or other public entities to enforce environmental protection do not use directly environmental regulation, but some proxies, and were more frequent some years ago, when environmental datasets were scarce and imprecise.

Proxies for public expenditure could consist in a mix between private and public expenditures (Pearce and Palmer 2001), in the state budget for enforcement and inspections (Gray 1997), or in the number of employees in environmental agencies in relation to the number of manufacturing industries (Levinson 1999).

The advantage of such approach is that it often includes the enforcement stage, which would allow assessing also the de facto scenario. However, if the aim is to assess the impact that environmental stringency has on firms, the method of looking at public expenses, which may include also environmental actions for consumers or householders, could bring to distorted results. Second, such proxies do not have a strong causal link with environmental stringency, making results not completely reliable.

Emissions

Polluting emissions can be used as a measure of stringency, in particular emissions in air and energy consumption. Scholars have considered the emission indicators in different ways; for some, a high number of emissions indicate that environmental regulation is too permissive; for others it demonstrates the tendency to a stricter environmental regulation. In particular, the literature on emissions as measure of stringency is divided in two branches: the first one considers regulation as exogenous, where environmental rules are given by external sources, such as a central government or an international organization. The second one considers emissions as a factor of production as any other in the firm's profit maximization dynamic.

With regard to the first case, some used the emissions' reduction as indicator for stringency (Smarzynska and Wei 2001; Gullop and Roberts 1983). For the second one, emissions produced by

firms are intended as a factor of production like any other, with the consequence that environmental stringency can be assessed through the amount of emissions that the firm decides to produce. This approach is based on the neoclassical assumption of the profit-maximizing firms: a firm that decides to maximize its profits would use its factors of production until the marginal revenue of the product equals its price (Levinson 2001).

Currently there is no global instrument that collects emissions for each country. Datasets on emissions are available only for Europe and the United States. In Europe the *European Pollutant Release and Transfer Register* (E-PRTR), which replaced the *European Pollution Emissions Register* (EPER), collected data on emissions into air and water from 2007 to 2014. In the United States, the US Environmental Protection Agency listed a new dataset, the *Trade and Environmental Assessment Model* (TEAM). TEAM considers emission produced as factor inputs in the production process, in that it combines these data with the industry sector, location, and trade agreements, being therefore able to assess the effect that a certain trade agreement could have on pollution discharge (Creason et al. 2005). Such database is available for three different periods, 1997, 2002, and 2007, only for the United States.

Indexes

Through the inclusion of individual indicators into a unique instrument, indexes can tackle many of the multidimensionality and complexity issues (Smarzynska and Wei 2001).

The first complete index is the one that Dasgupta et al. (2001) prepared in the occasion of the *United Nations Rio Conference on Environment and Development* on the basis of national reports prepared by public authorities. Public officials and NGO representatives had to answer questions regarding environmental awareness, the scope of environmental policies and regulations, the presence of control mechanism, and the implementation of environmental regulation. The analysis included four media (water, soil, air, biodiversity). However, such index was composed only for 1 year and afterward extended by Eliste and Fredriksson (2002) for the agricultural sector.

Another well-known index is the CLIMI (*Climate Laws, Institutions and Measures Index*) prepared for the European Bank for Reconstruction and Development (EBRD) in 2010 on the basis of the UN country reports and the UNFCCC (United Nations Framework Convention on Climate Change) reports. The CLIMI includes three dimensions: international cooperation, domestic climate framework, and fiscal or regulatory measures. The disadvantage of this index is that it considers only rules designed to address climate change and that it was drafted only for the year 2010 (Surminski and Williamson 2012).

The World Economic Forum used another method; businessmen and firm directors had to rate the regulatory stringency of the country in which the firm operates and to give them a grade from one to seven, where one represents very lax regulation and seven is the strictest environmental stringency. The result, called *Environmental Sustainability Index* (ESI), is complete, although it has the intrinsic shortcoming of being based on subjective perceptions and not on objective elements. The ESI has been used by Esty and Porter (2002) to build the *Environmental Regulatory Regime Index* (ERRI), together with the data on the Global Competitiveness Report 2001–2002 annual survey on business and government leaders (Esty and Porter 2005, p. 86).

EPI (Environmental Performance Index) measures environmental performance of countries in two main fields, human health and protection of the ecosystem. It is constructed through the use of 20 indicators reflecting national-level environmental data. The EPI represents a good source for panel data studies, since it examines more than 200 countries over 10 years. The benchmark for score application is the reaching of a “proximity target,” which corresponds to objectives fixed by international or national regulations, guaranteeing therefore a de facto assessment. Finally, the OECD has recently created a new index on environmental stringency, the *Environmental Policy Stringency Index* (EPS), which deserves a separate section for its peculiarities.

Cross-References

► Environmental Policy: Choice

References

- Albrizio S, Kozluk T, Zipperer V (2014) Empirical evidence on the effects of environmental policy stringency on productivity growth. OECD Economics Department working papers, vol 1179. OECD Publishing, Paris
- Becker RA, Henderson JV (2000) Effects of air quality regulations on polluting industries. *J Polit Econ* 108:379–421
- Botta E, Kozluk T (2014) Measuring environmental policy stringency in OECD countries: a composite index approach. OECD economic department working papers, n. 1177
- Brunel C, Levinson A (2013) Measuring environmental regulatory stringency. OECD trade and environment working papers, 2013/05. OECD publishing, Paris
- Brunnermeier SB, Cohen MA (2003) Determinants of environmental innovation in US manufacturing industries. *J Environ Econ Manag* 45:278–293
- Carraro C, De Cian E, Nicita L, Massetti E, Verdolini E (2010) Environmental policy and technical change: a survey. *Int Rev Environ Resour Econ* 4:163–219
- Creason J, Fisher M, Morin I, Stone SF (2005) Comparison of the environmental impacts of trade and domestic distortions in the United States. Environmental economics working paper series, EPA, Washington DC, USA
- Dasgupta S, Wheeler D, Roy S, Mody A (2001) Environmental regulation and development: a cross-country empirical analysis. *Oxf Dev Stud* 29:173–187
- De Serres A, Murtin F, Nicoletti G (2010) A framework for assessing green growth policies. OECD Economic Department working papers, vol 774. OECD publishing, Paris
- Eliste P, Fredriksson PG (2002) Environmental regulations, transfers and trade: theory and evidence. *J Environ Econ Manag* 43:234–250
- Esty, Daniel, and Michael E. Porter (2002) Ranking national environmental regulation and performance: a leading indicator of future competitiveness? In *The Global Competitiveness Report 2001–2002*, by Michael E. Porter, Jeffrey D. Sachs, Peter K. Cornelius, John W. McArthur, and Klaus Schwab, 78–101. New York: Oxford University Press
- Esty D, Porter ME (2005) Ranking national environmental regulation and performance: a leading indicator of future competitiveness? Available at http://www.hbs.edu/faculty/Publication%20Files/GCR_20012002_Environment_5d282a24-bb10-4a9a-88bd-6ee05e8c6678.pdf
- Gallaher MP, Murray BC, Nicholson RL, Ross MT (2006) Redesign of the pollution abatement costs and expenditures (PACE) survey: findings and recommendations from the pretest and follow up visits. Final report for the US Environmental Protection agency, EPA, Washington DC, USA
- Gray W (1997) Manufacturing plant location: does state pollution regulation matter? NBER working papers, vol 5880. National Bureau of Economic Research, Cambridge, MA
- Gullop F, Roberts J (1983) Environmental regulations and productivity growth: the case of fossil-fueled electric power generation. *J Polit Econ* 91:654–674
- Henderson JV (1996) Effects of air quality regulation. *Am Econ Rev* 33:789–813
- Heutel G (2010) Plant vintages, grandfathering and environmental policy. *Econ Fac Publ* 14:1–59
- Johnstone N, Hascic I, Kalmova M (2010) Environmental policy characteristics and technological innovation. *Econ Polit* 27:277–302
- Keller W, Levinson A (2002) Pollution abatement costs and foreign direct investments inflows to U.S. states. *Rev Econ Stat* 84(4):691–703
- Kozluk T (2014a) Measuring environmental policy stringency in OECD countries: a composite index approach. OECD economic department working papers, vol 1177. OECD Publishing, Paris, p 35
- Kozluk T (2014b) The indicators of the economic burdens of environmental policy design: results from the OECD questionnaire. OECD Economics Department working papers, vol 1178. OECD Publishing, Paris
- Levinson A (1999) An industry-adjusted index of state environmental compliance costs. NBER chapters behavioral and distributional effects of environmental policy. Cambridge, MA pp 131–158
- Levinson A (2001) An Industry adjusted index of state environmental compliance costs. In: Behavioural and distributional effects of environmental policy. University of Chicago Press, Chicago
- Levinson A, Taylor MS (2008) Unmasking the pollution haven effect. *Int Econ Rev* 49:223
- Malatu A (2008) Weighting the relative importance of environmental regulation for industry location. University of Manchester, Economics Discussion paper, EDP-0803
- McConnell VD, Schwab RM (1990) The impact of environmental regulation on industry location decisions: the motor vehicle industry. *Land Econ* 66:67–81
- Pearce D, Palmer C (2001) Public and private spending for environmental protection: a cross-country policy analysis. *Fisc Stud* 22:403–456
- Sauter C (2014) How should we measure environmental policy stringency? IRENE working paper 14–01. Institute of Economic Research, University of Neuchatel, pp 4, 5
- Smarzynska BJ, Wei SJ (2001) Pollution havens and foreign direct investment: dirty secret or popular myth? In: The BE journal of economic analysis and policy. Berkeley Electronic Press, 3(2):pp 1–34
- Surminski S, Williamson A (2012) Policy indexes. What do they tell us and what are their applications? The case of climate policy and business planning in emerging markets. Working paper, vol 101. Centre for Climate Change Economics and Policy

- Wing-Hung L, Zhan X, Liu N (2011) Understanding styles of corporate compliance with environmental regulation: towards a multidimensional conceptual framework. Paper prepared for the 11th national public management research conference Syracuse University
- Xing Y, Kolstad C (2002) Do lax environmental regulations attract foreign investment? *Environ Resour Econ* 21:1–22

Media

Peter T. Leeson and Joshua Pierson
Department of Economics, George Mason
University, Fairfax, VA, USA

Abstract

Citizens can use media to solve political agency problems. To be useful for this purpose, however, media must be free. Freer media are strongly associated with superior political-economic outcomes and may be especially important to fostering political-economic improvement in the developing world.

Synonyms

[Mass media](#)

Definition

Media refer to means of mass communication such as the Internet, television, radio, and newspapers.

Media as a Mechanism for Controlling Government

Citizens in democratic regimes face a principal-agent problem with respect to their elected governors. While citizens empower political officials to wield government's authority in citizens' interests, left unchecked, political officials are tempted to use that authority for personal benefit. Democratic elections provide a means by which

citizens may hold elected officials accountable for their uses of government authority. However, citizens' ability to use the voting both for this purpose depends crucially on the extent of their knowledge about political actors' behavior and their ability to coordinate responses to such behavior.

Media can improve democracy's ability to help citizens address the principal-agent problem they face with respect to elected officials (Besley and Burgess 2001, 2002; Coyne and Leeson 2004, 2009a, 2009b, Leeson and Coyne 2007). By reporting on political actors' behavior, media inform citizens about the activities of political actors that are relevant for citizens' evaluation of such actors as stewards of citizens' interests. Moreover, by providing such information to large numbers of citizens, media can help coordinate citizens' responses to what they learn about political actors' behavior, rewarding faithful stewards of their interests through, for example, reelection and punishing bad stewards through popular deposition and/or refusal to reelect them.

Media can also help citizens address political agency problems through the foregoing channel by influencing who seeks political office. Where would-be political officials know that the private benefits of holding political office are low because of media-provided information and media-facilitated citizen coordination, individuals who desire political power to further their own interests rather than citizens' are less likely to seek elected office.

Media Freedom and Government Control of Media

The extent to which media can assist citizens in addressing the principal-agent problem they face with respect to political officials depends on media's freedom. Media freedom (or independence) refers to the extent to which government can directly or indirectly control the content of media-provided information reaching citizens. Where media freedom is higher, government's ability to influence the content of media-provided information is weaker and vice versa.

Government control of media can take many forms (Leeson and Coyne 2005). The most direct form is state ownership of media outlets. For example, all North Korean media outlets are controlled by the Korean Workers' Party or other appendages of the North Korean government. Elsewhere, media outlets are not owned by government, but are nevertheless owned by powerful people in government, creating a similar situation. Italy under the prime ministership of Silvio Berlusconi, who was also an Italian media mogul, is one well-known example of this phenomenon.

Government may also control media indirectly through ownership of media infrastructure. For instance, Romania's only newsprint mill was state owned for years following the end of Romanian communism. Similarly, the Associated Press of Pakistan is owned by the Pakistani government. In other countries, government exercises indirect control over media outlets whose financial positions depend on state-supplied income, such as revenue from government advertising.

Another important source of indirect government control of media is regulation of the media industry. Many governments require licenses for newspapers, television stations, and even journalists to operate and may use this power to restrict entry into the media industry to individuals who are friendly to the government and/or use the threat of license revocation to silence media critics.

Where government exerts significant influence over media and thus media are unfree, media's usefulness as a mechanism for assisting citizens to solve the agency problems they face with respect to their political officials is seriously impaired. Rather than monitoring political actors' behavior and reporting accurately on their uses of authority, media are likely to avoid furnishing citizens with such information or, worse still, furnish citizens with misleading information that benefit those in power. Uninformed or misinformed citizens find it difficult to use media-provided information to hold political actors accountable or to coordinate appropriate responses to such actors' behavior. Media's ability to effectively control government thus hinges critically on its freedom from government.

Media's Influence on Politics

A strong link exists between media and citizen knowledge. Citizens who are exposed to more media coverage of their local politics and who live in regions where media are freer are more politically knowledgeable than citizens who enjoy less media coverage of their local politics and who live in regions where media are less free. For example, in Eastern Europe, citizens who live in countries where media are freer are more likely to correctly answer basic questions about political representation in the EU, and in the United States, citizens who are exposed to more media coverage of their local politics are more likely to know their congressman's name (Leeson 2008; Snyder and Stromberg 2010).

A strong link also exists between media and political-economic outcomes. Across countries, freer media are associated with higher voter turnout and other forms of political participation (Leeson 2008). Freer media are also associated with higher income, more democracy, more education, and more market-oriented economic policies (Djankov et al. 2003). In the United States, congressmen whose districts receive better media coverage are more likely to vote in a manner consistent with their party's line, stand witness before congressional committees, and procure more spending for their districts (Snyder and Stromberg 2010).

Media's Role in Developing/Transition Economies

In developing and transition economies, media freedom is especially critical for political-economic development. Here, the problem of dysfunctional and corrupt government is pronounced, rendering media as a mechanism for controlling such malfeasance – a mechanism that, as indicated above, requires media independence – of particular importance.

Peru provides a striking example of how even a small amount of media independence can have a large effect on political-economic outcomes in the developing world (McMillan and Zoido 2004).

Despite the Fujimori government's bribe-secured control of all major media outlets in the country in the late 1990s, a small independent station that had not been bribed, Channel N, managed to acquire a recording of a high-ranking government official bribing an opposition politician to switch parties. Channel N repeatedly broadcast the video, and following suit, other channels began doing so too, eventually generating a popular backlash against the Fujimori government and the downfall of the corrupt Fujimori administration. In this case, the existence of only a single free media outlet proved critical to exposing political malfeasance and catalyzing the removal of a self-serving government.

In Russia, in contrast, a dearth of media freedom has prevented media from controlling government. Under the Soviet regime, government completely controlled Russian media, which served as little more than tools for government propaganda. During glasnost and perestroika, laws guaranteeing media independence were passed, and Russia's media appeared to become significantly freer. Ties between Russian media and political elites were never completely severed, however, and Russian media independence suffered major blows during the economic downturn of the early 1990s.

During this period, Russian media circulation and revenue plummeted. In consequence, many media outlets became dependent on state subsidies. Today, the Russian government commonly interferes with freedom of the press (on an important exception, see Enikolopov et al. 2011). Owners of media outlets that are critical of the government are threatened with jail time or forced to sell, and journalists who are critical of the state have been intimidated and even killed (Zassoursky 2004). Government's influence on media in Russia has contributed to Russian political actors' ability to wield public office for private gain.

References

- Besley T, Burgess R (2001) Political agency, government responsiveness and the role of media. *Eur Econ Rev* 45:629–640
- Besley T, Burgess R (2002) The political economy of government responsiveness: theory and evidence from India. *Q J Econ* 117:1415–1451
- Coyne CJ, Leeson PT (2004) Read all about It! Understanding the role of media in economic development. *Kyklos* 57:21–44
- Coyne CJ, Leeson PT (2009a) Media, development, and institutional change. Edward Elgar, Northampton
- Coyne CJ, Leeson PT (2009b) Media as a mechanism of institutional change and reinforcement. *Kyklos* 62:1–14
- Djankov S, McLiesh C, Nenova T, Shleifer A (2003) Who owns the media? *J Law Econ* 46:341–382
- Enikolopov R, Petrova M, Zhuravskaya E (2011) Media and political persuasion: evidence from Russia. *Am Econ Rev* 101:3253–3285
- Leeson PT (2008) Media freedom, political knowledge, and participation. *J Econ Perspect* 22:155–169
- Leeson PT, Coyne CJ (2005) Manipulating the media. *Inst Econ Dev* 1:67–92
- Leeson PT, Coyne CJ (2007) The reformers' dilemma: media, policy ownership, and reform. *Eur J Law Econ* 23:237–250
- McMillan J, Zoido P (2004) How to subvert democracy: Montesinos in Peru. *J Econ Perspect* 18:69–92
- Snyder JM, Stromberg D (2010) Press coverage and political accountability. *J Polit Econ* 118:355–408
- Zassoursky I (2004) Media and power in post-Soviet Russia. M.E. Sharpe, New York

Medical Experimentation

► Human Experimentation

Medical Liability

Ben C. J. van Velthoven and
Peter W. van Wijck
Leiden Law School, Leiden University,
Leiden, The Netherlands

Abstract

This article concerns the preventive effects of medical liability. On theoretical grounds it is argued that medical liability does not necessarily lead to a socially optimal level of precaution, because the incentives are distorted in various ways. Since the 1970s, US states have enacted a variety of reforms in their tort

systems. This variation has provided highly useful data for empirical studies of medical liability issues. For one thing, it has become clear that only some 2% of the patients with negligent injuries gets compensation. The empirical evidence nevertheless suggests that medical liability pressure does affect the behavior of healthcare providers to some degree. It has a negative effect on the supply of services, and it encourages the ordering of extra diagnostic tests. At the margin, medical liability law does seem to have some social benefits that offset reasonable estimates of overhead and defensive medicine costs.

Definition

Medical liability essentially is tort law applied to healthcare providers. If negligent behavior of a healthcare provider causes harm to a patient, the healthcare provider has to pay damages to the patient. In this way, medical liability may lead to compensation of harm. Medical liability may also influence incentives to take care and consequently influence the probability and the size of harm. In the Law and Economics literature, the focus is on this preventive function.

Introduction

Medical treatment is supposed to improve the patient's health. But there is no guarantee. The patient's condition, bad luck, and medical errors may stand in the way of recovery. As adverse events occur within a market-type relationship, physicians and patients could write a complete contract to lay down the mutual understandings with respect to the specifics of the treatment and set the price of the contract accordingly. But this manner of controlling the level of care fails because of asymmetric information (Arlen 2013). Patients just do not know whether a physician delivers medically appropriate care. Moreover, individual bargaining about the level of care and the corresponding price brings along huge transaction costs and is even totally impossible in the

case of emergencies. Fortunately, society has several other institutions available for the control of medical care. The government can centralize control by installing a regulatory body to enforce safety standards. State licensing, disciplinary boards, and hospital credential committees may also motivate physicians to act with proper care. Here we focus on medical liability (see also Van Velthoven and Van Wijck 2012).

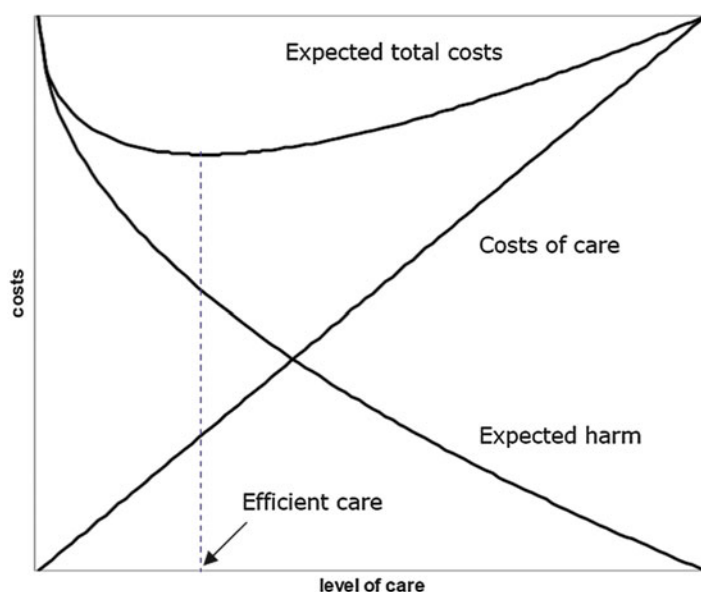
Medical liability has received much attention in the Law and Economics literature in the past decades. The main reason is that since the 1970s, the USA has experienced three medical malpractice crises, periods characterized by significant increases in the premiums and contractions in the supply of malpractice insurance. In response to these crises, US states have enacted a variety of reforms in their tort systems. As a result, the USA has seen a considerable variation, across time and space, in the pressure of the medical liability system on health care providers. That variation has provided highly useful data for empirical studies into the actual working of the tort law system. For the same reason, most of those studies focus on the USA.

The Standard Tort Model Applied to Medical Liability

In the economic analysis of tort law, a fundamental distinction is made between unilateral and bilateral accidents (Shavell 2004). Medical injuries can generally be taken to be *unilateral accidents*. A physician who wants to reduce the probability and severity of medical injury can increase the number of visits provided to his patient, perform additional diagnostic tests, refer the patient to a specialist, opt for more or less invasive procedures, and/or take more care in performing surgery. The patient is usually unable to influence expected harm from a medical injury.

Figure 1 presents the standard tort model for a unilateral accident case (Miceli 2004, pp. 42–45). As the (potential) injurer raises the level of care by taking additional precautionary measures, his *costs of care* increase. But at the same time, there is a reduction in the *expected harm* for the

Medical Liability,
Fig. 1 Efficient care



(potential) victim, as additional care may reduce the probability and/or the severity of accident losses. The social optimum is obtained if the expected total costs are at a minimum. The socially optimal level of precaution is frequently referred to as the *efficient level of care*.

Medical liability law quite universally holds an injurer liable for accident losses that are attributable to *negligence*. The negligence rule presupposes a norm of *due care*, specified by statutory law or jurisprudence, for the precautionary measures that the injurer should take at a minimum. If the injurer's level of care falls short of this minimum, the injurer is negligent and will be held liable for accident losses. On the other hand, if the injurer's level of care equals or exceeds the due care norm, accident losses will remain with the victim. In this way, the negligence rule creates a discontinuity in the injurer's expected costs at the level of due care, as shown by the fat curve in Fig. 2. The injurer minimizes his costs by just taking precautionary measures in conformity with the due care norm.

Whether the injurer will act in a socially optimal manner then depends on the proper choice of the due care norm. He will take socially optimal precaution if the level of due care coincides with

the efficient level of care, as in Fig. 2. If the due care norm is set below (above) the efficient level of care, the personal incentives will generally lead the injurer to behave in a suboptimal manner by taking too little (too much) precaution.

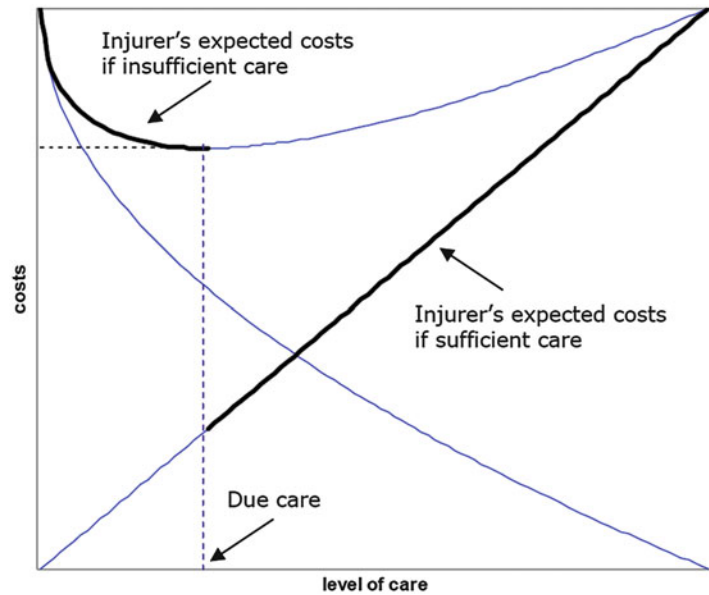
Apart from the level of care chosen by the (potential) injurer when engaging in a certain activity, total costs for society also depend on his *number of activities*. Under negligence, when the injurer conforms to the due care norm, he cannot be held liable for any accident losses. Consequently, a non-negligent injurer is not confronted with the full social costs of his activity. This negative externality may lead the injurer to undertake too many activities from a social point of view.

Problems in Applying the Standard Tort Model to Medical Liability

In practice, a number of specific characteristics of medical care pose serious problems to a straightforward application of the standard tort model.

Uncertainty About Due Care

The standard tort model suggests that injurers can be induced to provide socially optimal care if law

Medical Liability,**Fig. 2** Efficient care under the negligence rule

sets due care at the efficient level. This simple rule, however, is not easily met in the health care sector:

- It is difficult to determine the efficient level of care, even for specialists in the field. Instead, the courts generally evaluate the conduct of physicians in terms of customary standards of practice within the medical profession or a significant minority of such professionals (Weiler et al. 1993, p. 8; Danzon 2000, p. 1343; but see Peters (2002) for a somewhat different reading).
- Physicians have no exact insight into the due care norm that will be applied by the courts when confronted with a claim.
- In each specific case, the court has to decide whether the physician has been careful enough. Even if the due care norm is right, the court may err in its decision, as the information presented to the court will generally be incomplete and subject to mistakes in interpretation. Consequently, physicians do not know exactly how careful they have to be in order to escape liability.

Calfee and Craswell (1984) have studied the consequences. If there is uncertainty about the due care norm and its application in court, there is a

chance that a physician who has taken sufficient care may still be held liable for damages. The physician can try to reduce that chance by *over-complying*, that is, by raising his level of care beyond due care.

Ethical Norms

Most physicians swear (some variant of) the Hippocratic oath to use their best ability and judgment in treating their patients. This may affect the ethical values in the profession in such a profound manner that all other (financial) considerations are rendered more or less futile. If so, the result would be overcompliance, that is, delivering more care than strictly necessary.

Who Bears the Costs of Care?

The standard tort model starts from the premise that the costs of care are borne by the injurer. This is, however, not self-evident in the health care sector, as physicians are generally paid on a fee-for-service basis. Moreover, most patients carry a health insurance policy. When a physician decides to raise the level of care, he will not have much trouble in charging the additional costs to his patient, who will forward the bill to his insurance company.

Of course, this financial incentive to provide too much care will be mitigated if healthcare

services are financed in a different way, which hinders or prevents the passing on of additional care costs. One can think of: pure salary payment, capitation payment, or managed care plans (Glied 2000).

Litigation Problems

The standard tort model takes it for granted that the injurer pays damages, once he is found negligent. In general, however, this payment will not be made spontaneously. First, the victim has to decide whether it is in his interest to file a claim. The patient may be unaware of the negligence of his physician, the decision of the court may be too uncertain, the litigation costs may be too high, or the financial means of the victim may be insufficient. Second, if the claim is below a threshold of, say, \$250,000, the patient may not be able to find an attorney willing to accept his case on a contingent fee basis (Shepherd 2014). Third, even if a claim is filed, parties may decide to settle the case in order to save on litigation costs. If the case is settled, damages will not amount to a full compensation.

The implication of all this is that expected damage payments are smaller than the expected harm of the injurer's behavior. The result is a tendency toward insufficient care.

Liability Insurance

Most, if not all physicians, carry a liability insurance policy (GAO 2003, p. 6). By shifting the burden of the damages, liability insurance lowers the injurer's costs of insufficient care (Zeiler et al. 2007).

In theory, experience rating can redress this tendency. Experience rating refers to a variety of schemes that see to it that liability insurance premiums reflect each insured's expected loss (Danzon 2000, p. 1160). This is mostly done by varying premiums with past claims or loss experience. Shavell (1982) has shown that insurance need not interfere with the incentive effects of liability if premiums are perfectly experience rated. But experience rating is more easily said than done in the case of medical care. Medical malpractice claims occur too infrequently to give insurance companies enough information to

reliably set premiums in accordance with individual physician's care levels.

Medical liability insurance, however, does not completely eliminate incentives to take care. Malpractice may affect the physician severely, even if he is fully insured against the financial consequences, as claims also bring along other kinds of costs. The defense takes quite a lot of precious time, the experience is rather unpleasant, and it may cause serious reputational harm (Weiler et al. 1993, p. 126).

Patient Safety Investments

In the standard tort model, the individual physician makes a more or less informed decision on how to treat each particular patient. But that model does not fully capture the causes of medical negligence (Arlen 2013). Many medical injuries occur accidentally when physicians unknowingly err in the diagnosis and the selection of treatment or when the hospital equipment and staff fail in delivering the treatment. The probability of such errors can be reduced by investing in expertise, healthcare technology, and personnel. But such patient safety investments generally fall outside the factors courts consider in determining whether a physician has been negligent. The incentives for the physician, and the hospital, to invest in patient safety are therefore suboptimal.

Conclusion

Medical liability does not necessarily lead to a socially optimal level of precaution, because the incentives are distorted in various ways. For one thing, physicians generally do not bear the full accident losses of insufficient care. This distortion may act as an invitation to physicians to take less care than legally required. Still, the nonfinancial consequences of liability (time, hassle, reputation loss) and ethical considerations may provide some counterweight. Other distortions provide incentives to act on the safe side of the due care norm. Physicians generally do not bear the (full) costs of care due to specific methods of financing in the healthcare sector. And there is uncertainty about the due care norm and its application by the courts. On balance, there might be a bias toward excessive care.

Defensive Medicine

In the USA, the conviction has taken root among physicians and their liability insurers that the medical liability system has gone wrong. It is argued that the pressure has evolved to such a level that it has given rise to *defensive medicine*. The most common definition (OTA 1994, p. 21) reads: “Defensive medicine occurs when doctors order tests, procedures, or visits, or avoid high-risk patients or procedures, primarily (but not necessarily solely) to reduce their exposure to malpractice liability.” According to this definition, defensive medicine can take two forms. *Positive* defensive medicine involves supplying care that is not cost effective, unproductive, or even harmful. *Negative* defensive medicine involves declining patients that might benefit from care. It also includes physicians deciding to exit the profession altogether.

Our discussion of the standard tort model demonstrates that positive defensive medicine will not necessarily be found in practice, as liability pressure on the level of care is working in two opposite directions. Thus, the question whether physicians take excessive care is really an *empirical* question. Second, if malpractice pressure does produce a bias toward excessive care, it is excessive in comparison to the *due care* norm. But it is not at all certain that the due care norm has been set equal by law to the efficient level of care. That leaves the possibility, even if empirical research finds proof of excessive care, that level of care still falls short of the socially optimal amount (Sloan and Shadle 2009, p. 481).

The concept of negative defensive medicine is related to the number of activities, referred to earlier. If a physician takes at least due care, he will not be liable for any accident losses, which might give him the incentive to accept too many patients from a social point of view and/or to stay too long in the profession. On the other hand, the simple fear of malpractice claims, even if unwarranted, and the corresponding threat of time and reputation loss may work in the opposite direction. Hence, the existence and scope of negative defensive medicine is, once again, an

empirical question. Moreover, note that the interpretation of the findings may change if physicians exercise insufficient care. Then, malpractice law helping patients to file claims and to obtain damage payments may give negligent physicians a good reason to revise their conduct, not only by raising the level of care but also by accepting fewer patients or by early retiring. Such a behavioral response might be very welcome from a social point of view.

What makes the interpretation of the findings even more complicated is the interaction between the defensive medicine that may follow from medical liability pressure and the *offensive medicine* that is induced by physicians’ financial incentives (Avraham and Schanzenbach 2015). When physicians have the discretion to choose among different treatment regimens and health insurance adequately covers their patients’ medical costs, physicians may be tempted to opt for the more invasive procedures, which in general will be the more remunerative ones. But more invasive procedures are also riskier. Hence, medical liability pressure may counteract the tendency toward offensive medicine.

Medical Liability Litigation

This section surveys the empirical evidence concerning medical liability litigation. The different layers of the dispute pyramid (Galanter 1996) are discussed one by one.

Three large-scale surveys of medical records of hospitalized patients in the USA have investigated the incidence of injury due to negligent medical care. In the most recent survey in Utah and Colorado in 1992 (Studdert et al. 2000), it was found that 2.9% of all hospitalized patients had an adverse event that was related to medical care. Some 0.8% of the patients suffered a negligent injury, where negligence was defined as treatment that failed to meet the standard of the average medical practitioner. No attempt was made to define negligence by weighing marginal costs and benefits of additional precautions. So, the resulting count of negligent injuries does not necessarily correspond to inefficient injuries.

The second layer of the dispute pyramid discloses how many of the injury victims take steps to obtain compensation. In the Utah and Colorado study, only 3% of the patients who were identified as having sustained a negligent injury filed a malpractice claim. But there was also a significant number of “false positives.” Aggregate data from insurers’ records pointed out that a substantial number of malpractice claims do not correspond to an identifiable injury due to negligent medical behavior. Of course, all these plaintiffs may still have filed the claims in good faith, from a state of imperfect information, leaving it to the tort system to separate the rightful claims from the non-deserving ones.

The third layer of the dispute pyramid discloses how filed claims fare in the tort system. In a large study of malpractice claims closed between 1984 and 2004, Studdert et al. (2006) found that 61% of claims could be associated with injury due to medical error, while 39% of the claims had no merit. Only 15% of all the claims were resolved by trial verdict; the rest was settled in the “shadow of the law” or dropped. Most of the claims involving injuries due to medical error (73%) received compensation; most claims not involving medical error (72%) did not receive compensation. Moreover, when claims involving error were compensated, payments were significantly higher on average than were payments for non-error claims.

With respect to the payment amounts, two observations are in place. First, compensation in most cases falls short of plaintiff’s losses, especially for more serious injuries (Sloan and Chepke 2008). Second, the costs of administering the tort system (legal expenses, overhead costs) are considerable. According to calculations by Mello et al. (2010), it costs US society overall more than \$1.70 to deliver \$1 of net compensation.

The tort litigation system is not perfect, then. It sometimes makes physicians – or their insurers – pay damages for non-negligent care. But the system is clearly not a random lottery (see also Eisenberg 2013). Negligent injuries are at least ten times as likely to end up in compensatory payments as non-negligent injuries. As a result, there is a strong association between the

numbers of adverse patient safety events in hospitals and paid medical malpractice claims (Black et al. 2017).

More disturbing for the proper working of the system is the high rate of “false negatives.” From the figures above, it follows that just some 2% of the patients with negligent injuries gets compensation, mainly because a large fraction of valid claims is not filed, but also because not all valid claims that are filed get honored. And even that 2% is quite likely a serious overestimation, as an observational study of healthcare providers has shown that hospital records may miss over 75% of serious medical errors (Andrews 2005). Combining the high rate of false negatives with the finding that compensation generally falls short of victims’ losses suggests that the deterrent function of the system must be rather limited.

Tort Reform

In the introduction, it was noted that the USA experienced three “crises” in the medical liability insurance market in the past decades. These were periods of deterioration in the financial health of carriers, followed by sharp increases in premiums and contractions in supply. This is not the place to delve deeply in the causes of these crises (cf. Danzon 2000; Sloan and Chepke 2008). But one factor can be singled out: the “long-tail” character of this line of insurance. Claims may be filed many years after an adverse event causes injury. And from there, it may take many more years before the insurance company finally knows how much compensation it has to pay. If, for whatever reason, there is a gradual rise in claim frequency and/or in average payments, for instance, because of pro-plaintiff adaptations in common law doctrines or because patients are becoming more assertive toward healthcare professionals, insurance companies will tend to lag behind. They will develop unexpected losses, and over-react in raising premiums and curtailing supply.

In response to the malpractice crises, most US states have adopted tort reform measures. The objective of these measures is to reduce the

overall costs of medical liability. The extent and specifics of tort reform vary from state to state (cf. www.atra.org). Some reforms aim at a reduction of damage awards, other reforms make it more costly or difficult to file tort cases in the first place. The most commonly adopted tort reforms are: caps on non-economic damages, pretrial screening panels, contingency fee reform, joint-and-several liability reform, collateral source rule reform, periodic payment, and shorter statutes of limitation. The effects of these tort reforms on the frequency and the size of claims and on malpractice insurance premiums have been studied extensively. A detailed review by Mello and Kachalia (2016) concludes that there is no convincing evidence that any other reform than caps on non-economic damages has had a significant impact. As to damage caps: the weight of the evidence suggests that they reduce claims frequency, achieve substantial savings in average damage payments, and modestly constrain the growth of malpractice insurance premiums. So one would be tempted to conclude that at least this specific kind of tort reform can help to relieve malpractice pressure, if so desired. But even that conclusion is called into question. Zeiler and Hardcastle (2013) point out that thus far no one study has employed a consistently solid set of empirical research methods.

More recently, also other kinds of reform measures have been proposed, such as apology laws and disclosure programs, presuit notification periods, health courts, and safe harbors for adherence to evidence-based practice guidelines. As these reforms are relatively new in use, if at all, the empirical literature on their effects is as yet very small.

Preventive Effects of Medical Liability

It is an empirical question whether medical liability leads physicians to take appropriate precautions or to engage in defensive medicine. In the literature, four main research lines can be distinguished.

The first line of research surveys physicians and asks their opinion on the role of malpractice

pressure in clinical practice. These studies (e.g., Carrier et al. 2010) unequivocally point out that concerns about malpractice liability are pervasive among physicians. Indeed, in a survey of high-risk specialists, 93% of the interviewees reported practicing defensive medicine (Studdert et al. 2005). Yet, the results should be handled with caution. First, the relationship between perceived malpractice threat and objective liability risk is found to be very modest. Physicians systematically overestimate the risk that malpractice action will be brought against them. Second, the relationship between malpractice threat and clinical response is a self-reported one.

The second line is about the actual relevance of positive defensive medicine. How do treatment choices by physicians, and the health outcomes of their patients, respond to malpractice pressure? Much attention has gone to obstetrics, the field that has one of the highest levels of premiums, claim frequency, and damage payments. Typically, studies examine the impact of tort reform on cesarean section rates. Some studies have also looked at the impact on infant health at birth. Overall, the results are inconclusive. Thus far, there is no decisive evidence for positive defensive medicine in obstetrics (Eisenberg 2013). Some other studies focus on cardiac illness or take a look at broader sets of ailments or total healthcare expenditures. The results are mixed. As far as physicians are found to practice positive defensive medicine, the excessive care appears to be related to rather elementary diagnostic tests such as imaging, not to major surgical procedures. The overall picture is that the total effect on healthcare costs, if any, is rather small (Thomas et al. 2010).

The third line of research is on negative defensive medicine and analyzes how tort reform affects the supply of healthcare services. The evidence with respect to obstetrics is mixed. Other studies analyze the overall supply of physician services. Their results generally point out that higher malpractice pressure tends to diminish healthcare supply, be it the number of physicians, statewide or in local areas only, or their hours worked. That finding seems to be proof of negative defensive medicine. But note that the

interpretation is not so obvious. A smaller supply of physicians in itself can be presumed to contribute negatively to social welfare, but there may also be offsetting effects if the quality of the physicians that stop or reduce their practice is below average. Indeed, Dubay et al. (2001) and Klick and Stratmann (2007) find no evidence that the changes in supply had negative health effects.

Finally, an interesting new line of research tries to assess the preventive impact of medical liability forces by drawing on variations in the negligence standard facing physicians. The majority of US states have over time moved from a due care norm based on the customary practices of local physicians to a national standard of care. The first empirical results (Frakes and Jena 2016) indicate that treatment quality improves when the clinical standards go up.

Cost-Benefit Analysis

Empirical evidence suggests that medical liability pressure does affect the behavior of healthcare providers. It does seem to encourage the ordering of extra diagnostic tests, and it tends to reduce the supply of services. However, positive defensive medicine does not have a clear-cut effect on health. If the additional tests and procedures have any value, it is only a marginal one. Furthermore, changes in the supply of services do not affect health adversely. This suggests that the physicians that are driven out of business have a below average quality of performance. Hence, at the margin, medical liability law may have some social benefits (see also Zabinski and Black 2015).

These benefits must be weighed against the costs of the additional tests and procedures. The costs of administering malpractice claims also deserve attention. Both Danzon (2000) and Lakdawalla and Seabury (2012) have made a shot at a back-of-the-envelope calculation of the costs and benefits. They conclude that under quite general assumptions, the benefits of even a modest reduction in injury rates suffice to offset reasonable estimates of overhead and defensive medicine costs. This follows from the large social costs

of medical injuries and the low rate of claims per negligent injury.

Yet, instructive as these calculations may be, they mainly have a heuristic value. First, a full cost-benefit evaluation of the medical liability system is impossible in the current state of affairs. Second, even if the marginal benefits of the current system do outweigh the costs, the search for improvements and alternatives is open (see, e.g., Sloan and Chepke 2008). It is argued that the impact of the medical liability system can be substantially improved by shifting liability from the individual physician to the medical entity involved (Arlen 2013) and by restructuring the financial incentives in the healthcare sector (Frakes 2015).

Cross-References

- ▶ [Litigation and Legal Expenses Insurance](#)
- ▶ [Mass Tort Litigation: Asbestos](#)
- ▶ [Medical Malpractice](#)
- ▶ [Strict Liability Versus Negligence](#)
- ▶ [Tort Damages](#)

References

- Andrews L (2005) Studying medical error in situ: implications for malpractice law and policy. *DePaul Law Rev* 54:357–392
- Arlen J (2013) Economic analysis of medical malpractice liability and its reform. In: Arlen J (ed) *Research handbook on the economics of torts*. Edward Elgar, Cheltenham, pp 33–68
- Avraham R, Schanzenbach M (2015) The impact of tort reform on intensity of treatment: evidence from heart patients. *J Health Econ* 39:273–288
- Black BS, Wagner AR, Zabinski Z (2017) The association between patient safety indicators and medical malpractice risk: evidence from Florida and Texas. *Am J Health Econ* 3:109–139
- Calfee JE, Craswell R (1984) Some effects of uncertainty on compliance with legal standards. *Virginia Law Rev* 70:965–1003
- Carrier ER et al (2010) Physicians' fears of malpractice lawsuits are not assuaged by tort reforms. *Health Aff* 29:1585–1592
- Danzon PM (2000) Liability for medical malpractice. In: Culyer AJ, Newhouse JP (eds) *Handbook of health economics*, vol 1B. North-Holland, Amsterdam, pp 1339–1404

- Dubay L et al (2001) Medical malpractice liability and its effect on prenatal care utilization and infant health. *J Health Econ* 20:591–611
- Eisenberg T (2013) The empirical effects of tort reform. In: Arlen J (ed) *Research handbook on the economics of torts*. Edward Elgar, Cheltenham, pp 513–550
- Frakes MD (2015) The surprising relevance of medical malpractice law. *Univ Chicago Law Rev* 82: 317–391
- Frakes M, Jena AB (2016) Does medical malpractice law improve health care quality? *J Public Econ* 143: 142–158
- Galanter M (1996) Real world torts: an antidote to anecdote. *Md Law Rev* 55:1093–1160
- GAO (2003) Medical malpractice insurance. Multiple factors have contributed to increased premium rates. US General Accounting Office, Washington, DC. Report GAO-03-702
- Glied S (2000) Managed care. In: Culyer AJ, Newhouse JP (eds) *Handbook of health economics*, vol 1A. North Holland, Amsterdam, pp 707–753
- Klick J, Stratmann T (2007) Medical malpractice reform and physicians in high-risk specialties. *J Leg Stud* 36: S121–S142
- Lakdawalla DN, Seabury SA (2012) The welfare effects of medical malpractice liability. *Int Rev Law Econ* 32:356–369
- Mello MM, Kachalia A (2016) Medical malpractice: evidence on reform alternatives and claims involving elderly patients. A report prepared for the Medicare Payment Advisory Commission. MedPAC, Washington, DC
- Mello MM et al (2010) National costs of the medical liability system. *Health Aff* 29:1569–1577
- Miceli TJ (2004) *The economic approach to law*. Stanford University Press, Stanford
- OTA (1994) Defensive medicine and medical malpractice. US Congress, Office of Technology Assessment, Washington, DC. Report OTA-H-602
- Peters PG Jr (2002) The role of the jury in modern malpractice law. *Iowa Law Rev* 87:909–969
- Shavell S (1982) On liability and insurance. *Bell J Econ* 13:120–132
- Shavell S (2004) *Foundations of economic analysis of law*. Harvard University Press, Cambridge, MA
- Shepherd J (2014) Uncovering the silent victims of the American medical liability system. *Vanderbilt Law Rev* 67:151–195
- Sloan FA, Chepke LM (2008) *Medical malpractice*. MIT Press, Cambridge, MA
- Sloan FA, Shadle JH (2009) Is there empirical evidence for “defensive medicine”? A reassessment. *J Health Econ* 28:481–491
- Studdert DM et al (2000) Negligent care and malpractice claiming behavior in Utah and Colorado. *Med Care* 38:250–260
- Studdert DM et al (2005) Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *JAMA* 293:2609–2617
- Studdert DM et al (2006) Claims, errors, and compensation payments in medical malpractice litigation. *New Engl J Med* 354:2024–2033
- Thomas JW, Ziller EC, Thayer DA (2010) Low costs of defensive medicine, small savings from tort reform. *Health Aff* 29:1578–1584
- Van Velthoven BCJ, Van Wijck PW (2012) Medical liability: do doctors care? *Recht der Werkelijkheid* 33(2):28–47
- Weiler PC et al (1993) *A measure of malpractice: medical injury, malpractice litigation and patient compensation*. Harvard University Press, Cambridge, MA
- Zabinski Z, Black BS (2015) The deterrent effect of tort law: evidence from medical malpractice reform. Northwestern University Law School, Chicago, IL. Law and Economics Research Paper No. 13-09
- Zeiler K, Hardcastle L (2013) Do damages caps reduce medical malpractice insurance premiums? A systematic review of estimates and the methods used to produce them. In: Arlen J (ed) *Research handbook on the economics of torts*. Edward Elgar, Cheltenham, pp 551–587
- Zeiler K et al (2007) Physicians’ insurance limits and malpractice payments: evidence from Texas closed claims, 1990–2003. *J Leg Stud* 36:S9–S45

Medical Malpractice

Veronica Grembi

Copenhagen Business School, Copenhagen, Denmark

Abstract

MM first came to the attention of policy makers primarily in the USA where, from the 1970s, healthcare providers denounced problems in getting insurance for medical liability, pointing out to a *crisis* in the MM insurance market (Sage WM (2003) *Understanding the first malpractice crisis of the 21st century*. In: Gosfield AG, (ed) *Health law handbook*. West Group, St. Paul, pp 549–608). The crisis was allegedly grounded in an explosion of requests of compensations based on suffering iatrogenic injuries. Since then, MM problems have been identified with scarce availability of insurance coverage and/or its affordability, the withdrawal from the MM insurance of commercial insurers, the growth of MM public insurance or self-insurance solutions, the choice of no-fault

rather than negligence liability, the adoption of enterprise liability for hospitals, the concerns for defensive medicine, and the implementation of tort reforms so to decrease MM pressure (i.e., frequency of claims and the levels of their compensation) on healthcare practitioners. While the initial contributions to the topic are mainly based on the US healthcare and legal system experience, a growing attention to these problems has raised in the last decades also among European countries (Hospitals of the European Union (HOPE) (2004) Insurance and malpractice, final report. Brussels, www.hope.be; OECD (2006) Medical malpractice, insurance and coverage options, policy issues in insurance n.11; EC (European Commission, D.G. Sanco) (2006) Special eurobarometer medical errors).

Definition

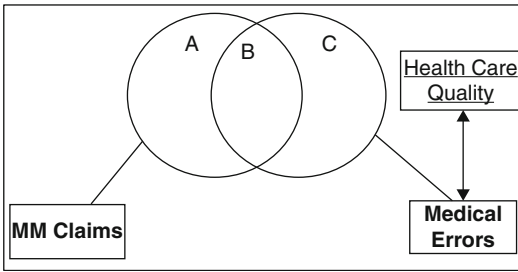
Medical malpractice (MM) deals with two kinds of problems that healthcare providers, patients, and insurance companies have to deal with in different legal and healthcare systems: (1) medical errors without legal consequences and (2) medical errors with legal consequences. The contributions to this field aim at defining an efficient level of (1) and to efficiently deter and compensate (2). Empirical evidence on the actual trends of medical errors, medical liability claims, and the link of the former and the latter with both the trend of MM insurance premium and the treatment decisions of healthcare providers are not always unambiguous.

Medical Malpractice

Starting from the 1970s, healthcare practitioners, lawyers, and insurance companies in the USA experienced so-called crises of MM. The characteristics of the first crisis, joint to those which followed during the 1980s and the beginning of the New Millennium, could be summarized as follows (Mello et al. 2003): first, an increase in MM claims not always justifiable by a similar increase in medical errors; second, an increase in

MM premium for healthcare providers joint to the scarcity of available insurers in the market; and third, the tendency of healthcare practitioners to adopt behaviors which can minimize the probability to be targeted by a legal claim but not the probability of an error or bad quality medicine, also known as defensive medicine. All in all, there is an increase in legal, insurance, and healthcare expenditures, which have consequences for the quality of the delivered healthcare (Arlen 2013). Similar concerns and complaints have started to be common also among several European countries, notwithstanding their quite different healthcare and legal systems compared to the USA. The contribution of the economic analysis of law to this subject is at first a theoretical analysis of the role of legal rules as incentivizing mechanisms to achieve efficiency across different institutional settings. As such, the law and economics of MM provides a common ground of theoretical elements, which, in time, have been accompanied by the production and discussion of often controversial empirical evidence.

From a theoretical viewpoint, MM has to do with the definition of the *efficient* medical error as it has been shaped by the contributions of the economic analysis of tort law (Shavell 1987; Arlen 2013). According to this approach, the goal of legal rules is not to reduce the probability to have errors to zero, but to deter *inefficient* errors and compensate the innocent victim of an *inefficient* error. Suppose that the probability that an error i takes place is equal to p_i and that in case i happens, victim v will suffer damages equal to D (both economic and noneconomic). The expected damages are equal to $p_i D$. The healthcare provider-potential injurer in this framework can decrease p_i investing in precaution, which will cost her c , with c increasing as p decreases (i.e., $c(p)$). The efficient level of precaution minimizes the sum of the expected damages, $p_i D$, and the cost of precaution, c . For that optimal level of precaution, let us call it x^* , the level of medical errors allowed into the system is considered efficient and expected to be different from zero (i.e., the social optimum). We have tried to summarize this basic intuition in the diagram presented in Fig. 1. Areas B and C represent the



Medical Malpractice, Fig. 1 Relationship between MM claims and medical errors

universe of medical errors within a given institutional setting. Consistently with what we have just stated, legal rules need to define the prerequisites to put an error in either B or C: in an ideal setting, if the legal system is properly designed, an error in C is efficient and an error in B is inefficient, and the dimension of B has to be efficient in a social welfare perspective.

In the same setting, a policy maker can set the boundaries between C and B using either a negligence/fault rule or a strict liability/no-fault rule. In the first case the stress is on the threshold of precaution below which the potential injurer is held liable. As in the case of a car accident liability, a speed threshold is set and if an accident occurs, you will be liable only if your speed was higher than the imposed limit. In the second case (i.e., No Fault) a key role is played by the causation link between the negative outcome and the injurer action. In other words you are responsible if you injure somebody in a car crash, but you are not accountable for injuries that your potential victim got before the crash. Both liability rules find a place in the realm of MM. Both can generate less than or more than efficient precaution in the real world.

The problem of transposing the basic liability model on medical “accidents” is that it is not always clear how precaution is related to expected outcomes. While in the case of a car accident, the sequence is clear: you speed, the car crashes, and a person that before the impact was sound is now injured. In the case of a medical accident, there might be a less clear sequence. A physician does not take the necessary precaution; she treats a sick patient, and the sick patient does not heal.

Unfortunately it is not so simple. It could well be that the sick patient heals or that he/she heals but it takes a week more than expected. In other words, defining precaution in this context is not so straightforward. We could think about appropriate hygienic conditions and basic working environment, but then the real issue at stake is more related to the case of a misdiagnosis or to a mistreatment and to the elements causing the former rather than the latter. But this is a case-by-case call. Hence, the problem is how we define a case-by-case standard of care. For instance, the consensus on precaution levels (or treatment choices) can be high on routinely practiced interventions and low on more complicated procedures. Whenever the standard of care to be adopted is not accurately specified, the system might end up characterized by less or more than efficient standard of care. One consequence is the so-called defensive medicine, a modification in care decisions triggered by MM pressure (i.e., frequency of claims and the levels of their compensation), which can be positive or negative (Danzon 2000; Kessler 2011). Positive defensive medicine consists in the use of treatments or diagnostic tools that are not able to improve the quality of care delivered to patients, but apt to decrease the probability to be targeted by a legal claim. It is a form of supplied induced demand: if the patient has the same information that her physician has, she would have not chosen the recommended care. Negative defensive medicine coincides with forms of cream skinning of patients or procedure. Less risky patients are selected into treatment so to decrease the probability of negative outcomes, and needed risky treatments can be avoided due to the fear of the legal consequences.

A second challenge to the liability model is represented by the organization of the healthcare system. The healthcare organization matters as much as legal rules, to achieve or miss an efficient level of precaution. For instance, a physician could have different incentives to practice defensive medicine depending on whether she is employed by the hospital or she is an independent practitioner. Finally, the possibility to buy insurance for medical liability and the type of insurance

can alter the structure of incentives produced by the legal rules. Insurance is available in a private/commercial form, often not experience rated, or in a public form (However, the distortions generated by the insurance for medical liability should be attenuated by reputational concerns, supposedly more for private rather than public healthcare providers). Forms of public insurance can come together with hospital self-insurance within a public healthcare system, or through a specific public fund. Public coverage can solve some of the problems connected to the insurance market for MM, but they have the potential to generate new kind of problems, such as common pooling of individual risks among the covered healthcare providers, basically incentivizing under-precaution.

In other words the perceived costs/benefits of taking precaution are affected by the certainty of the standard of care, the organization of the healthcare system, and the availability and form of insurance for MM. In principle a fault system with a private MM insurance and no unanimity on the standard of care could trigger positive defensive medicine. Such consequence could be mitigated if the practitioner is not directly responsible, but as an employee of a hospital is covered by an enterprise liability. However, in this case less than efficient precaution could be held. For the same token, a no-fault liability system with private MM insurance could more easily generate negative defensive medicine behaviors. A public insurance scheme could cope with this problem, but again at the expenses of incentive to take precaution. In reality, MM liability comes together with other institutional elements able to affect the structure of incentives foreseen by the economic analysis of law, and this is why defining the best solution according to a social welfare perspective is a debated issue, which needs to be supported by sound empirical work.

Empirical evidence on MM can be grouped in two sets: (1) evidence on the incidence of medical errors, useful to assess the dimension of the iatrogenic injury problem, the relationship between actual errors and claims, and the inner causes of medical errors (Weiler et al. 1993; Nys 2009), and (2) evidence on the effectiveness of

policies directly or indirectly oriented to decrease MM pressure on the healthcare providers as well as changing the structure of costs and benefits of potential claims (i.e., tort reforms) on (a) liability measures such as the frequency of MM claims and medical liability insurance and (b) care-related measures as defensive medicine, the quality of care, or the supply of physicians (Kachalia and Mello 2011; Kessler 2011).

Overall, the empirical assessment of MM is not an easy task. The number of claims filed every year against hospitals and medical practitioners might not necessarily stem from negligent mishaps, while many negligent behaviors and their outcomes are not actually prosecuted (Weiler et al. 1993). Defining for empirical purposes, a medical error or, more properly, an injury due to unacceptable medical negligence is not a straightforward task either: as stated it requires an implicit assumption on the expectations of the outcome of a specific treatment or procedure conditional to many variables describing the conditions of the patient (Danzon 2000). Expectations on a normal distribution of outcomes are not always set before running empirical investigations. The US studies are usually the benchmarks in this field. The first surveys had been run during and in the aftermath of the first malpractice crisis. The 1974 California Study is worth mentioning even though the most famous work is definitely the 1984 Harvard Study, dealing with New York Hospitals data (Weiler et al. 1993). Later on similar initiatives have been undertaken in other countries: the 1995 Australia Study on healthcare quality had been drawn on the Harvard blueprint (Weingart et al. 2000), and a similar approach has been followed in a 1998 study on New Zealand public hospitals (Davis et al. 2002), a 1999–2000 English study (Vincent et al. 2001), a 2004 Dutch study (Zegers et al. 2009), and a 2005 study on Spanish hospitals (Ministero de Sanidad Y Consumo 2006), just to mention a few. Overall they assess a level of incidence of adverse events between 2.7% and 16.6% for the public systems and between 3.7% and 17.7% for the American system (Weiler et al. 1993; Andrews 2005). Around half of them are judged to be

preventable¹. The idea underpinning risk/error management solutions is that bad things do not happen to bad people (Reason 2000); rather “it is weak systems that create the conditions for error” (Department of Health 2002). According to this managerial awareness, the Department of Health in the UK has promoted initiatives to implement the error report system between local and national level. In the aftermath of the 1999 Institute of Medicine Report *To Err is Human: Building a Safer Health System*, many countries felt the urgent need to introduce a “safety culture” within their health systems (Barach and Small 2000). Improve error disclosure, a better interaction among different branches of the same health system, and homogenization of some basic medical procedures are among the undertaken steps. Both insurance companies and medical associations have stirred up such urgency. Additionally, it has been frequently recalled the importance of looking to “high reliability organizations” (i.e., aviation, nuclear power plants) in order to import or shape risk/error management policies that had been widely tested, especially to deal with near misses. Leape and Berwick (2005) undertook a study on the changes in hospital practice to improve safety in the USA 5 years after the publication of *To Err is Human*. Although a national inquiry is still missing, a local level analysis has been characterized by a decrease of adverse drug-related events and infections. Similar investigations (*how* the policy enforcement has effectively reduced *what*) would be desirable in other countries. An accurate analysis of the administrative costs of such policies is missing both at local and national level.).

However, the adopted definition of injury/adverse event, starting with the US studies, has been quite criticized. Identifying an injury as “any (negative) deviation from the expected outcome” and adverse event as “an unintended injury that

was caused by medical management and that result in measurable disability” (Danzon 2000, p. 1352) has been judged either a too broad or a too narrow approach. Some authors, generally not economists, think that such definitions do not allow the inclusion of errors that are not associated to any injury either because the doctors were extremely lucky (and the patients too) or because the doctors could catch the mistake in time (again, a matter of luck) (Weingart et al. 2000). According to this interpretation, the findings should be regarded as the lower bound of a much striking number. Other authors, generally economists, disagree with the previous view, arguing against a loose definition used in the recalled studies and raising the problem of the efficient error, that for which the cost of precaution is equal to the expected damages. In other words, since we have not decided *ex ante* what is the efficient slot of errors related to the practice of a high risk activity as healthcare, we could not state for sure whether those numbers represent efficient or inefficient errors. Consequently, the empirical findings could be viewed as an exaggerated upper bound of a much less sensational phenomenon. Is the “preventable” number of adverse events really representative of inefficient errors? Does “preventable” mean efficiently preventable (precaution costs < potential benefits)? These questions still need to be addressed in practice.

Despite the fact that the two views are clashing, they are extremely useful to get a flavor of the range of factors we should evaluate and weigh when we try to empirically assess the incidence of iatrogenic injuries. A further contribution in this respect comes from a strand of less sound econometric analysis, run sometimes by physicians, which proposes an issue raised by other studies in the field: why injured patients do not always file a claim? Michael Rowe (2004), for example, supports with several case studies the evidence that the probability that a patient will file a suit will decrease whenever he/she has been told about the error. Paradoxically, wards – who are more exposed to the eventuality of error like Emergency Rooms, but where the doctors have a closer and constant supervision of the

¹It might be worthwhile to mention the “risk management” approach on this point. Indeed a central issue is “whether negligent injuries are caused largely by occasional inadvertent lapses of many, normally competent providers or by a minority of incompetent, physicians and low quality hospitals” (Danzon 2000)

patient – tend to get less malpractice suits than other safer wards.

A final concern related to the assessment of medical errors is how we should judge the relationship between errors in medicine and quality of the healthcare service. This is in a way a sort of paradox. The actual improvement in the medical technology and a better knowledge of tackling endemic pathologies, considered fatal just a couple of decades ago, have transformed medical risk in two ways. On the one hand the new technology creates risks that did not exist previously (This belongs to the well-known path of human progress. The automobile invention had caused both development and accidents!), increasing also the skills required for its use. On the other hand the medical science progress makes the “time” factor crucially relevant to diagnose lethal pathology and provides solutions at the early stages of its development (i.e., cancer; Grady 1992). In other words, *ceteris paribus*, an increase in the service quality, at least in terms of adoption of technology, can hold an increase in risk of being suited and in expansion of the object of compensatory claims (For instance, a 2004 decision of the Italian Supreme Court established the right of compensation for iatrogenic *loss of chances* as a juridical independent category. Hence it is possible to file suit both for iatrogenic injury and iatrogenic loss of chances). On this respect a crucial role can be played by liability rules, which can or cannot favor the decision of adopting new technology.

The second strand of empirical literature is not always unambiguous as well and mainly concerns the efficacy of the policies adopted to decrease MM pressure. Decrease the pressure in this context means affecting the probability to be the target of a claim (the physician/hospital), the probability to compensate a claim (insurers), and the probability to get compensation (patient). The main target of these policies is to relieve healthcare practitioners and insurance companies from MM pressure so to decrease problems such as defensive medicine or the availability and affordability of insurance coverage. These goals should be pursued while the quality of the healthcare system is preserved or, at best,

enhanced. The policies at stake are twofold: (1) reforms in the realm of tort law, not designed specifically to tackle MM problems, but that can directly or indirectly affect the behavior of the parties struggling with MM problems, and (2) reforms affecting the organization of the MM insurance market.

A first group of policies/reforms related to tort law has been the most studied in the USA from an empirical perspective. These policies have been distinguished in *direct* (i.e., caps on damage compensations) and *indirect* (i.e., pretrial screening panels) (Kachalia and Mello 2011; Kessler 2011). Although the evidence on the final impact of these policies on both liability and care-related measures is mixed, direct reforms seem to be associated to a stronger effect and in particular caps on damages are regarded among the most effective adoptable measures to decrease MM pressure².

Besides these two groups, tort reforms might include also the shift from negligence to no-fault liability and the adoption of form of enterprise liability. These reforms, differently from the previous, have often been adopted with specific reference to MM, as in the case of the UK (i.e., enterprise liability of hospitals; Fenn et al. 2004), Virginia, and Florida (i.e., no-fault system for severe birth-related neurological damages), as well as the Scandinavian countries (i.e., no-fault liability system for every medical injury). In the European cases the change in the liability regime was accompanied by a change in the MM insurance structure. It coincided with a change from a private/commercial to a public-taxpayers'-paid insurance coverage. The two shifts not always coincided. For example, Denmark adopted left the fault liability system for a no-fault system in 1992, but only from 2004

²However, the empirical results are ambiguous, and they depend (1) on the caps' target, as punitive damages, rather than economic or noneconomic damages, and (2) on the period of caps' introduction, the reforms implemented in the 1970s, in the 1980s, or in the 1990s. For a review see Kachalia and Mello (2011). Many European countries adopt schedules of noneconomic damages rather than caps, as in the case addressed by Bertoli and Grembi (2013)

the insurance system became completely public (also for private providers until 2013). So far, it has been difficult to empirically disentangle the effect of the change in the liability regime from the change in the insurance regime, so the attention has been focused especially on some elements of the latter, as in the English case (Fenn et al. 2007). The problems related to systems of public insurance as the case of the English *Clinical Negligence Scheme for Trusts* under a fault system or, for instance, of the Swedish *Patient Compensation Insurance* under a no-fault system are that public insurance can trigger common pool of risks more than private insurance. This means that often other institutional elements are requested to play a crucial role when adopting a public coverage, as for instance a proper monitoring mechanism on the flow of MM claims by a proper authority (Towse and Danzon 1999; Amaral-Garcia and Grembi (2014)).

A sound evaluation of the effects of a shift from negligence to strict liability has not been produced yet, and therefore, the main references are often translated by the experience of car accidents. Strict liability would allow for more liquidated claims, an average level of compensation lower than under alternative regimes, and a higher incidence of fatal accidents. However, given the peculiarities of the healthcare context, there are no strong arguments to infer that we should always expect the same results.

Changes at the fault regimes of MM remain one of the most debated issues lately. For instance, a recent document of the English Department of Health released on February 2014 (Department of Health UK 2014) addresses the concerns that the liability system did not foster medical innovation. Empirical analyses, which can help to address this and the use of additional tools to incentive efficient levels of precaution given specific institutional settings, are fundamental. Gathering and critically analyzing data on medical error remains priority one for many administrations (As underlined in the General Accounting Office (GAO) (USA) reports (GAO-04-128 T). The GAO has addressed several times the Congress to take initiatives in order to collect reliable data from the States regarding malpractice trends and effects).

Cross-References

- [Medical Liability](#)
- [Strict Liability Versus Negligence](#)

References

- Amaral-Garcia S, Grembi V (2014) Curb your premium: the impact of monitoring malpractice claims. *Health Policy* 144(2):139–146
- Andrews L (2005) Studying medical error in situ: implications for malpractice law and policy. *De Paul Law Rev* 54:357–392
- Arlen J (2013) Economic analysis of medical malpractice liability and its reform. New York University public law and legal theory working papers. Paper 398
- Barach P, Small SD (2000) Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. *Br Med J* 320:759–763
- Bertoli P, Grembi V (2013) Courts, scheduled damages, and medical malpractice insurance. Baffi center working paper. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2367218
- Ministero de Sanidad Y Consumo (2006) National study on hospitalisation-related adverse events. ENEAS 2005. Report. Madrid
- Danzon P (2000) Liability for medical malpractice, Chap 26. In: Newhouse J, Culyer A (eds) *Handbook of health economics*. Elsevier, New York, pp 1339–1404
- Davis P, Lay_Yee R, Briant R, Ali W, Scott A, Schug S (2002) Adverse events in New Zealand public hospitals I: occurrence and impact. *N Z Med J* 115(1167):1–9
- Department of Health (UK) (2014) Legislation to encourage medical innovation: a consultation. Available at <https://www.gov.uk/government/organisations/department-of-health>. Accessed Feb 2014
- European Commission, D.G. Sanco (EC) (2006) Special eurobarometer medical errors
- Fenn P, Gray A, Rickman N (2004) The economics of clinical negligence reform in England. *Econ J* 114:F272–F292
- Fenn P, Gray A, Rickman N (2007) Liability, insurance, and medical practice. *J Health Econ* 26:1057–1070
- GAO-04-128T (2003) Medical malpractice insurance. Multiple factors have contributed to premium rate increases, testimony before the subcommittee on wellness and human rights, committee on government reform, house of representatives. U.S. Government Printing Office, Washington, DC
- Grady MF (1992) Better medicine causes more lawsuits, and new administrative courts will not solve the problem. *Northwest Univ Law Rev* 86:1068–1093
- Hospitals of the European Union (HOPE) (2004) Insurance and malpractice, final report. Brussels, www.hope.be
- Kachalia A, Mello MM (2011) New directions in medical liability reform. *N Engl J Med* 364:1564–1572
- Kessler DP (2011) Evaluating the medical malpractice systems and options for reform. *J Econ Perspect* 25(2):93–110

- Leape LL, Berwick DM (2005) Five years after *To Err is Human*. What have we learned? JAMA 293:2384–2390
- Mello MM, Studdert DM, Brennan TA (2003) The new medical malpractice crisis. N Engl J Med 348:2281–2284
- Nys H (2009) The factual situation of medical liability in the member states of the council of Europe, reports from the rapporteurs, conference ‘The ever-growing challenge of medical liability: national and European responses’, Strasbourg, 2–3 June 2008, pp 17–28
- OECD (2006) Medical malpractice, insurance and coverage options, policy issues in insurance, no. 11
- Reason J (2000) Human error: models and management. Br Med J 320:768–770
- Rowe M (2004) Doctor’s responses to medical errors. Crit Rev Oncol Hematol 52:147–163
- Sage WM (2003) Understanding the first malpractice crisis of the 21st century. In: Gosfield AG (ed) Health law handbook. West Group, St. Paul, pp 549–608
- Shavell S (1987) Economic analysis of accident law. Harvard University Press, Cambridge
- Towse A, Danzon PM (1999) Medical negligence and the NHS: an economic analysis. Health Econ 8:93–101
- Vincent C, Neale G, Woloshynowych M (2001) Adverse events in British hospitals. Preliminary retrospective record review. Br Med J 322:517–519
- Weiler PC, Hiatt HH, Newhouse JP, Johnson WG, Brennan TA, Leape LL (1993) A measure of malpractice: medical injury, malpractice litigation and patient compensation. Harvard University Press, Cambridge
- Weingart SN, McL Wilson R, Gibberd RW, Harrison B (2000) Epidemiology of medical error. Br Med J 320:774–777
- Zegers MD, Bruijne MC, Wagner C, Hoonhout LHF, Waaijman R, Smits M, Hout FAG, Zwaan L, Christiaans-Dingelhoff I, Timmermans DRM, Groenewegen PP, Van Der Wal G (2009) Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care 18:297–302

Mercantilism

Günther Chaloupek

Austrian Chamber of Labour, Vienna, Austria

Definition

Mercantilism is a system of economic policy and a corpus of economic doctrines which developed

side by side from the sixteenth to the eighteenth century. The main goal was to increase a nation’s wealth and power by imposing government regulation to promote the nation’s commercial interests by maximizing exports and limiting imports.

Mercantilism, Colbertism, Cameralism

Mercantilism is a system of economic policy and a corpus of economic doctrines which developed side by side from the sixteenth to the eighteenth century. As a theory, mercantilism marks the decisive step in the emancipation of thinking about economic phenomena from scholastic theology to political economy and economics as a social science of its own. With respect to economic policy, mercantilism took a variety of different forms according to the different political, economic, and social conditions prevailing in European states during the early modern period. As a consequence, there are national variants of mercantilist economic literature focusing on different aspects of the economic process.

In a more general perspective of history of ideas, mercantilism as a body of theoretical doctrines as well as a system of economic policy is part of the emergence of a rationalist worldview with its understanding of natural phenomena in terms of cause and effect, instead of purpose inherent in the substance of things. Central to the new worldview is the concept of law as a force independent of human intention applicable to external physical nature and to human nature. In the spirit of the Baconian sentence *scientia est potentia*, knowledge of such laws brings with it the power to influence the course of events according to desired goals.

The early modern period was the time when nation states took shape on the European continent and in England. It was in this context in which the new political and economic doctrines acquired practical relevance. As a consequence, the focus of mercantilist authors was primarily on relations between states and on collectives within states (Pribram 1983, p. 83). That the perspective of the individual agent was only relevant in this context explains the contempt of classical and

neoclassical economics for mercantilism, whereas other schools of thought, such as the German Historical School or Keynesianism, are more prepared to acknowledge the merits of some of its doctrines.

None of the mercantilist authors has produced a compact theoretical system of the working of the economy. Rather, “the economics of mercantilism” is an *ex post* construct of history of economic theory which has identified a number of core issues and policy problem upon which the debates centered. Among those issues, the balance of trade, or, in a wider sense, the balance of payments of a nation is the most important one. But it would be wrong to identify the emphasis on achieving an export surplus with the opinion that national wealth amounts to nothing else than the accumulation of treasure, as was and still is often suggested. Rather, an improvement of the balance of payments was in various ways seen as a policy strategy to develop a country’s economic potential and thereby enhance its power in the context of rivalry among European nations. The relevance of specific ramifications of the external balance issue, such as import restrictions, export promotions, creation of legal monopolies, acquisition of colonies, exchange controls, varied according to the different circumstances under which individual nations sought to establish themselves in the competition of European powers. The key role for economic development which is assigned to trade extends beyond external to domestic trade and the means of its promotion. Increasing awareness of interdependence of a multitude of policy instruments lead to the formulation of policy concepts with the claim to regulate the economy as a whole (Colbertism). Given the prevalence of policy aspects in mercantilist literature, for some historians of economic theory the main relevance of mercantilism consisted in providing a political doctrine for formation of national states through the replacement of the local and regional government by the central government (Schmoller 1883) or as a system of power politics (Heckscher 1935). In contrast, Schumpeter (1954) carefully elaborated the substantial contributions of mercantilist authors to theoretical economic analysis.

If mercantilist economics had an early start in England, this can be attributed to the fact that the government had reached a comparatively high degree of centralization at the beginning of the Modern Age. In addition, consolidation of the central government coincided with the buildup of a colonial empire with its rapidly expanding trade. Conflicting interests of commercial capital are reflected in books and pamphlets whose authors often “clothed their views in the garb of a policy designed to strengthen the nation” (Roll 1942, p. 58f). Gerard de Malynes (ca. 1555–1643) warned against the loss of precious metal due to the fall of the exchange rate below silver parity which he attributed to the lack of foreign exchange controls and to the privilege of the East India Company for limited export of bullion. This focus on exchange rate was contradicted by Edward Misselden (ca. 1608–1654) who introduced the concept of balance of payments which should be seen as true indicator whether trade was beneficial for a country. A surplus should be achieved through promotion of exports and discouragement of imports, especially imports of luxuries. Interests of commercial capital found their “fullest expression” (Roll, p. 75) in the work of Thomas Mun (1571–1641). Building upon Misselden’s balance of payments theory, Mun argued that the export surplus augmented the capital (“stock”) that could be invested in trade and production and thus enhanced wealth and power of the nation. The debate between Josiah Childs (1630–1699) and Sir Dudley North (1641–1691) focused on the interest rate as a possible cause for the depression which hit England during the period of naval warfare against the Netherlands. Childs argued that the high interest rates which English merchants had to pay in comparison with Dutch competitors were responsible for depression and called for limits to be enforced by the state. North reversed the argument by saying that it was an increase in the volume of trade which would lead to an increase of the quantity of money and thus lower the rate of interest. If North proposed to do away with measures of trade protection and prohibition for that purpose, this foreshadows the end of mercantilist thinking in England. Of the important

contributions of Sir William Petty (1623–1687), mention should be made of his *Political Arithmetick* in which he advocated the use of *number, weight, and measure* in debates about economic issues – an early example of quantitative empiricist method in the attempt to establish laws of nature. English mercantilist authors extensively reflected also on prices and wages, taxation, population, etc., in the context of their respective causes. First attempts to assemble these elements of economic analysis into a coherent system from a mercantilist perspective, i.e., from the perspective of the state, were undertaken by Richard Cantillon in his *Essai sur la nature du commerce en general* (1755, originally written in English) and by Sir James Steuart in his *Principles of Political Economy* (1767).

If the **Netherlands** are often cited as model by English mercantilists, this is due the close identification of the country's government with the interest of the merchant class. While the political influence of the landed aristocracy in politics was still strong in England, in the seventeenth century Holland is the merchant state *par excellence*. The interests of Dutch merchants were best served by free trade for which the legal sciences (Hugo Grotius) provided the best arguments.

Following a path different from England in its formation of a national state, in **France** political and administrative power had been concentrated the hands of the king ruling as absolute monarch. French mercantilism was above all a comprehensive system of administration which sought to develop the economic powers of the country through “retablissement des manufactures,” advocated by Bartéhelemy Laffemas (1545–ca. 1612) who served as *controleur* under King Henri IV (Sommer 1920/1925, p. 29). A similar approach was pursued by Antoine de Monchrétien in his *Traicté de l'oeconomie politique* (1615) in which this term appears for the first time. French mercantilism was fully developed in practice, much less in theory, during the reign of Louis XIV by his finance minister Jean Baptiste Colbert (1619–1683). Under Colbert, population policy was adjusted to the aims of power policy, external trade was conducted as a kind of warfare against England and the Netherlands, the acquisition of

precious metal was proclaimed as the main objective of external trade, all measures of commercial policy were ruled by the endeavor to promote exports and prevent imports of final products (Pribram 1983, p. 51). “Colbertism” came to be used as synonym for mercantilism. It was the model for the “cameralist” system established in Austria in the eighteenth century.

In the territorial states of the **German Empire**, mercantilist practice and theory appears in different forms. In Prussia and in the Austrian monarchy mercantilist policies were deliberately used to create a unified internal market, thereby also strengthening political control of the central government over the heterogeneous provinces. Among the cameralist authors who offered their advice to the Habsburg Emperors, Johann Joachim Becher (1635–1683) made the most important theoretical contributions with his doctrine of market forms which distinguishes between monopoly, “polypolium,” i.e., free competition with free access to markets, and “propolium,” by which he means various kinds of restrictive or speculative practices. Rejecting all three forms, Becher pleads for some kind of organized and supervised competition. Becher assigns a key role to commerce in the efforts to make the economy more dynamic and recommends the foundation of sectoral trading companies by the state as instrument to encourage industrial activities. Philipp Wilhelm von Hörnigk, in his book *Österreich über alles, wenn es nur will* (1684) drew up a comprehensive program to create a national economy in the Habsburg crownlands. Hörnigk's tract as well as later cameralist literature highlight the essentially defensive orientation of Austrian mercantilist policies, which aim at an improvement of the external balance through a strategy of import substitution, in contrast to offensive export promotion by Western European states. Veit Ludwig von Seckendorff (1626–1692) who served as chancellor in the small state of Sachsen-Weimar puts significantly more emphasis on general conditions of production, such as reliability of legal framework, a stable monetary system, moderate taxation, education and training, investment in infrastructure, improvement of sanitary conditions, etc., while on

the other hand he devotes much less attention than Becher to interventionist measures of promotion of trades.

Johann Heinrich Gottlob Justi (1717–1768), the most important cameralist author of the eighteenth century, came close to recommending autarky. If external commerce was beneficial, it was not a necessity. “An empire may be very powerful, wealthy and flourishing without having external commerce with other peoples; alone, never can there be a state of such a character if its manufactures and industries are not flourishing.” It is in this context where Justi developed his central concept of *Universalkommerz* for which Colbert’s system served as model: “The sovereign has to direct all trades according to the needs of the country and to the requirements of its external commerce, of the promotion and augmentation of the livelihood of its subjects, and – in brief – of the general welfare.” Earlier than in England, chairs were established at German universities for mercantilist economics under the title “cameral sciences” or “police sciences,” the first one 1727, in Halle, Prussia. German textbooks, such as Justi’s *Grundsätze der Polizeywissenschaft* (2 vols., 1756) and Joseph von Sonnenfels’ *Grundsätze der Polizey, Handlung und Finanz* (3 vols., 1765ff), are handbooks for practical policy, rather than syntheses of analytical knowledge.

Well into the twentieth century, popular perceptions of mercantilism have been shaped by Adam Smith’s attack on what he called the “commercial” or “mercantile” system for its comprehensive regulations of imports and exports which he considered undue restrictions of freedom and obstacles to augment the wealth of a nation. Meanwhile, economic history and history of economic theory have corrected Smith’s verdicts in important respects and produced a more balanced picture of the merits and errors of this early stage of economics. It is widely recognized that free trade cannot under any circumstances be considered the best strategy to foster economic development, which can be supported by well-designed, temporary state interventions. On the other hand, a strategy of export-led growth which can turn into some kind of “new mercantilism” has

remained a powerful temptation not only for newly industrializing nations, with the risk of accumulating large-scale international imbalances. It appears that important issues of mercantilist economics will remain relevant in the twenty-first century.

References

- Heckscher EF (1935) Mercantilism. Allen & Unwin, London
- Pribram K (1983) A history of economic reasoning. Johns Hopkins University Press, Baltimore/London
- Roll E (1942) A history of economic thought. Prentice-Hall, New York
- Schmoller G (1883) Das Merkantilssystem in seiner historischen Bedeutung. In: Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft, vol VIII
- Schumpeter JA (1954) History of economic analysis. Allen & Unwin, London
- Sommer L (1920/1925) Die österreichischen Kameralisten in dogmengeschichtlicher Darstellung, Heft XI und XII der Studien zur Sozial-, Wirtschafts- und Verwaltungsgeschichte, Carl Grünberg C (ed), Vienna, reprint Scientia Verlag, Aalen 1967

Merchants' Law

► [Lex Mercatoria](#)

Merger Control

Tim Reuter

RBB Economics, Brussels, Belgium

Abstract

Merger control is at the heart of competition law institutions throughout the world. Firms, before they complete a merger or an acquisition, are required to get the transaction approved by competition authorities. The objectives of merger control, limiting the accrual of market power and protecting the welfare-generating competition forces of

the market, are well in line with economic theory. Economic theory has identified several types of effects that can arise from a merger, in particular unilateral and coordinated horizontal effects that result from an elimination of competition between firms supplying substitutable goods, and non-horizontal and conglomerate effects that result if firms are active on vertically or otherwise linked markets. We discuss the reasoning behind these effects and how they are assessed by competition authorities. We also discuss market definition as a first step in the assessment of mergers and the legal framework under which mergers are controlled.

Introduction: Objective of Merger Control and a Classification of Merger Effects

In most jurisdictions throughout the world, merger control procedures are a central element of competition law institutions. The objective of merger control is to limit the accrual of market power and to maintain the process of competition in the market in order to protect (consumer) welfare. To achieve these ends, merger control procedures must anticipate the competitive effects mergers will bring and establish enforceable rules according to which mergers are blocked or approved. In certain jurisdictions, merger control rules might have objectives other than maintaining competition, for example public interest objectives. Such objectives are however usually not considered as part of competition law in the narrower sense. We will not discuss these aspects in this note.

From an economic theory perspective, the first general theorem of welfare economics states that Pareto efficient allocations will only be achieved if firms behave as price takers, i.e., act as if their own output decisions have no influence on price or, in other words, have no market power. In comparison to such models with atomistic firms, models with imperfect competition feature losses of consumer welfare. The objectives of merger control are hence in line with a welfare

maximization objective. It is also general consensus that it is preferable to limit the accrual of market power via merger control over regulating firms not to exercise their market power once acquired (though competition laws have provisions to also limit the exercise of market power).

While an increase in market power leads to a lessening of (consumer) welfare from an economic theory viewpoint, mergers may be conducted without any effect in market power. They may even entail procompetitive effects. For example, it may be easier to generate economies of scale or scope after a merger. Mergers may also be conducted to increase productive efficiency or to merge two complementary businesses. The trade-off between allowing firms to realize efficiencies and allowing them to accrue market power has been denominated as “Williamson trade-off” after a famous paper by Williamson 1968. As such, any merger control regime that would block mergers per se would be overly interventionist. Instead, in competition law regimes throughout the world, case-by-case assessments are conducted, in order to predict the competitive effects the mergers entail.

From an economics perspective, mergers can lead to consumer harm in several ways. First, harm can arise from a lessening of competition if firms merge that are active on the same market, i.e., supply substitutable products. Such harm is labelled a horizontal effect. Two types of horizontal effects are distinguished: The first type arises if the merged entity has an incentive to increase prices as a result of a reduction of direct competitive pressure by the removal of a competitor. This is called a “unilateral effect.” The second type arises if the merger facilitates for the remaining firms in the market to coordinate on some anti-competitive behavior. This is called a “coordinated effect.”

A second category of potential harm is called vertical effects, i.e., effects that arise if the firms are active on related segments of the same supply chain. Vertical effects typically require that some competitor is foreclosed from accessing either inputs or customers. Finally, conglomerate effects can arise. Under conglomerate effects, merging firms neither supply competing goods nor are

they vertically linked. Harm can arise, because the merged firm can leverage its broader scope (for instance by bundling complementary goods). As a rule of thumb, horizontal effects are the most frequent concern. Vertical and conglomerate mergers are less likely to entail anticompetitive effects and even often give rise to procompetitive effects.

The Legal Framework of Merger Control

Before discussing the substantive economic assessment of mergers, based on which it will be decided whether a proposed merger is prohibited or not, we describe in this section certain procedural aspects of merger control. We focus on aspects that are common to most jurisdictions throughout the world.

First of all, usually not all mergers are assessed by competition authorities. Depending on the jurisdiction, usually some notification thresholds exist, below which mergers do not need to be cleared. These can be traction volume based (for example, in the USA, generally transactions in excess of 78.2 million USD are notifiable) or turnover based (under European Union law, mergers are notifiable if the combined worldwide turnover exceeds 5,000 million EUR or if the EU turnover of at least one concerned entity exceeds 100 million EUR). In most countries, notification means that a transaction must be approved before it is completed (where gun-jumping fines may be applied).

Regarding the scope of activities undertaken, the word merger control might suggest a too narrow definition of what it covers. In reality, in some jurisdictions, also concentrations short of a merger or an acquisition fall under merger control rules. For example, in the European Union, joint ventures have to be notified under certain conditions, and in some jurisdictions, also minority shareholding acquisitions have to be notified.

The role of authorities and courts can also differ fundamentally between jurisdictions. For example, under the European Union Merger Regulation, the European Commission can challenge mergers directly. Courts play only a role should

any party (including third party complainants) appeal the European Commission's decision. In the USA in contrast, mergers are investigated by the Department of Justice or the Federal Trade Commission. Should they decide to challenge a merger however, they have to seek an injunction in front of a court. Either way, usually precise time tables govern the merger control procedures.

Finally, the choice of competition authorities (or courts where applicable) is not limited to a choice between approving a merger and blocking it. Firms can also propose to remedy the concerns raised by competition authorities. These can be structural, i.e., the merging parties agree to divest certain activities (where the divestment transaction may be itself subject to merger control) or behavioral, i.e., the merging parties commit themselves to some postmerger behavior as agreed with the competition authorities.

The Substantive Assessments of Mergers

As discussed in the Introduction, the objective of merger control comes from the notion that restricting the accrual of market power enhances welfare. To implement this into enforceable rules, different legal tests have been established according to which mergers can be blocked. While traditionally, in many jurisdictions, the legal test for prohibiting a merger was whether it would establish or strengthen a dominant position; nowadays most jurisdictions have adopted a more nuanced test.

For example, according to the European Union Merger Regulation enforced by the European Commission (and most member states have similar rules), a merger is to be prohibited if it causes a *significant impediment of effective competition* (the so-called SIEC test), in particular if it gives rise to a dominant position (until 2004, the creation or strengthening of a dominant position was the sole reason for a prohibition under European merger control). Similar, in the USA and the UK, a merger is to be prohibited if it gives rise to a *substantial lessening of competition* (the so-called SLC test). While the SIEC test explicitly mentions the dominance criterion as a particular example

under which mergers can be blocked, the SIEC and SLC tests are generally considered to be similar and remaining differences for practical purposes are nuanced.

In the following, we describe how competition authorities economically assess and decide whether a proposed merger is to be cleared or blocked (and whether remedies might be appropriate) according to the SIEC or SLC test. This description is based on guidelines that various competition authorities have published to provide guidance on their assessment of mergers. In particular, the European Commission published a “Commission notice on the definition of the Relevant Market for the purposes of Community competition law” (1997), “Guidelines on the assessment of horizontal mergers” (2004), and “Guidelines on the assessment of non-horizontal mergers” (2008), while the US Department of Justice published “Horizontal Merger Guidelines” (2010) and “Non-horizontal merger guidelines” (1997).

A first step in the assessment of likely effects is the definition of the relevant market. Then, if necessary, horizontal and nonhorizontal effects are assessed. We discuss each of these steps in turn.

Market Definition

The first step of the assessment of mergers under both the SIEC and the SLC test is the definition of the relevant market. The objective of defining the relevant market is to identify the competitive constraints firms face and to provide a framework for the competitive assessment. In particular, defining the market allows the measurement of market shares and other concentration-based statistics.

The definition of a relevant market comprises the combination of both a product market definition and a geographic market definition. With respect to the product market definition, a given product market is constituted by all products that are interchangeable or substitutable by consumers. The substitutability is to consider all product characteristics, prices, and the intended use of the products. With respect to the geographic market definition, the market comprises all areas

in which the conditions for supply and demand for the given services are homogenous and can be distinguished from other areas in which conditions of competition are sufficiently different.

In practice, to assess the relevant market definition, the Hypothetical Monopolies Test (“HMT”) is implemented. The HMT is a sequential test that starts with the narrowest possible candidate market and asks whether it would be profitable for a hypothetical monopolist in that market to implement a small but significant non-transitory increase in price (“SSNIP”). If so, it is worth monopolizing the market, i.e., a monopolist in this market does not face significant competition from outside this market and the relevant market is found. If it is not profitable to increase prices, the hypothetical monopolist faces some competition from outside the market. The candidate market definition is discarded. As next sequential step then, the candidate market definition is widened and the SSNIP question is reapplied assuming a hypothetical monopolist for the wider market. The process is reiterated until a market is wide enough such that the price increase is profitable. For this market, the hypothetical monopolist faces no significant competition from outside the market.

The HMT can be implemented using several empirical techniques to answer the question whether a price increase is profitable for the hypothetical monopolist, for example critical loss analysis, demand estimation, analysis of geographic sales patterns, analysis of price levels and price correlation, and stationarity analysis (see Gore et al. 2013 for a description of these techniques).

Finally, it should be noted that certain commentators have argued that in some cases, market definition may not be a necessary step and that economic effects can be directly assessed (see Kaplow 2010). This notion seems to be better received in some jurisdictions (in particular in the USA and the UK) than in others.

Horizontal Effects

Horizontal effects may arise under mergers of firms that supply substitutable goods, i.e., firms

that are competing directly with one another. Two types of horizontal effects are distinguished. First, a merger may eliminate direct competition between the merging firms (called a “unilateral effect” or a “noncoordinated effect”). Second, a merger may lessen or impede competition by making coordinated behavior more likely, i.e., behavior of multiple firms that is profitable for each of them only as a result of the reactions of the other firms, i.e., (tacit or explicit) collusion (called a “coordinated effect”). We discuss unilateral and coordinated effects in turn.

Unilateral Effects

A first step in the assessment of unilateral effects is usually the calculation of market shares and other concentration measures, such as the Herfindahl concentration index (“HHI”). These measures are in particular useful evidence if products are homogeneous, if capacity can be easily expanded, and if goods are traded in a spot-market fashion. In these cases, concentration measures often constitute a presumption for the competitive effects of the merger (for instance, the European Commission presumes that there is a significant impediment of effective competition if the combined market shares of the merging parties are above 40%; in US enforcement, mergers involving an HHI increase of more than 200 points are presumed as likely to enhance market power). In some cases, competition authorities conduct price-concentration analyses, investigating whether there is an empirical relationship between the concentration in a market and its given price level.

However, concentration-based measures are in many cases only the starting point of the competitive assessment of mergers. Further elements are particularly relevant if products within the market are not homogeneous, if firms in the market face capacity constraints, or if customers procure the goods in a nonstandard way, for example, via auctions.

Concentration-based criteria are unable to inform competition authorities in a precise way about the competitive constraints firms face, in particular in markets, in which products are differentiated. For instance, if the market definition has revealed that the three products A, B, and

C are in the relevant market, and if B and C have the same market shares, the concentration-based criteria presume that B and C exert the same competitive constraint on the pricing of product A. However, consumers may actually be more willing to substitute product A with product B rather than with product C. In such a case, competition authorities investigate typically the precise substitutability between products to determine the relative importance of all products in the defined market as competitive constraints to the products of the merging firms.

For instance, the European Commission investigates the closeness of competition of goods within the relevant market. One piece of evidence for this purpose can be an analysis of diversion rates, e.g., the fraction of unit sales lost by a price increase of a certain goods that would be diverted to the sales of a second product. A high diversion ratio from a product of one of the merging firms to a product of the other merging firm is an indication of a high likelihood of an anticompetitive effect (*ceteris paribus* the market shares).

Competition authorities use a number of empirical techniques to assess the competitive constraints imposed by particular products within the relevant market, for example, diversion rates can be assessed using survey evidence, customer switching data from firms’ market intelligence databases (e.g., mobile number portability data in the case of mobile network operator mergers – see European Commission 2016), assessments of natural experiments (e.g., switching as the result of a closure of a plant – see Coate 2012), and analysis of win/loss data (see Botteman 2006). Advanced tools to estimate the constraints of particular products are upward pricing pressure tests (see Farrell and Shapiro 2010) and merger simulation models (see Budzinski and Ruhmer 2009).

Other factors that play a role in the assessment of unilateral effects are entry, capacity constraints of suppliers, countervailing or buyer power, efficiencies generated by the merger, product repositioning, failing firm defense, effects of the merger on innovation and product variety, and a loss of potential competition (e.g., one merging party planning to enter a given market where the other is already active). These factors may be

treated differently in different jurisdictions however. For example, Brouwer (2008) points out that the main difference between EU and USA merger policy lies in the greater scope for efficiency arguments in the USA.

Coordinated Effects

The notion of coordinated effects relies on the idea that a merger can make (tacit) collusion more likely to arise or can increase its effectiveness. As is well established by economic theory, collusive situations are difficult to establish for firms, because in a static setting, firms have generally an incentive to deviate from such situation by competing more aggressively. However, in a dynamic setting, collusive situations can be stable such that no firm has an incentive to deviate.

For the implementation in merger control, coordinated effects are assessed by judging whether the relevant market is generally vulnerable to coordinated conduct and whether the vulnerability increases by the merger.

For the vulnerability of the market in general, it is usually assessed (i) whether the firms in the market are likely to reach a common understanding of the terms of the coordination, (ii) whether firms can monitor whether all firms adhere to the coordination conduct, (iii) whether credible deterrent mechanisms are available that prevent firms from deviating from the coordination, and (iv) whether outside firms (competitors or customers) have ways of destabilizing the coordination.

As shown in economic theory, (tacit) collusion in dynamic settings often has many equilibria (see Friedman 1971). It is less clear which equilibrium may be a focal point and how firms coordinate on a specific tacit collusion equilibrium to play. This is reflected in the assessment of coordinative effects that firms must reach an understanding of the terms of the coordination conduct. While in some markets, for example, a market with few competitors, where competitors are symmetric, with homogenous products and that is not strongly affected by entry and innovation, reaching such an understanding may be simple. In markets which are more complex, reaching a tacit understanding is less likely. Hence, merger

control takes these market characteristics into account when assessing coordinative effects.

The ability of firms to monitor and punish deviating firms depends on market characteristics such as transparency and stability of demand conditions (i.e., to determine whether an unexpected market outcome is result of deviating behavior or result of demand fluctuation), frequency and concentration of orders (it becomes more difficult to punish if orders are lumpy), and whether firms compete in a single- or multimarket setting (as punishment can also occur on markets other than the one on which coordination takes place).

Finally, competition authorities take reactions by other firms into account, for example whether competing providers that do not take part in the coordination have the ability to increase capacity, whether third parties can enter in the market and whether customers have countervailing power that can destabilize the coordination.

For the facilitation of coordination by the merger, after all what matters in whether coordination becomes more likely or more effective by the merger, the reduction of the number of firms active can be enough if market concentration is sufficiently high. However, even if the merger does not cause a significant increase in concentration, coordinative effects can be found, for example, if one of the merging firms is found to be a 'maverick,' i.e., a firm that is known to be disruptive to the market, e.g., by pricing aggressively or by innovating regularly.

Nonhorizontal Effects

A merger can also affect market outcomes if the merging firms do not compete on the same market, but are active on different levels on the same supply chain (i.e., a vertical merger) or are active on different markets that are somehow related, for example, two markets for complement goods (i.e., a conglomerate merger).

In general, nonhorizontal mergers tend to be less likely to raise competitive concerns than horizontal mergers, because unlike the latter, they do not lead to a loss of direct competition between the merging firms. Often vertical mergers solve

inefficiencies in the supply chain and are hence procompetitive; for example, they reduce double-marginalization (i.e., lower downstream prices, because the merged entity internalizes increased upstream profits from an expansion of output if downstream prices decrease), decrease transaction cost, or reduce prices for complement goods.

However, vertical mergers may not only solve inefficiencies and lead to procompetitive effects, but can under certain circumstances have anticompetitive effects as well. Two types of such anticompetitive effects are input foreclosure and customer foreclosure. Input foreclosure relies on the notion that a firm active on a downstream market, by vertically integrating upwards, can start supplying its competitors on the downstream market on unfavorable terms only, thereby weakening downstream competition and increasing downstream prices. Customer foreclosure, on the other hand, means that a firm on an upstream market, by integrating downward, can shrink the customer base of its upstream competitors, thereby decreasing the ability of the upstream competitors to compete. In result, upstream prices may increase, whereby competition on the downstream market is hurt, leading to an increase of downstream prices.

Whether it is for the merging firms possible and profitable to foreclose competitors from accessing inputs or customers is however dependent on the specific market characteristics. For example, a downstream competitor cannot be foreclosed by a vertical merger if the merged entity does not have sufficient market power in the upstream market, because the downstream competitors can get access to the input from competing providers (at competitive terms). Similar, while input foreclosure may increase the integrated firm's profits on the downstream market, it pays a cost for the foreclosure by lost profits resulting from less sales on the upstream market. As such, the merged entity might not have an incentive to engage in foreclosure, for example, if upstream margins are high and downstream margins are low.

In a similar vein, customer foreclosure is not necessarily possible and profitable for vertically integrated firms. First of all, for customer foreclosure to be possible, the integrated firm must have a sufficient degree of market power in

the downstream market, such that competing upstream providers may actually face a significant loss of sales opportunities. Second, the loss of customers for the upstream competitors must entail that they are not able to compete any more effectively, which requires some form of economies of scale or scope for the upstream providers or that they operate at or close to minimum efficient scale. The profitability of consumer foreclosure depends on a similar trade-off as that of input foreclosure: By customer foreclosure, the integrated firm can lower competition and raise prices on the downstream market. To do so, it has however to carry additional costs on the upstream market (by eventually not procuring from the cheapest provider).

Foreclosure concerns can also matter in conglomerate mergers, i.e., for mergers in which the merging firms are active neither on the same market, nor on vertically linked markets. Similarly to vertical cases, conglomerate mergers will often entail procompetitive effects, for example, when the merging firms sell complementing goods (because they will lower prices for both goods, taking into account that a price decrease for one good will trigger an increase in demand for the complement). As well similar to vertical mergers, they can however trigger foreclosure concerns. For example, by tying or bundling two goods together, the conglomerate firm may leverage market power it holds in one market to another market to achieve above-competitive prices in that market. However, the firm will only have the possibility to leverage the power of one market to the other, if its market power in the first market is sufficiently high and if a large enough share of the consumers who buy one good also want to buy the second. The profitability of such foreclosure depends on the trade-off between losses in the market with market power (because some customers might refrain from buying this good if it is bundled to the other good) and gains in the market to which the power is leveraged.

For all types of foreclosure, competition authorities are hence comprehensively assessing market power of the integrated firm in at least one market, the cost of leveraging the market power that is entailed on the same market, and the benefit the leveraging causes on the second market.

Cross-References

- Abuse of Dominance
- Horizontal Effects
- Market Definition
- Merger Remedies
- Type-I and Type-II Errors

References

- Botteman Y (2006) Mergers, standard of proof and expert economic evidence. *J Compet Law Econ* 2(1):71–100
- Brouwer M (2008) Horizontal mergers and efficiencies; theory and anti trust practice. *Eur J Law Econ* 26(1): 11–26
- Budzinski O, Ruhmer I (2009) Merger simulation in competition policy: a survey. *J Compet Law Econ* 6(2): 277–319
- Coate MB (2012) The use of natural experiments in merger analysis. *J Antitrust Enforc* 1(2):437–467
- Department of Justice and Federal Trade Commission (1997) Non-horizontal merger guidelines, <https://www.justice.gov/sites/default/files/atr/legacy/2006/05/18/2614.pdf>
- Department of Justice and Federal Trade Commission (2010) Horizontal merger guidelines, <https://www.ftc.gov/sites/default/files/attachments/merger-review/100819hmg.pdf>
- European Commission (1997) Commission notice on the definition of the relevant market for the purposes of community competition law. *Off J Eur Union C* 372:155–163
- European Commission (2004) Guidelines on the assessment of horizontal mergers under the council regulation on the control of concentrations between undertakings. *Off J Eur Union C* 31:5–18
- European Commission (2008) Guidelines on the assessment of non-horizontal mergers under the council regulation on the control of concentrations between undertakings. *Off J Eur Union C* 265:6–25
- European Commission (2016) Recent developments in telecoms mergers. *Compet Merger Brief* 3(2016): 1–8
- Farrell J, Shapiro C (2010) Antitrust evaluation of horizontal mergers: an economic alternative to market definition. *BE J Theor Econ Pol Perspect* 10(1), Article 9
- Friedman J (1971) A non-cooperative equilibrium for Supergames. *Rev Econ Stud* 38(1):1–12
- Gore D, Lewis S, Lofaro A, Dethmers F (2013) The economic assessment of mergers under European competition law. Cambridge University Press, Cambridge
- Kaplow L (2010) Why (Ever) define markets? *Harv Law Rev* 124:437–440
- Williamson O (1968) Economics as an anti-trust defense: the welfare trade-offs. *Am Econ Rev* 58(1):18–36

Merger Remedies

Patrice Bougette

Department of Economics, Université

Côte d’Azur, CNRS, GREDEG,

Nice, France

Abstract

Merger remedies are used by competition agencies to prevent the harm to the competitive process that may result as a consequence of a merger. They allow for the approval of mergers that would otherwise have been prohibited, by removing the anticompetitive concerns that a given transaction may pose to competition. First, we present the typology of merger remedies generally used. Second, we analyze the link between the size of the offered remedies and the level of efficiency gains announced in a context of asymmetric information. Third, we summarize the results of several retrospective merger studies in which remedies have been used.

Competition agencies use merger remedies when a notified merger is likely to raise some competitive concerns. In this case, the merging firms may make a remedial offer, which may be accepted or rejected by the agency. In this entry, we derive examples mainly from the European Commission (hereafter “the EC”) even though most mechanisms can be found in any competition agencies worldwide.

Actually, remedies are relatively less used with respect to the total number of notified merger proposals to agencies. According to the EC’s data, less than 10% of notified mergers are eventually conditioned upon merger remedies (see the EC’s website for detailed data on the European merger control, <http://ec.europa.eu/competition/mergers/statistics.pdf>). The majority are approved without any conditions. However, these remedies may be found in large or complex merger cases. In the absence of remedies, such mergers would be rejected.

Types of Merger Remedies

Two types of remedies are commonly used by competition agencies, structural and behavioral remedies. We first detail them and then discuss the pros and cons of each in terms of implementation.

Structural remedies refer to a transfer of ownership rights. For instance, merger firms commit themselves to selling a portion of their assets to one or several buyers. The selected assets may be located in markets where the level of concentration is very high. These assets are often the merging parties' overlap. For instance, in 2016, the EC approved the acquisition of beer group SABMiller by Anheuser-Busch InBev provided that, among other commitments, the new group divested SAB's brands Peroni, Pilsner Urquell, and Grolsch (*ABI/SAB* merger case, M.7881; see the EC press release, "Mergers: Commission approves AB InBev's acquisition of SABMiller, subject to conditions," 24 May 2016).

By nature, structural remedies are difficult to reverse and may largely apply to horizontal practices. The sale of an independent activity operating in the market appears more effective than a transfer of heterogeneous assets from the two merging firms (see the European Commission (2005)'s study for a categorization of these remedies).

Behavioral remedies may be defined as constraints on the property rights of the new entity. These commitments do not affect the market structure. They include third-party access to infrastructure or technology, or the parties commit themselves to putting an end to any exclusive vertical agreements. For instance, in the context of the acquisition of Ariespace by Airbus Safran Launchers (ASL), a joint venture between Airbus and Safran, the EC had concerns that the transaction would give rise to flow of potentially sensitive information between the two companies. This would have been at the expense of rival satellite manufacturers and launch service providers. Among other behavioral remedies, the companies committed not to share information about third parties with each other (*ASL/Ariespace* merger case, M.7724; see the EC press release, "Mergers: Commission approves

acquisition of Ariespace by ASL, subject to conditions, 20 July 2016).

In general, competition agencies have a preference for structural commitments. Indeed, a behavioral remedy involves direct monitoring costs, whereas in the case of a structural remedy, once transferred, the assets no longer need special monitoring. However, in this case, it may be useful to set up a follow-up mechanism until the assets are sold. For instance, with regard to complex merger cases, an independent trustee may be appointed to monitor the smooth transfer of assets and strict compliance with time limits (see the Commission Notice on remedies acceptable under the Council Regulation (EC) No 139/2004 and under Commission Regulation (EC) No 802/2004 Official Journal C 267, 22.10.2008, p. 1–27, http://ec.europa.eu/competition/mergers/legislation/files_remedies/remedies_notice_en.pdf).

Nevertheless, the preference for a structural option may be flexible and adapted to the case context. First, the two types of remedies are not necessarily exclusive, behavioral commitments may be useful to a structural solution (see, e.g., the EC *Orange/Jazztel* merger case, M.7421, 19 May 2015). Then, in some cases, perhaps no competitor is interested in the proposed assets, in which case the agency turns to a behavioral commitment. In addition, in highly changing market circumstances, nonstructural remedies may constitute a more flexible and a revisable solution (Motta et al. 2007). Lastly, transfer of assets may facilitate collusion if they lead to a more symmetric industry structure. Compte et al. (2002) show that due to the use of remedies, coordinated effects emerged in the context of the 1992 *Nestlé–Perrier* merger case.

Efficiency Gains and Asymmetric Information

A number of theoretical models have focused on the effects of structural merger remedies. For instance, such remedies enlarge the scope for approvable mergers in the presence of merger synergies. Nonetheless, if the merger does not lead to any efficiency gains (i.e., cost synergies),

only under very restrictive conditions, reallocation of assets through structural remedies may satisfy the criterion of consumer surplus (Vergé 2010).

With regard to the nature of the *ex ante* merger control, asymmetric information problems arise between the merging firms and the competition agencies. Firms know precisely the level of their own efficiency gains expected, while the competition agency may doubt the real level of expected synergies (adverse selection). The merger remedies may be used to signal the true type of efficiency gains (see, e.g., Dertwinkel-Kalt and Wey (2016)).

The strategic trade-off for the merging firms is the following. On one hand, the greater the synergies of a proposed merger, the more companies may value their assets, and therefore are reluctant to propose significant commitments. Thus, theoretically, efficiency gains reduce the size of the proposed commitments (Cosnita and Tropicano 2009; Bougette 2010).

On the other hand, this argument does not take into account the temporal dimension of the trade-off. The EC's investigations are costly for merging firms. In this sense, firms may judge that it is preferable to divest more assets ("overfixing" strategy) than what exactly would suffice to resolve the competition concern, in order to avoid a long, and therefore, costly control procedure. Thus, in spite of the high level of synergies generated, the merging companies may then be inclined to give up on strong commitments simply because they are delay averse.

Cosnita-Langlais and Tropicano (2012) analyze the link between the efficiency defense and the use of remedies. They model the quality of information held by the competition agency and show that it may be best for the agency to prohibit the efficiency defense when the information quality is poor.

Merger Remedies Retrospective Studies

With regard to empirical applications dedicated to remedies, several levels of study in the literature may be specified.

First, empirical analysis has focused on specific merger cases. These studies aim to assess the effectiveness of the decision made by competition agencies. Structural remedies may have been used and could be evaluated as such. In most cases, a difference-in-differences approach is adopted. The chosen econometric method allows for estimating a counterfactual, namely, what would have occurred in the absence of the studied merger, or in this case, in the absence of remedies. For example, Tenn and Yun (2011) show that the structural remedies used in the 2006 *J&J-Pfizer* merger resulted in the return of the premerger situation.

Second, another approach to evaluating selected remedies is to build a database of merger decisions including the ones with remedies and to analyze the determinants of these remedies. Several studies on the EC's data have been released in this perspective (see, e.g., Bougette and Turolla (2008)). Some of them have studied the merger control process in general, while others focus exclusively on the remedial decisions. For instance, Duso et al. (2011) show that remedies may be more effective when anti-competitive concerns are not too harmful and when applied to the first rather than the second investigation phase.

Time actually plays a considerable role in companies' strategy to provide more or less strict merger remedies, when needed. By studying 254 cases of mergers from the EC over the period 1999–2010, Ormosi (2012) shows that companies that do not use an efficiency defense strategy are actually more likely to reach a deal relatively early. The experience of law firms hired for the case has a high impact on the probability of using efficiency defense, but only for European cabinets. Less restrictive remedies are more likely to be proposed by the parties at the outset of the proceedings.

Based on the EC's data, Garrod and Lyons (2016) also show that the probability of early settlement is increasing in delay costs of the merging parties, decreasing in the uncertainty associated with the complexity of the economic assessment, and decreasing in the case load of the EC when resources are plentiful.

Finally, studies and reports have been conducted and prepared by the agencies themselves in a self-evaluation exercise of their decisions. The most notable of these is the European Commission (2005) that pointed out a number of practical problems in terms of monitoring, divested asset selection, and potential buyers, among others.

Cross-References

- [Competition Policy: France](#)
- [Difference-in-Difference](#)
- [Merger Control](#)

References

- Bougette P (2010) Preventing merger unilateral effects: a nash–cournot approach to asset divestitures. *Res Econ* 64:162–174
- Bougette P, Tuolla S (2008) Market structures, political surroundings, and merger remedies: an empirical investigation of the EC’s decisions. *Eur J Law Econ* 25:125–150
- Compte O, Jenny F, Rey P (2002) Capacity constraints, mergers and collusion. *Eur Econ Rev* 46:1–29
- Cosnita A, Tropéano JP (2009) Negotiating remedies: revealing the merger efficiency gains. *Int J Ind Organ* 27:188–196
- Cosnita-Langlais A, Tropéano JP (2012) Do remedies affect the efficiency defense? An optimal merger-control analysis. *Int J Ind Organ* 30:58–66
- Dertwinkel-Kalt M, Wey C (2016) Structural remedies as a signaling device. *Inf Econ Policy* 35:1–6
- Duso T, Gugler K, Yurtoglu BB (2011) How effective is European merger control? *Eur Econ Rev* 55: 980–1006
- European Commission (2005) Merger remedies study. Technical report
- Garrod L, Lyons BR (2016) Early settlement in European merger control. *J Ind Econ* 64:27–63
- Motta M, Polo M, Vasconcelos H (2007) Merger remedies in the European Union: an overview. *Antitrust Bull* 52:603–631
- Ormosi PL (2012) Claim efficiencies or offer remedies? An analysis of litigation strategies in EC mergers. *Int J Ind Organ* 30:578–592
- Tenn S, Yun JM (2011) The success of divestitures in merger enforcement: evidence from the J&J-Pfizer transaction. *Int J Ind Organ* 29:273–282
- Vergé T (2010) Horizontal mergers, structural remedies, and consumer welfare in a cournot oligopoly. *J Ind Econ* 58:723–741

Merit Goods

Valérie Clément¹ and Nathalie Moureau^{2,3}

¹MRE, EA 7491, Université de Montpellier, Montpellier, France

²Université Paul-Valéry MONTPELLIER 3, Montpellier, France

³ARTdev, UMR 5281, Université Paul Valéry, Montpellier, France

Merit goods are a category of goods, introduced in the debate by Musgrave (1957), which individuals tend to under- or over-consume because their preferences are “irrational” or “defective.” This leads individuals to make suboptimal choices, which are detrimental to their well-being. Now, if they exist, merit goods must be produced by the government that must so to speak force individuals to consume the correct amount of these goods. In other words, the government must behave paternalistically.

The concept of merit goods was a precursor to the debates on paternalism within welfare economics. In particular, the interpretation of the merit goods concept through the meta-preferences approach helps in legitimizing legal intervention and achieving a more efficient regulation.

When Musgrave introduced the term “merit goods” (originally called *merit wants*), it was in an attempt to create a normative definition for government functions. Nevertheless, only three of the functions he studied in his article have gone down in history: (i) the provision of public goods (*service branch*), (ii) the redistribution of income (*distribution branch*), and (iii) economic regulation (*stabilization branch*). Yet, in this groundbreaking article, Musgrave also mentioned another category of goods which he called merit wants. He was referring to goods which are subject to “transfers in kind” (e.g., social housing) and for which the regulator’s preferences override individual choices (Musgrave 1957 p. 341).

In 1959, Musgrave returned to this concept of merit goods by explicitly linking it to the issue of consumer sovereignty. In some cases, when choices made by people on the markets do not

lead to a situation that maximizes their well-being, the regulator intervenes in order to address the limitations of individual preferences and correct people's choices in their own best interest.

It is nevertheless in his 1987 Palgrave article that Musgrave strengthened the definition he had introduced 30 years earlier. He clarified two points in particular which attracted most comments since they were first published.

Firstly, Musgrave confirms his initial theoretical claim that the justification for government intervention through merit goods is distinct from that linked to market failures and redistribution. Indeed, while links between merit goods, public goods, and externalities may have caused some confusion in his initial papers (Head 1966, 1969; Ver Eecke 2001), the Palgrave article provides clarification. In no way should merit goods be confused with public goods or externalities. Whereas in the case of public goods there is a link between consumers' willingness to pay and consumption levels, this link is broken in the case of merit goods. Furthermore, merit goods refer to situations where people's choices are detrimental to their own well-being without third parties being involved, as is the case with externalities.

Secondly, at the heart of the definition of merit goods lies the fact that if choices are detrimental to individual, it is because their current preferences are defective. Thus, choices then expressed in the market no longer equate with welfare. These *individual failures* could justify government interventions (Jones and Cullis 2002).

The reasons why choices made on the market may lead to a suboptimal situation have been the subject of extensive debate. In the article he wrote for the Palgrave Dictionary, Musgrave takes the view that situations in which people voluntarily delegate their choice to a more informed party, in a principal agent relationship, do not relate to merit goods. However, in his early works, he did not take this stance and had in fact used education as a prime example of merit goods. Indeed, at first he considered that the reason why education was compulsory was because people were not able to forecast the profit they would earn of such an investment. He nevertheless changed his mind, stating that it was simply an information issue

encountered by the individual which justified a delegation of choice to another better informed party (For sure, one must admit that the government is better informed; that is not at all accepted in the economic literature.) (Musgrave 1987; West and McKee 1983). Defining the concept of merit goods is rather about highlighting the inconsistency of the preference standard in order to form judgements on individual well-being. Hence, it seems that even when full information is available, wrong choices can be made and lead to a suboptimal situation for the individual.

By definition, merit goods infringe consumer sovereignty and for this reason were put aside the standard welfare economics framework as the golden standard for paternalism (McLure 1968). However, there have been attempts to model merit goods in the context of welfare economics (Pazner 1972; Roskamp 1975; Wenzel and Wiegard 1981; Salanié and Treich 2009). These attempts perhaps reflect the need to justify an extremely widespread regulatory practice. For example, OECD data shows that two-thirds of European government bodies expenditure are somehow justified in terms of merit goods (Fiorito and Kollintzas 2004) and cannot be explained by standard market failure arguments.

If current short-term preferences are disqualified, the question arises of how "authentic" preference could be defined and what it stands for. The theoretical issue underlying this question lies in the possibility of articulating merit goods with the classical liberal principle of normative individualism. Musgrave did not evade the issue. In some of his papers, he noted that there is an elite who is in a position to know people's "true preferences" or "authentic preferences" (Musgrave 1969); in other papers, he refers to collective norms or "community preferences" (Musgrave 1987).

Another way to justify the concept of merit good in the economic framework, which seems more in line with the theoretical issue, involved expanding the area of individual preferences beyond market preferences, displayed through the willingness to pay and choice, by introducing the notions of "multiple-selves" and "meta-preference." The economic agent is then defined

by a collection of different and independent personalities (Harsanyi 1955; Elster 1979; Etzioni 1986), each of which leads to a separate classification of available options. The individual is no longer a unified person and may struggle to control his/her behavior (Schelling 1984). Equally, the individual may have the ability to assess and reflect on his/her own tastes and preferences that are expressed through second-order preferences or meta-preferences (Frankfurt 1971; Jeffrey 1974; Sen 1977; Hirschman 1984; George 1998). These help reflect the individual's dissatisfaction with a choice that he/she nevertheless made. The regulator then appears as a mediator between preferences displayed on the market on the one hand and reflexive preferences on the other, this mediation being then carried out under merit goods (Brennan and Lomasky 1983).

Incorporating into the analysis this idea of the existence of several ranges of preferences enables decisions made by policymakers, legislators, and judges to be perceived as an expression of second-order or meta-preferences, which, in turn, allows for the regulation implemented under merit goods to lead to higher efficiency than that implemented by the market, while respecting the individualistic foundations of collective choice. In this sense, the interpretation of Musgrave's concept through reflexive preferences is particularly relevant when analyzing economic policies and regulation policies within a law and economics approach. As noted by Kirchgässner (2017), this ties in with an important tradition in political philosophy which, from Buchanan and Tullock (1962) to Rawls (1971), combines the choice of a constitution or of the general principles on which society is organized with higher-order preferences.

With the developments of libertarian paternalism (Sunstein and Thaler 2003), the question of the role played by merit goods in the economic framework remains topical. By exploring faulty reasoning and rationality defaults, behavioral economics actually deepen the empirical content of Musgrave merit goods' rationale. Nevertheless, this strand of literature quite surprisingly did not refer to the concept of merit good in its developments of a new framework for the state regulation. Obviously, behavioral economists argue against

merit goods' intervention as it represents hard paternalism restricting individual choices through prohibition or taxation. Behavioral economists favor nudge practices, where the regulator helps people to make the best choice through a change in the frame or in the environment of choices (Thaler and Sunstein 2008). In any case, there should not be any restriction of the available options provided through the market allocation.

Then, behavioral economics did not pay heed to the contribution of the merit goods' argument to the normative justification for the government intervention. In this respect, the continuity with the pioneering concept of Musgrave is certainly to be found in the role and definition devoted to the merit goods in law and economics following Calabresi (2016). Calabresi complements and extends the definition of Musgrave adding two reasons why merit goods should not be (and actually are not) allocated through markets: the refusal of the "commodification" and pricing of certain goods and the refusal of an allocation based on people's willingness to pay given the vast inequalities in wealth distribution in our societies.

In the first category, people object to the use of monetary evaluation and measurement for being conducive toward unacceptable trade-offs, for example, trade-offs implying life or safety and money. The second category of merit goods following Calabresi includes those goods whose measurement in monetary term is no longer objectionable, but people oppose the use of the pure market mechanisms because the allocation thus depends on the prevailing unequal distribution of wealth; examples are military service or the right to obtain body parts (blood, kidney, etc.) or the right to a basic education. Taking seriously people's actual preferences embedded in these two merit goods categories allows the society to avoid "indirect external moral costs," Calabresi argues, that arise from the denial of people's objection to commodification and to the neglecting of the distributional consequences of the pure market allocation. Hence the allocation of merit goods should rest on hybrid mechanisms involving either modified market or modified command schemes if people preferences for merit goods are taken into account. Tort laws

provide a prominent example of such hybrid mechanisms in the case of objection to commodification that lead to lessen the externalities created by merit goods, that is, “moral costs” people would bear were the merit goods (life and safety) priced directly through the market, Calabresi points out.

Merit goods lie at the heart of law and economics as Calabresi conceived it, first of all because the inclusion of people’s preferences about commodification and equality enables regulation policies to be efficient in the sense that third-party moral costs are fully integrated and lastly because the thorough study of the law and the legal institutions should serve to identify merit goods and to elicit people’s preferences about merit goods. This renewal of the merit goods’ argument confirms the initial statement of Musgrave that merit goods were a category of goods that called for the expansion of the standard economic model.

Cross-References

- [Libertarian Paternalism](#)
- [Nudge](#)

References

- Brennan G, Lomasky L (1983) Institutional aspects of ‘merit goods’ analysis. *Finanzarchiv* 41(2):183–206
- Buchanan JM, Tullock G (1962) *The calculus of consent: logical foundations of constitutional democracy*. University of Michigan Press, Ann Arbor
- Calabresi G (2016) *The future of law and economics*. Yale University Press, New Haven
- Elster J (1979) *Ulysses and the sirens: studies in rationality and irrationality*. Cambridge University Press, New York
- Etzioni A (1986) The case for a multiple utility conception. *Econ Philos* 2(2):159–183
- Fiorito R, Kollintzas T (2004) Public goods merit goods and the relation between private and government consumption. *Eur Econ Rev* 48:1367–1398
- Frankfurt HG (1971) Freedom of the will and the concept of a person. *J Philos* 68(1):5–20
- George D (1998) Coping rationally with unpreferred preferences. *East Econ J* 24(2):181–194
- Harsanyi J (1955) Cardinal welfare individualistic ethics and interpersonal comparisons of utility. *J Polit Econ* 63(4):309–321
- Head JG (1966) On merit goods. *Finanzarchiv* 25(1):1–29
- Head JG (1969) Merit goods revisited. *Finanzarchiv* 28(2):214–225
- Hirschman AO (1984) Against parsimony: three easy ways of complicating some categories of economic discourse. *Am Econ Rev* 74(2):89–96
- Jeffrey R (1974) Preferences among preferences. *J Philos* 71(13):377–391
- Jones P, Cullis J (2002) Merit want status and motivation: the knight meets the selfloving butcher brewer and baker. *Public Financ Rev* 30(2):83–101
- Kirchgässner G (2017) Soft paternalism merit goods and normative individualism. *Eur J Law Econ* 43(1):125–152
- McLure CE (1968) Merit wants: a normative empty box. *Finanzarchiv* 27(2):474–483
- Musgrave RA (1957) A multiple theory of budget determination. *Finanzarchiv* 17(3):333–343
- Musgrave RA (1959) *The theory of public finance: a study in public economic*. McGraw-Hill Book Company, New York
- Musgrave RA (1969) Provision for social goods. In: Margolis J, Guitton H (eds) *Public economics*. McMillan, London, pp 124–144
- Musgrave RA (1987) Merit goods. In: Eatwell J, Milgate M, Neuman P (eds) *The New Palgrave: a dictionary of economics*. Macmillan, London, pp 452–453
- Pazner EA (1972) Merit wants and the theory of taxation. *Public Financ* 27(4):460–472
- Rawls J (1971) *A theory of justice*. Harvard University Press, Cambridge, MA
- Roskamp KW (1975) Public goods merit goods private goods Pareto optimum and social optimum. *Public Financ* 30(1):61–69
- Salanié F, Treich N (2009) Regulation in Happyville. *Econ J* 119(537):665–679
- Schelling TC (1984) Self-command in practice in policy and in a theory of rational choice. *Am Econ Rev* 74(2):1–11
- Sen AK (1977) Rational fools: a critique of the behavioural foundations of economic theory. *Philos Public Aff* 6(4):317–344
- Sunstein CR, Thaler RH (2003) Libertarian paternalism is not an oxymoron. *University Chicago Law Rev* 70(4):1159–1202
- Thaler RH, Sunstein CR (2008) *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven
- Ver Eecke W (2001) The concept of a merit good: the ethical dimension in economic theory and the history of economic thought or the transformation of economics into socio economics. *J Socio-Econ* 27(1):133–153
- Wenzel HD, Wiegard W (1981) Merit goods and second-best taxation. *Public Financ* 36(1):125–139
- West EG, MacKee M (1983) De gustibus Est disputandum the phenomenon of merit wants revisited. *Am Econ Rev* 7(5):1110–1121

Method of Merchants

► Lex Mercatoria

Mises, Ludwig von

Karl-Friedrich Israel

Faculty of Economics and Business

Administration, Humboldt-Universität zu Berlin,
Berlin, Germany

Department of Law, Economics, and Business,
University of Angers, Angers, France



Mises, Ludwig von, Fig. 1 Ludwig von Mises

Abstract

Ludwig Heinrich Edler von Mises (September 29, 1881, in Lemberg – October 10, 1973, in New York City) was a classical liberal philosopher, sociologist, and one of the most influential adherents to the Austrian school of economics. He made major contributions to the epistemology of the social sciences and to many areas of general economics, especially the fields of value theory, monetary theory, and business cycle theory. In his habilitation thesis of 1912, *The Theory of Money and Credit*, he laid down the foundations of what would later become Austrian business cycle theory. The whole body of Misesian economics which is based on praxeology, the rational investigation of human decision making, is comprised in his 1949 *Human Action: A Treatise on Economics*. Among his disciples were such noteworthy economists as Friedrich August von Hayek (1899–1992) and Murray N. Rothbard (1926–1995).

Life, Work, and Influence of Ludwig von Mises

Life, Family, and Personal Background

Ludwig von Mises (Fig. 1), son of Arthur Edler von Mises and Adele Landau, was born in the city of Lemberg, Galicia, in the former

Austro-Hungarian Empire (today Lviv, Ukraine) on September 29, 1881. He is the older brother of the famous Harvard mathematician Richard Martin von Mises (1883–1953). His youngest brother Karl died of scarlet fever at the age of 12 in 1903.

In 1881 Emperor Franz Joseph granted Ludwig's great grandfather Meyer Rachmiel Mises a patent of nobility and the right for him and his lawful offspring to bear the honorific title "Edler" (Hülsmann 2007, p. 15). Ludwig then was the first member of his family to be born a nobleman. The whole Mises family was heavily involved in the construction and financing of Galician railways after the 1840s. So it came to pass that Ludwig's father was a construction engineer for the Czernowitz railway company in Lemberg before moving to Vienna no later than 1891 (Hülsmann 2007, p. 21).

Ludwig entered the *Akademische Gymnasium* of Vienna in 1892. There, he was given an education stressing among other things the classical languages, Latin and ancient Greek. Reading Virgil, he found a verse that he chose to be his motto for life: "Tu ne cede malis sed contra audentior ito" (translated: Do not give in to evil, but proceed ever more boldly against it) (Mises 2009, p. 55). Much later he would point out the importance of the classical literature, and in particular of the ancient Greeks, for the emergence of liberal social philosophy in *The Anti-Capitalistic Mentality*

(Mises 1956; Hülsmann 2007, p. 34). After having completed the Gymnasium in May 1900, when his major interests were in politics and history, he enrolled in the Department of Law and Government Science at the University of Vienna.

Initially, Mises studied under the Marxist sociologist and economist Carl Grünberg (1861–1940) who was an adherent of the German historical school. At the age of only 20, Mises published his first scholarly work in one of Grünberg's books on the peasants' liberation and agrarian reforms in the Bukovina (Hülsmann 2007, p. 68). He graduated with the first *Staatsexamen* in 1902 in history of law from the University of Vienna. In 1906, after completion of the military service, he passed the second and third *Staatsexamina* in law and government science. Thereafter, he was awarded a doctorate degree, *doctor juris utriusque*, from the same university by passing his juridical, political science, and general law exams.

Mises in the course of his studies shifted gradually away from the influence of historicism which was at the time the predominant method in the social sciences in the German-speaking world. Highly influenced by economists Carl Menger (1840–1921) and his follower Eugen Böhm von Bawerk (1851–1914), he published his habilitation thesis in 1912 under the title *Theorie des Geldes und der Umlaufsmittel* (translated in 1934 as *The Theory of Money and Credit*). In fact, it was Menger's *Grundsätze der Volkswirtschaftslehre* (*Principles of economics*, Menger 1871) and Böhm-Bawerk's *Kapital und Kapitalzins* (*Capital and Interest*, von Böhm-Bawerk 1890) originally published in 1884, two foundational works of the Austrian school of economics, that had a tremendous impact on Mises's thinking. Mises wrote in his *Memoirs* that it was through Menger's book that he became an economist (Mises 2009, p. 25). After his habilitation, he worked as a civil servant for the Austrian chamber of commerce and became a *Privatdozent*, an unsalaried lecturer, at the University of Vienna – the highest academic rank he would ever achieve in his native Austria.

During the First World War, Mises served as an artillery officer in the Austro-Hungarian army and

spent several months at the front. It was only in December 1917 that he was ordered to join the department of war economy in the War Ministry of Vienna. He resumed his teaching activities at the University in the spring of 1918. Among his students at that time was Richard von Strigl (1891–1942) who became an important Austrian economist of the interwar period. In 1919, he finished a manuscript under the working title *Imperialismus* which would contain his analysis of the causes of the Great War and the political challenges for postwar Europe as well as his personal war experiences. The book was eventually published under the title *Nation, Staat und Wirtschaft* (translated in 1983 as *Nation, State and Economy*, von Mises 1919). This book, although less well known today, established Mises as the foremost champion of classical liberalism in the German-speaking world and eventually in all of Europe (Hülsmann 2007, p. 300).

Mises made a new personal acquaintance with the famous Max Weber (1864–1920), who came to teach at the University of Vienna in September of 1917. Their professional relationship was characterized by mutual respect and admiration. Weber had a remarkable impact on Mises's writings on the methodological and epistemological problems of the social sciences in the 1920s. On the other hand, Weber praised Mises's monetary theory as “the most acceptable” of the time (Hülsmann 2007, p. 288).

Only a few years later, in 1922, Mises published “the most devastating analysis of socialism yet penned” (Hazlitt 1956, p. 35), *Die Gemeinwirtschaft* (translated in 1936 as *Socialism* (von Mises 1951), a book in which he gives a detailed and thorough explanation of the problem of economic calculation under socialism. In 1927, his *Liberalismus* (translated in 1962 as *Liberalism* (von Mises 1985) was first published). Two years later, *Kritik des Interventionismus* (translated in 1977 as *A Critique of Interventionism*, von Mises 1996) appeared, an anthology of articles he wrote in the early 1920s.

Throughout the 1920s until 1934, Mises held a weekly private seminar in his office at the chamber of commerce to which students and scholars not only from Vienna but from all over Europe

and the whole world were attracted. Among the numerous participants were Austrian economist Gottfried Haberler (1900–1995), Frenchman Francois Perroux (1903–1987), American economist Frank Knight (1885–1972), and even four Japanese economists (Itschitani, Midutani, Otaka, Takemura) (Hülsmann 2007, p. 674). However, as one of the best-known students and colleagues of Mises, one has to mention Friedrich August von Hayek who attended the seminar on a regular basis. Hayek became the director of the *Austrian Institute for Business Cycle Research* that Mises had established (Hülsmann 2007, p. 454). Later, Hayek would take a position at the *London School of Economics* which was an important step in his academic career.

The political developments made life in Vienna increasingly unpleasant for Mises. Being of Jewish descent and an ardent critic of socialism, he had to flee Vienna in 1938 after occupation and annexation of Austria by the German national socialists. With great foresight he had already accepted a position at the Geneva-based *Institut de Hautes Études Internationales* (Graduate Institute of International Studies) in 1934, where he became visiting professor. There, under the company of William E. Rappard (1883–1958) and Paul Mantoux (1877–1956) who led the school together for about 20 years, he had some of the most productive and fruitful years of his scholarly career, culminating in the publication of his opus magnum *Nationalökonomie* in 1940. In 1938, he would marry his longtime companion Margit Serény in Geneva.

As history moved on, Geneva as well seemed no longer to be a safe place for Mises and his wife, due to the rising threat of national socialism and ultimately the outbreak of World War II. Mises was high on the list of wanted men. In March 1938, unidentified men broke into Mises's apartment in Vienna. A few days later, the Gestapo confiscated his books, correspondences, personal records, as well as other belongings. At the end of the war, the Red Army found his files in a train in Bohemia. They were brought to a secret archive in Moscow, where they would be rediscovered in 1991, 18 years after his death (Ebeling 1997). Mises never saw his documents

again. He thought that they had been destroyed during the war.

Ludwig and Margit von Mises decided to flee Europe. After a long and arduous trip, they arrived at the docks of New York City on the third of August, 1940. At almost 60 years of age, Mises had to start a new life. He quickly sought contact with potential supporters. His brother was already professor at Harvard University. Henry Hazlitt (1894–1993), journalist at the *New York Times* and a great admirer of Mises, turned out to be very helpful. With the financial support of the *William Volker Fund*, he arranged a visiting professorship for Mises at *New York University* (NYU). Mises would remain a visiting professor for more than 20 years until he retired in the Spring of 1969, as the oldest active professor in the United States (Rothbard 1988, p. 44). His salary during that period would always be paid from private funds.

Mises set up a seminar at NYU in the spirit of his Vienna private seminar, where he gathered a diverse group of journalists, businessmen, scholars, and young university and even high school students. Hans Sennholz (1922–2007) and Louis Spadaro (1913–2008) were his first doctoral students in the United States. Other seminar members were William Peterson (1921–2012), George Reisman (born 1937), his classmate Ralph Raico (born 1936), Israel Kirzner (born 1930), and most notably Murray N. Rothbard. According to his wife, Mises encouraged all of his students to pursue scholarly work, hopeful that one of them might develop into a second Hayek (Mises 1976, p. 135). He might have found his “second Hayek” in Rothbard, although both Hayek and Rothbard developed Misesian ideas in quite different directions.

Mises's first books written in English appeared in 1944, *Omnipotent Government* (von Mises 1944a) and *Bureaucracy* (von Mises 1944b), both published by Yale University Press. *Human Action*, the extended English edition of his opus magnum *Nationalökonomie*, was published in 1949. It contains the culmination of Misesian economic theory. Among several other books and numerous articles, he published his last great work in 1957, *Theory and History* (von Mises

1957), a philosophical treatise that builds a bridge between economic theory and human history and explains the true relation between those two disciplines (Rothbard 1988, p. 110). Having devoted his whole life to the enhancement of economic theory and the social sciences, as well as the promotion of peace and individual liberty, Ludwig von Mises died at the age of 92 on October 10, 1973, in New York City.

Work and Influence

When Ludwig von Mises entered university in 1900, the social sciences in the German-speaking world, including economics, were predominantly influenced by historicism – ideologically a forerunner of positivism. Gustav Schmoller (1838–1917), under whom Mises’s first university teacher Carl Grünberg studied, was the leading intellectual of the German historical school. Mises recognized fundamental flaws within this school of thought and the academic tendencies of his time, namely, the state orientation:

It was my intense interest in historical knowledge that enabled me to perceive readily the inadequacy of German historicism. It did not deal with scientific problems, but with the glorification and justification of Prussian policies and Prussian authoritarian government. The German universities were state institutions and the instructors were civil servants. The professors were aware of this civil-service status, that is, they saw themselves as servants of the Prussian king. (Mises 2009, p. 7, as cited in Rothbard 1988, p. 51)

Mises was interested in positive social sciences, not in the promotion of political measures by the government. He was convinced that there are absolutely and uncompromisingly, in his words “apodictically,” true statements in the realm of social sciences that are not mere tautologies or conventions – an idea that historicists and positivists would decidedly reject. Mises on the other hand rejected the relativism of the historicists. He first found such uncompromisingly true statements in Menger’s *Grundsätze der Volkswirtschaftslehre*, which is widely considered to be the foundational work of the Austrian school of economics. This book made Mises an economist and an “Austrian.” Mises, among the other Austrian economists of the second and third

generation after Menger, dwelled on this epistemological and methodological point the most vehemently. He considered economics to be an “a priori” science, rooted in the broader discipline of the rational investigation of human behavior and decision making that he would term praxeology.

The Misesian body of economic theory is derived from the undeniable fact that human beings exist and act, that is, they are consciously employing means to attain ends. One cannot argue that human beings do not act, that they do not employ means to attain ends, since such an argument would precisely be an action as described above. By acting, human beings make choices, they demonstrate preferences, and they engage in exchanges. They make value judgments – they decide for one opportunity and they forgo others, that is, they pay a price. Mises shows that all fundamental economic concepts, such as value, exchange, preferences, prices, costs, and gains, are inherent to human action:

Action is an attempt to substitute a more satisfactory state of affairs for a less satisfactory one. We call such a willfully induced alteration an exchange. A less desirable condition is bartered for a more desirable. What gratifies less is abandoned in order to attain something that pleases more. That which is abandoned is called the price paid for the attainment of the end sought. The value of the price paid is called costs. Costs are equal to the value attached to the satisfaction which one must forego in order to attain the end aimed at.

The difference between the value of the price paid (the costs incurred) and that of the goal attained is called gain or profit or net yield. Profit in this primary sense is purely subjective, it is an increase in the acting man’s happiness, it is a psychological phenomenon that can be neither measured nor weighed. There is a more and a less in the removal of uneasiness felt; but how much one satisfaction surpasses another one can only be felt; it cannot be established and determined in an objective way. A judgment of value does not measure, it arranges in a scale of degrees, it grades. It is expressive of an order of preference and sequence, but not expressive of measure and weight. Only the ordinal numbers can be applied to it, but not the cardinal numbers. (Mises 1998, p. 97)

As one can see from the above quote, Mises’s value theory is completely subjective, with the consequence that he would not only refuse the

Marxian labor theory of value but also the welfare economics of a Léon Walras (1834–1910). For Mises, value or utility is something that only the individual attaches to a good. It cannot be measured and therefore interpersonal utility comparisons and concepts like “aggregate welfare” are invalid. On this issue, Mises takes the same line as Czech economist Franz Cuhel (1862–1914) (Rothbard 1988, pp. 19 and 59). The only common denominator of individual preferences and evaluations, through which heterogeneous goods and services can be compared somewhat objectively, is the price system of a free market. However, this is not to be confused with measuring utility in terms of money prices.

From this insight, Mises derives his critique of central planning. In a centrally planned economy, where the means of production are not privately owned and therefore are not sold and bought on the market, there are no market prices for production or capital goods. Consequently, there is no way for the central planners to determine which production processes or which combinations of capital resources, for the production of some desired good, are economically efficient and in line with consumer preferences. In short, economic calculation is impossible under socialism. *Socialism: An Economic and Sociological Analysis*, the relevant book for this topic, is to a large extent built on an earlier article that started the socialist calculation debate, *Economic Calculation in the Socialist Commonwealth*. According to Friedrich August von Hayek, this debate and in particular Mises’s contribution had a remarkable impact on the social scientists and economists of his generation, who predominantly favored socialism over the free market (Hayek 1981).

In *A Critique of Interventionism*, Mises provides us with a very distinctive clarification of another crucial question concerning political economy. If socialism is inherently dysfunctional, what about a middle-of-the-road solution – a mixed economy that is neither a complete free market system nor full-blown socialism? Can we combine the benefits of both systems in a suitable way? Mises’s answer is no. Such a system of interventionism would be inherently unstable. For every problem that the government detects

and attempts to solve through intervention into the economy, it will most of the time not only not solve the problem but also create new ones. After each intervention the government can then either take the initial intervention back or go on to intervene further into the economy. Mises argued that interventionism is a slippery slope and ultimately leads to socialism. The only alternative is provided by a genuine free market system (Bagus 2013). This idea of the self-reinforcing character of interventionism has been picked up by Hayek in his best-selling popular book *The Road to Serfdom* (von Hayek 2005). Hayek does not deny the impact that Mises had on his political philosophy, although some economists and philosophers who see themselves more in line with the Misesian tradition and the Rothbardian interpretation thereof, such as Walter Block, criticized Hayek sharply for his “lukewarm” defense of laissez-faire capitalism (Block 1996).

Although his critique of the state is extensive, Mises would not deny the necessity of the state altogether. As unmistakably explained in *Liberalism*, he assigns a unique task to the state, namely, the enforcement of law and in particular the protection of private property. Mises thereby remains in the tradition of classical liberalism (Hoppe 2013). It was Murray N. Rothbard, certainly Mises’s most influential American student, who would enrich the Misesian analysis of the state with his theory of rational ethics developed in his philosophical treatise *The Ethics of Liberty* (Rothbard 2003) and push it to its ultimate conclusion: the state should be abolished. Rothbard spearheaded the anarcho-capitalistic movement in the United States that today is more vibrant than ever.

When it comes to pure economics, Mises probably made his most important contributions in the area of monetary theory and business cycle theory. In his habilitation thesis, *The Theory of Money and Credit* (von Mises 1912), he elaborates on Menger’s account of the origins of money out of barter trade to solve a challenge that has been laid down to the Austrian economists by Karl Helfferich (1872–1924) in 1903. In his work *Money*, Helfferich correctly pointed out that the “Austrians,” despite their comprehensive

microeconomic analysis of markets and prices for goods and services, had not yet managed to solve the problem of money. Money had been treated separately from the rest of economics in a “macro-box,” independently of utility, value, and relative prices (Rothbard 1988, p. 55). When trying to explain the value of money, one ended up in a circular argument. Money is a special good that is demanded not for consumption but for exchange – in the present or some point in the future. It is demanded because it has an exchange value, but it has its exchange value only because it is demanded. This circular reasoning posed a problem that Mises was able to solve by incorporating the time dimension into this interdependency (Hülsmann 2013).

The demand for money today is determined by the exchange value of money in previous periods. The demand for money in previous periods was determined by its exchange value in still earlier periods, and so the chain of reasoning goes backward in time. Do we end up in an infinite regress? No, because at some point back in time, the money good has been demanded, as any other good, for its value in consumption. At this point, we are back in a barter economy. This is the starting point. In a transition from direct to indirect exchange, in order to overcome the problem of the double coincidence of wants, some goods and eventually one single good will become universally accepted as a medium of exchange. This good will become money. Traditionally, precious metals such as gold and silver have played this role, since they are rare, homogeneous, highly divisible, transportable, and durable. Mises’s explanation, which we call the *Regression Theorem*, “was a remarkable achievement, because for the first time, the micro/macro split that had begun in English classical economics with Ricardo was now healed” (Rothbard 1988, p. 57). Money was incorporated into the rest of economic theory and had no longer to be treated separately. For a more detailed outline and interpretation of Mises’s monetary theory, the interested reader may have a look at *Theory of Money and Fiduciary Media* edited by Guido Hülsmann (2012), an anthology of essays in celebration of the centennial of Mises’s major work.

The next fundamental contribution by Ludwig von Mises is his theory of business cycles that emerged out of his monetary theory. Thinking about the way money evolves on the market according to Menger and Mises, one recognizes that our current monetary system is quite special, in the sense that our money is not backed by any real good. We are living in a fiat money system. Our money is money by government decree. However, this regime has been transformed into what it is today by government intervention out of a gold-backed monetary system, and its mere existence does therefore not disprove the Mengerian-Misesian account. That it is advantageous for any government to possess the monopoly over the production of an unbacked currency is quite obvious. It enables the government to create money out of thin air through credit expansion and to lower interest rates artificially. The Keynesian view holds that this is a necessary political tool for anti-cyclical economic stabilization and therefore the cure of the business cycle which is inherent to free market capitalism. Mises, on the contrary, sees the root cause of the business cycle not in the free market itself but precisely in artificial credit expansion. Mises’s explanation is as follows.

He built his theory out of three preexisting components, the business cycle model of the Currency School, the differentiation between the “natural rate of interest” and the “bank rate of interest” by Swedish economist Knut Wicksell (1898), and the Böhm-Bawerckian capital and interest theory (Rothbard 1988, p. 63; Hülsmann 2007, Chapter 6). Mises argues that pumping money into the economy by expanding credit and lowering interest rates under the “natural” time preference level cause excess malinvestments in capital goods industries.

In Mises’s view, the interest rate is not an arbitrary number that should be interfered with. Instead, it is the price that tends to accommodate the roundaboutness of production processes or investment projects to the available subsistence fund in the economy. Usually, interest rates fall, when consumers save more and thereby increase the subsistence fund. However, in the case of artificial credit expansion, the decrease in interest

rates and the subsequent excess investments are not justified or covered by real savings. Therefore, some of these investments, sooner or later, have to be liquidated, when it turns out that the subsistence fund in the economy is too small to finish them all.

Note that artificial credit expansion is not a purely political phenomenon that can only be brought about by the state. It can and did happen on the market, on behalf private banks. However, it is only through the institutionalization of credit expansion, by establishing a central bank system, that it has reached today's magnitude.

Hayek won the Nobel Memorial Prize in economics in 1974, the year after Mises's death. He received it precisely for his contributions to business cycle theory, that is, the work he did in the 1920s and 1930s as an "ardent Misesian" (Rothbard 1988, p. 112). This decision by the Nobel Prize committee gave a boost to Austrian economics in general and Austrian business cycle theory in particular. According to Rothbard, it marked a "revival of Austrian economics."

The financial crisis of 2007 has attracted more attention to the Austrian explanation of booms and busts. Subsequent to this event, Ron Paul, presidential candidate in the United States, gained increased recognition for his critique of the Federal Reserve System and his call to abolish it. His arguments are essentially backed by Misesian ideas.

Mises's "last great work" and "by far the most neglected" was *Theory and History*. Yet, it "provides the philosophical backstop and elaboration of the philosophy underlying *Human Action*" (Rothbard 1985). He had a lot of opponents concerning his political recommendations, but it was his methodological uniqueness that made him an academic outlier for his whole life. In *Theory and History*, Mises provides the philosophical justification for considering the social sciences, including economics, as a completely different branch than the natural sciences. The subjects of investigation in the social sciences are human beings, with minds, who have preferences and make choices – who have goals and try to attain these goals. They act purposefully and they change their minds constantly. Human action does not follow mechanical and quantifiable

laws like atoms or molecules do in physics. The empirical approach to the social sciences is ignorant of the uniqueness and the individuality of human decisions and the environment in which each decision is made. Human action is not replicable. By pointing out this insight, Mises made the case for methodological dualism. In the natural sciences, the researcher looks at data of repeated identical experiments in which all relevant variables can be controlled to find and isolate causal relationships. In economics, we already know the ultimate cause: humans act to attain their ends. This knowledge is not hypothetical. It is in Mises's words apodictically true and is not subject to empirical falsification – and neither are logical deductions from this primordial fact. Mises has been called unscientific and mystical for calling economics an "a priori" science, in the same way as mathematics is an "a priori" science. He has been criticized for being ignorant of economic history. His rebuttal is contained in *Theory and History*. It is the historicist-positivist-empiricist economist who overlooks the unique character of historical events by trying to draw generalized conclusions from observable data and thereby ignores the single common feature of all economic data: purposefully acting individual human beings. For Mises, it is not history that can provide us with a reliable theory. It is only by means of a theoretical framework that we can make sense out of historical data.

Today, the ideas of Ludwig von Mises and his intellectual followers are promoted more effectively than ever before by the *Ludwig von Mises Institute* in Auburn, Alabama. It was founded in 1982 by Llewellyn H. Rockwell Jr., Burton Blumert, and Murray N. Rothbard. Its website, www.mises.org, makes a large number of books, journal articles, and other writings available for free. It is worth a look for every interested reader. There exist a number of professional journals and periodicals published by the institute, including *The Journal of Libertarian Studies* (1977–2008), *The Review of Austrian Economics* (1987–1998), and *The Quarterly Journal of Austrian Economics* (since 1998).

Today, without formal ties among them, there exist more than 14 Mises Institutes worldwide,

including those in Belgium, Switzerland, Germany, Portugal, Sweden, Finland, Czech Republic, Romania, Poland, Russia, Canada, Brazil, Ecuador, and Japan. There also exist a Spanish Language Institute and a Mises Institute Europe.

References

- Bagus P (2013) Mises' Staats- und Interventionismuskritik, published in Ludwig von Mises – Leben und Werk für Einsteiger, Finanzbuchverlag
- Block W (1996) Hayek's road to Serfdom. *J Libert Stud* 12(2):339–365. Available online http://mises.org/journals/jls/12_2/12_2_6.pdf
- Ebeling R (1997) Mission to moscow: the lost papers of Ludwig von Mises. *Liberty Magazine* 10(5). A presentation by R. Ebeling at Universidad Francisco Marroquin is available online: http://newmedia.ufm.edu/gsm/index.php/Mission_to_Moscow:_Discovering_the_%22Lost_Papers%22_of_Ludwig_von_Mises,_and_their_significance
- Hazlitt H (1956) Two of Ludwig von Mises' most important works. In: Sennholz M (ed) *On freedom and free enterprise – essays in honor of Ludwig von Mises*. D. Van Nostrand. Available online <http://mises.org/document/3327/On-Freedom-and-Free-Enterprise-Essays-in-Honor-of-Ludwig-von-Mises>
- Helfferich K (1903) *Das Geld*, 1th edn. Hirschfeld, Leipzig
- Hoppe H-H (2013) *Ludwig von Mises und der Liberalismus*, republished in Ludwig von Mises – Leben und Werk für Einsteiger, Finanzbuchverlag
- Hülsmann J-G (2007) Mises – the last knight of liberalism. Ludwig von Mises Institute, Alabama. Available online <http://mises.org/document/3295/Mises-The-Last-Knight-of-Liberalism>; for a presentation by the author see <http://www.youtube.com/watch?v=h9BgKL5kX4U>
- Hülsmann J-G (ed) (2012) *Theory of money and fiduciary media*. Ludwig von Mises Institute, Alabama. Available online <http://mises.org/document/7018/Theory-of-Money-and-Fiduciary-Media>
- Hülsmann J-G (2013) Mises' Geldtheorie, published in Ludwig von Mises – Leben und Werk für Einsteiger, Finanzbuchverlag
- Menger C (1871) *Grundsätze der Volkswirtschaftslehre*, Wilhelm Braumüller – k.k. Hof- und Universitätsbuchhändler. Available online http://docs.mises.de/Menger/Menger_Grundsaeetze.pdf; for the English translation see <http://mises.org/document/595/Principles-of-Economics>
- Rothbard MN (1985) Preface to theory and history by Ludwig von Mises. Ludwig von Mises Institute, Alabama
- Rothbard MN (1988) *The essential von Mises*. Ludwig von Mises Institute, Alabama. Available online <http://mises.org/document/3081/The-Essential-von-Mises>
- Rothbard MN (2003) *The ethics of liberty*. New York University Press, New York and London
- von Böhm-Bawerk E (1890) *Capital and interest*. Macmillan, London/New York. Available online <http://mises.org/document/164/Capital-and-Interest>; for the German original see <https://archive.org/details/kapitalundkapita01bh>
- von Hayek FA (1981) *Foreword to socialism: an economic and sociological analysis*. Liberty Fund, Indianapolis
- von Hayek FA (2005) *The road to Serfdom with the intellectuals and socialism*. Institute for Economic Affairs. Available online <http://mises.org/document/2402/The-Road-to-Serfdom>
- von Mises L (1912) *Theorie des Geldes und der Umlaufsmittel*. Verlag von Duncker und Humblot, München/Leipzig. Available online <http://mises.org/document/3298/Theorie-des-geldes-und-der-Umlaufsmittel>; for the English edition see <http://mises.org/document/194/The-Theory-of-Money-and-Credit>
- von Mises L (1919) *Nation, Staat und Wirtschaft. Beiträge zur Politik und Geschichte der Zeit*. Manzsche Verlags- und Universitäts-Buchhandlung, Vienna/Leipzig; eager as nation, state, and economy (trans: Leland B). New York University Press, New York (1983). Available online <http://mises.org/document/1085/Nation-State-and-Economy>
- von Mises L (1944a) *Bureaucracy*. Yale University Press, New Haven. Available online <http://mises.org/document/875/Bureaucracy>
- von Mises L (1944b) *Omnipotent government: the rise of the total state and total war*. Yale University Press, New Haven. Available online <http://mises.org/document/5829/Omnipotent-Government-The-Rise-of-the-Total-State-and-Total-War>
- von Mises L (1951) *Socialism – an economic and sociological analysis*. Yale University Press, New Haven. Available online <http://mises.org/document/2736/Socialism-An-Economic-and-Sociological-Analysis>; for the German original see <http://mises.org/document/3297/Die-Gemeinwirtschaft>
- von Mises L (1956) *The anti-capitalistic mentality*. Van Nostrand, Princeton. Available online <http://mises.org/document/1164/The-AntiCapitalistic-Mentality>
- von Mises L (1957) *Theory and history – an interpretation of social and economic evolution*. Yale University Press, New Haven. Available online <http://mises.org/document/118/Theory-and-History-An-Interpretation-of-Social-and-Economic-Evolution>
- von Mises M (1976) *My years with Ludwig von Mises*. Ludwig von Mises Institute, Alabama. Available online <http://mises.org/document/3199/My-Years-with-Ludwig-von-Mises>
- von Mises L (1985) *Liberalism: in the classical tradition*. The Foundation for Economic Education, Inc. Irvington-on-Hudson, New York 10533. Available online <http://mises.org/document/1086/Liberalism-In-the-Classical-Tradition>; for the German original see <http://mises.org/document/5452/Liberalismus>
- von Mises, L (1996) *A critique of interventionism*, Irvington-on-hudson. Foundation for Economic Education, New York. Available online <http://mises.org/document/877/Critique-of-Interventionism-A>

- von Mises L (1998) Human action (the Scholar's edition). Ludwig von Mises Institute, Alabama. Available online <http://mises.org/document/3250/Human-Action>; for the German predecessor *Nationalökonomie* see <http://mises.org/document/5371/Nationalökonomie-Theorie-des-Handelns-und-Wirtschaftens>
- von Mises L (2009) Memoirs. Ludwig von Mises Institute, Alabama. Available online <http://mises.org/document/4249/Memoirs>
- Wicksell K (1898) Geldzins und Güterpreise. Gustav Fischer Verlag, Jena. The English version is available online <http://mises.org/document/3124/Interest-and-Prices>

Money Laundering

Donato Masciandaro
Department of Economics, Bocconi University,
Milan, Italy

Abstract

Money laundering is any activity aimed to hide the origin and/or the destination of a flow of money in order to reduce the probability of sanctions. In order to describe the economics of money laundering, the starting point is the definition of its microeconomic foundations, which are based on the existence of a rational actor who derives revenues from a criminal activity and from the assumption that his/her expected utility depends on four key elements: expected revenues, laundering costs, likelihood of being caught, and magnitude of the sanction.

The micro basis of the money laundering can explain its macroeconomic effects. Money laundering can function as a multiplier mechanism of the weight of the illegal sector in a given territory or country. In order to prevent and combat the polluting effects of money laundering, an effective regulation has to be designed, based on a correct incentives alignment between the supervisors and the financial intermediaries.

Definition

Money laundering is any activity aimed to hide the origin and/or the destination of a flow of money in order to reduce the probability of sanctions.

Only recently the economic analysis has developed a peculiar focus on the financial issues related to the study of crime, thus far completely absent (Masciandaro 2007; Unger and Van der Linde 2013). In this entry, a simple framework to explain the micro foundations of the money laundering activities in order to analyze its macroeconomic effects on the relationship between the illegal activities and the economic sector as a whole will be offered.

Microeconomics

The emphasis on the study of money laundering has progressively increased, recognizing its potential role in the development of any crime that generates revenues and or in financing a crime. In fact, the conduct of any illegal activity may be subject to a special category of transaction costs, linked to the fact that the use of the relative revenues increases the probability of discovery of the crime and therefore incrimination. The analysis zooms on the economics of concealing illegal sources of revenues, but the same reasoning can be applied in discussing the hidden financing of illegal activities (money dirtying).

However, we should stress that in terms of economic analysis, the financing of illegal activities (money dirtying) is a phenomenon not perfectly equivalent from the recycling of capital (money laundering).

The money dirtying resembles money laundering in some respects and differs from it in others. The objective of the activity is to channel funds of any origin to individuals or groups to enable illegal acts – for example, terrorism – or activities. Again in this case, an organization with such an objective must contend with potential transaction costs, since the financial flows may increase the probability that the financed crime will be discovered, thus leading to incrimination. Therefore, an effective money dirtying action, an activity of concealment designed to separate financial flows from their destination, can minimize the transaction costs.

The main difference between money laundering and money dirtying is in the origin of the

financial flows. While in the money laundering process the concealment regards capitals derived from illegal activity, the illegal actor or organization can use both legal and illegal fund for financing their action.

Money laundering and money dirtying may coexist. A typical example is the financing of terrorism with the proceeds from the production of narcotics. In those specific situations, the importance of the transaction costs is greater, since the need to lower the probability of incrimination concerns both the crimes that generated the financial flows and the crimes for which they are intended. The value of a concealment operation is even more significant.

The definition of money laundering points up its specialness with respect to other illegal or criminal economic activities involving accumulation and/or reinvestment. Now, given that the conduct of any illegal activity may be subject to a special category of transaction costs, linked to the fact that the use of the relative revenues increases the probability of discovery of the crime and therefore incrimination, those transaction costs can be minimized through an effective laundering action, a means of concealment that separates financial flows from their origin.

In other words, whenever a given flow of purchasing power that is potential – since it cannot be used directly for consumption or investment as it is the result of illegal accumulation activity – is transformed into actual purchasing power, money laundering has occurred.

Focusing our attention on the concept of incrimination costs enables us to grasp not only the distinctive nature of this illegal economic activity but also its general features. The definition here adopted maintains basic unity among three aspects that, according to other points of view, represent three different objects of the anti-laundering action: the financial flows (layering), the wealth and goods intended as terminal moments of those flows (integration), and the principal actors, or those who have that wealth and goods at their disposal (placement).

In this general framework of analysis, there will always be an agent who, having committed

a crime that has generated accumulation of illicit proceeds, moves the flows to be laundered, so as to subsequently increase his/her financial assets, by investment in the legal sector or re-accumulation in the illegal sector. The agent can be an individual or a criminal organization.

By criminal organization, we mean a group of individuals and instrumental assets associated for the purpose of exclusively exchanging or producing services and goods of an illicit nature or services and goods of a licit nature with illicit means or of illicit origin.

In general, following the classic intuition à la Becker, it can be claimed that the choices of an economic agent to invest his/her resources in illegal activities – as money laundering is – will depend, *ceteris paribus*, on two peculiar magnitudes, given the possible returns: the probability of being incriminated and the punishment he/she will undergo if found guilty.

Now, to undertake money-laundering activity, the agent possessing liquidity coming from illegal activity will decide whether to perform a further illicit act, in a given economic system – i.e., money laundering – assessing precisely the probability of detection and relative punishment and comparing that with the expected gains, net of the economic costs of this money-laundering activity.

The choice of the agent requires that the crime in question, and the relative production function, be basically autonomous with respect to other forms of crime, those that generated the revenues in the accumulation phase.

Assigning a monetary utility to the crime of money laundering, by giving it a unitary expression, actually summarizes the value of a series of more general services that stimulate the growth of demand for money-laundering services on the part of the agents that accumulate illegal resources. Money laundering, in fact, produces for its users:

1. The economic value, in the strict sense, of minimizing the expected incrimination costs, transforming into purchasing power the liquidity deriving from a wide range of criminal activities (transformation); transformation, in turn, produces two more utilities for the criminal agent.

2. The possibility of increasing his/her rate of penetration in the legal sectors of the economy through the successive phase of investment (pollution).
3. The possibility of increasing the degree to which the criminal actors and organizations are camouflaged in the system as a whole (camouflaging).

Having defined the micro choices in the most general terms possible, we can now investigate the macro effects of money-laundering choices.

Macroeconomics

To define a macro model of the accumulation-laundering-investment process, we focus on the behavior of a general criminal sector that derives its income from a set of illegal activities and that, under certain conditions, must launder the income to invest it. We will highlight the role of money laundering as an overall multiplier of the criminal sector endowment.

Let us assume – as in Masciandaro 1999 and then in Barone and Masciandaro 2011 – that in a given economic system, there is a criminal sector that controls an initial volume of liquid funds ACI , fruit of illegal activities of accumulation. Let us further suppose that, at least for part of those funds, determined on the basis of the optimal microeconomic choices already discussed in the previous pages, there is a need for money laundering. Without separating these funds from their illicit origin, given the expected burden of punishment, they have less value. Money-laundering activity is therefore required.

To underscore the general nature of the analysis, we can claim that the demand for money-laundering services could be expressed – distinguishing the different potential components of a criminal sector according to their primary illegal activity – by organized crime in the strict sense, by white collar crime, or by political corruption crime, also considering the relative crossover and commingling.

Each laundering phase has a cost for the criminal sector, represented by the price of the

money-laundering supply. The price of the money-laundering service, all other conditions being equal, will depend on the costs of the various money-laundering techniques. Let us suppose that in the money-laundering markets, the criminal sector is price taker and that the cost of money-laundering cR is constantly proportional to the amount of the illicit funds; designating the costs with c , both regulatory and technical, we can write:

$$cR = cACI \quad (1)$$

If the first laundering phase is successful, the criminal sector may spend and invest the remaining liquid funds $(1 - c)yACI$ in both legal economic activities (investment) and illicit activities (re-accumulation).

The trend to use specialist money launderers – with their explicit or implicit fees – is increasing. These operators use their expertise to launder criminal proceeds. In general, the professionals may be witting or unwitting accomplice; but in any case, the buildup of the overall procedure represents a cost.

We assume that in general, the money-laundering procedures represent a cost for organized crime, notwithstanding it is well known that the criminal groups can implement legal businesses also for the concealment of their illegal proceeds and these businesses can produce profits. As it will be evident below, the smaller the money-laundering costs will be, the greater the multiplier effect.

The criminal sector will spend part of the laundered liquidity in consumer goods, equal to d , while a second portion will be invested in the legal sectors of the economy, for an amount of f , and then a third portion, equal to q , will be reinvested in illegal markets (giving, of course, $d + f + q = 1$).

On the one side share of illegal funds needs to be spent: minimizing incrimination risks comes at a price; the criminal sector has to pay a price. On the other side, we suppose that a share of dirty money will be reinvested in the illegal market without concealment. For example, in all illicit services, cash is by definition the currency of

choice, running in a closed circuit separate from the legitimate market.

Reinvestment in criminal markets is a distinctive feature of actual organized crime groups, given their tendency toward specialization. Organized crime tends to acquire specialist functions to augment their illegal businesses.

If the criminal sector makes investment choices according to the classical principles of portfolio theory, indicating with $q(r, s)$ the amount of laundered funds reinvested in illegal activities, with r the actual expected return on the illegal re-accumulation, and with s the relative. Finally, we can assume that the re-accumulation of funds in the illegal sector requires their laundering only in part, thus indicating with the positive parameter y the portion of illegal re-accumulation that requires laundered liquidity.

The criminal sector reinvests both clean and dirty money, and then a new flow of illegal liquidity will be created. The illegal revenues will be characterized again by incrimination costs, which will generate a new demand for money-laundering services. It will be therefore equal to:

$$(1 + r)(1 - c)^2 q y^2 ACI \quad (2)$$

The crucial assumption is that both the lawful investment and part of the unlawful re-accumulation require financing with “clean” cash. This assumption can be supported by the presence of rational, informed operators in the supply of services to the criminal sector for the illegal re-accumulation or by rationality of the criminal himself/herself, who wishes to minimize the probability of being discovered.

Repeating infinite times, the demand for money-laundering services, which each time encounter a parallel supply, with the values of the parameters introduced remaining constant, the total amount of financial flows generated by money-laundering activity AFI will be equal to:

$$AFI = \frac{yACI(1 - c)}{1 - yq(1 - c)(1 + r)} = mACI \quad (3)$$

with $0 < c, q, y < 1$.

The flow AFI represents the overall financial endowment generated by the money-laundering activity, and m can be defined as the multiplier of the model. Doing comparative static exercises, it is easy to show that the amount of liquidity laundered will increase as the price of the money-laundering service declines:

$$\frac{\partial AFI}{\partial c} = - \frac{ACI y}{[1 - yq(1 - c)(1 + r)]^2} < 0 \quad (4)$$

Therefore, the more effective the money-laundering action, the greater the cash flows available to the criminal sector for reinvestment, illegal and legal, will be. Money laundering represents the multiplier of the illegal sector.

Regulation

Summing up the key features of money laundering, the micro foundations have been based on the existence of a rational criminal actor or organization, which derives revenues from an illicit or criminal activity. The criminal actor or organization wants to maximize the expected utility of his/her or its illicit proceeds. The expected utility increases with the average return, and it decreases with the costs for laundering, with the probability of being caught, and with the severity of the sanction when being caught.

Further, the macroeconomic effects of money laundering have been captured using a multiplier model. Money laundering triggers a multiplier process, which ends up with higher laundering and higher criminal activities. Money laundering harms the economy. Therefore, it can be useful to design and implement a regulation to prevent and combat the money-laundering phenomena.

On this respect, the starting point has to be to recognize that the banking and financial industry usually play a pivotal role for the development of the criminal sector as a preferential vehicle for money laundering.

The main actors are on the one hand the regulatory agents, who want to combat money

laundering, and on the other hand the financial intermediaries, who can be either honest and compliant or dishonest and noncompliant. Asymmetric information and principal-agent problems are typical for this market. The design of anti-money-laundering regulations must take four aspects into consideration: the difference in information assets between the individual intermediaries and the agency, the non-verifiability of bankers' efforts to comply, the costliness of that effort for the intermediaries, and the non-verifiability of the influence of the effort on the performance of the regulation.

To deal with such difficulties, the best way of analyzing the regulatory issues is to implement a principal-agent methodology, managing the incentive problems that arise at least in a three-layer hierarchy, which includes public authorities, financial institutions, and supervisors. It is possible to show – Dalla Pellegrina and Masciandaro (2009) – that the under asymmetric information, the effectiveness of the regulation depends on three crucial conditions.

First, the participation constraint of the financial institutions requires that the incentive scheme is well balanced, meaning that both rewards and penalties must be defined, in order to minimize the difference between the private costs in implementing a regulatory model and the public gains in collecting useful information against money laundering.

Second, excessive fines per se do not necessarily provide incentives to the financial institutions to improve their action. In particular, given the incentive scheme of the financial institutions, the quality of the supervision can be a good substitute for the severity of punishments: the more effective the (potential) supervisory action in monitoring ex post the money-laundering risk, the more likely the effectiveness of the financial institutions in building up ex ante their monitoring models.

Third, other things being equal, if the cost of supervision depends on its quality, also the efficiency of the supervisory agencies matters. Again, the importance of the quality and the efficiency of supervision can be particularly relevant.

References

- Barone R, Masciandaro D (2011) Organized crime, money laundering and legal economy: theory and simulations. *Eur J Law Econ* 32(1):115–142
- Dalla Pellegrina L, Masciandaro D (2009) The risk based approach in the new European anti-money laundering legislation: a law and economics view. *Rev Law Econ* 5(2):932–952
- Masciandaro D (1999) Money laundering: the economics of regulation. *Eur J Law Econ* 7(3):225–240
- Masciandaro D, Takats E, Unger B (2007) *Black finance. The economics of money laundering*. Edward Elgar, Cheltenham
- Unger B, Van der Linde D (eds) (2013) *Research handbook of money laundering*. Edward Elgar, Cheltenham

Multiple Tortfeasors

Samuel Ferey
CNRS, BETA, University of Lorraine,
Nancy, France

Definition

Multiple tortfeasor issues cover cases where the loss is jointly caused by several people acting in a common purpose or not. A lot of examples illustrate these situations including accident law, environmental damage, product liability, etc. The law and economics literature has devoted much attention on these situations in order to understand the properties of different liability rules. Two levels of discussion should be distinguished: first, the negligence/strict liability debate, and second, the joint or no-joint liability rules meaning that the victim may get compensation back by any of the tortfeasors. The article surveys the most important results in law and economics and insists on the fact that multiple tortfeasor cases lead to paradoxes which challenge the very basis of the tort law and the economic rationality.

From One Defender to Several

At the very beginning, tort law has been one of the most promising fields in law and economics: Coase, Calabresi, and Posner were particularly interested in the efficiency of liability rules.

However, the economic basic model implied only one tortfeasor (Posner 1973; Coase 1960; Calabresi 1970). The scope of this model was limited and did not cover more complex cases where several tortfeasors jointly cause the harm suffered by the victim, what is called multiple tortfeasor issues. Sometimes, cases of multiple tortfeasors refer to harm due to a common purpose of several people. However, the scope of multiple tortfeasors is broader and covers also harm due to independent wrongdoers' behaviors. Many examples of multiple tortfeasors come in mind: accident law, environmental law, products liability, etc. the law and economics literature about multiple tortfeasor cases started at the beginning of the 1980s and does not necessarily converge on clear principles. The reason is that two levels of discussion are involved: the first is about the regime of liability (strict liability or negligence), the second about the joint or non-joint liability.

To handle with this complexity, legal scholars have fallen into the habit of forming their reasoning based on typologies to classify different cases (Hart and Honoré 1985; Landes and Posner 1980, 1983). The first case is simple: several tortfeasors independently act at the same time, and their joint behaviors simultaneously lead to harm (two polluters pouring toxic waste in a river). The second case is sequential: a first tortfeasor causes at time t_1 a first harm that is enhanced at time t_2 by another wrongdoer, etc. The third case covers the example of contribution from the victim which has participated to its own harm (e.g., because he has not taken sufficiently care level). The fourth case implies uncertain causation: harm occurred without knowing with certainty who is the true wrongdoer (two hunters shooting while only one bullet actually harms the victim). In the following, we focus on the third first cases, and we do not deal with uncertain causation (see the article ► “Causation” in the *Encyclopedia*).

These complex cases have been extensively studied by legal scholars (Hart and Honoré 1985; Wright 1985; Stapleton 2013) because they challenge the usual way to consider liability, causation, and torts. Indeed, imagine two people lighting a fire in a forest. Suppose that these two fires merge together and destroy the house of the

victim. In such a case of overdetermined causation, the “but for test” criterion advocated by the law is useless: each of the tortfeasors could escape from his responsibility by pretending that, without his own action, the harm would have occurred anyway. It would be unfair and inefficient to follow the black letter of the “but for test,” and these cases require other legal solutions.

From the economic point of view, these cases are also difficult to handle with. We will survey one of the most key elements of the findings of law and economics literature on this topic. First, we insist on the efficiency and incentives aspect. We also deal with the fairness of the sharing rules applied by judges to apportion damage. Then, we compare three rules from an economic perspective: several liability, joint and several liability, and channeling liability. Last, we conclude with dealing with some of the most striking paradoxes raised by multiple tortfeasor cases.

Incentives, Efficiency, and Fairness

As we will see, in many instances, multiple tortfeasor cases challenge the result of efficiency which holds for a simple one victim/one tortfeasor case. First, it is now well-known that one of the most important findings in law and economics is lost when several tortfeasors are implied. In a paper published in 1981 by Aivazian and Callen, and entitled “Coase Theorem and the Empty Core,” these authors take a simple example where two polluters cause harm to a victim. Under the general conditions of the Coase theorem – perfect delineation of property rights and zero transaction costs – the result of invariance and efficiency cannot be proved. The reason is that the core of the game may be empty meaning that there is no stable allocation which could emerge from negotiation: victim and tortfeasors negotiate again and again without any convergence of their offers.

In case of positive transaction costs, liability rules are studied to implement an efficient output in terms of care, activity, and actual harm (Shavell 2004). But when it comes to efficiency in cases of multiple tortfeasors, it is difficult to conciliate two principles: the principle that every tortfeasor should pay for the harm he is responsible

for and the fact that the total amount of compensatory damage be equal to the harm. Most of times, if principle 2 is followed, underdeterrence is expected. On the contrary, if principle 1 is followed, it requires implementing a sanction which is above the amount of harm. This result is easy to show in case of strict liability. Assume that two tortfeasors 1 and 2 cause harm to a victim. The objective function of social cost to be minimized is $x_1 + x_2 + p(x_1, x_2) \cdot h$, with x_1 the cost of care of 1, x_2 the cost of care of 2, h the amount of harm, and p the probability of harm. To induce optimal care, it would be necessary that each tortfeasor bears damage of h . This is impossible insofar as h will be split among the two liable tortfeasors (usually, the compensatory damage is equal to harm, no more, no less). Other authors have elaborated on this issue by providing alternative rules. For example, Miceli and Segerson imagine a rule where each tortfeasor is responsible for the marginal damage that it causes. In that case, efficiency is achievable, but it requires that the total amount jointly paid by tortfeasors be above the loss caused. Miceli and Segerson consider that a mixed system requiring on the one hand, the recovery of the loss to the victim and, on the other hand, a fine to be paid to the state could achieve efficient incentives (Miceli and Segerson 1991; Miceli 1997).

While under strict liability, tortfeasors are threatened to pay less damage than required by efficiency; under negligence rule, incentives are different. If the efficient care standard (x_1^*, x_2^*) is implemented by the law, a tortfeasor will not be held responsible if his level of care is above the efficient level. The best way to analyze the behaviors of tortfeasors is to use game theory and to wonder whether the equilibrium of the game exists and reaches social efficiency. If tortfeasor 1 has chosen x_1^* , the best response of tortfeasor 2 is to choose the efficient level x_2^* . If not, he would pay for the entire damage. The situation is symmetric for the agent 1. As Shavell states, "The superiority of the negligence rule has to do with the fact that, by the nature of the negligence rule, each party is threatened with damages equal to the *entire* harm h when other parties act optimally; whereas under strict liability, each

party is threatened with a lesser amount" (Shavell 2007). This result holds when tortfeasors are jointly liable (they expect to pay for the entire loss) but does not hold in case of several liability where they would expect to bear only a share of the loss (Kornhauser and Revesz 1989). In the previous paragraphs, we have implicitly supposed that all the tortfeasors are solvable. In case of insolvency, the previous results may be altered, and we will focus on this issue in part III.

In a different perspective, it is possible to have a more fairness-oriented point of view and be interested in the contribution of tortfeasors. The idea is to have a scheme of payment which is in line with the causal contribution of each tortfeasor. Some legal scholars have elaborated on this idea of causal contribution (Stapleton 2013). In law and economics, comparative causation is such a rule (Parisi and Singh 2010). This idea dates back to Calabresi for whom negligence has an unexpected effect to promote a one or no liability. He suggests splitting the loss among parties following their causal contribution. Parisi and Singh follow this argument and have carefully studied the properties of a causal contribution rule mixed with a negligence rule. They show that such a rule has interesting properties in terms of optimal care level and optimal activity level and consider that it is a fair and equitable rule. In a different perspective focused on ex post issues, other scholars use cooperative game theory and consider that the Shapley value is a good candidate to provide an evaluation of the contribution of each tortfeasor involved in the case of indivisible harm. Moreover, the Shapley value has axiomatic properties interesting for the law, and it could be shown that most of principles advocated in the *Restatement of Tort* (American Law Institute 2012) in the USA to apportion damage are close to the elementary principles of the Shapley value (Ferey and Dehez 2016a).

Several, Joint and Several, and Channeling Liabilities

Even though it was possible to clearly apportion harm among tortfeasors, a separate issue raises about how the victim could trigger responsibility against each of them. In most countries, three

rules are implemented by the law: several liability, joint and several liability, and channeling liability (Faure 2016). Several liability states that each tortfeasor is responsible for his share and the victim has a separate claim against each of them but only for their respective shares. Several liability has two consequences: first, the victim has to bear the costs of several suits to recover all the compensation of his harm from all the tortfeasors separately; second, the risk of insolvency of one of the tortfeasors is bear by the victim.

Several and joint liability is different: the victim has a claim for the entire compensatory amount against any of the tortfeasors and gets all its recovery damages back from him. Then, in most of legal systems, the tortfeasor who has compensated the victim has a claim against the other tortfeasors to get their respective shares of responsibility back. Joint and several liability has two consequences: first, litigation costs are decreasing for the victim (and increasing for the defender); second, it is expected that the victim will choose the most wealthy defender to get its recovery back (following a deep pocket argument), and the risk of insolvency is bear by the solvable defender, not by the victim. A lot of fields in torts around the world are organized under a joint and several liability. In Europe, for example, the *European Principles of Tort Law* (see European Group on Tort Law 2005) mainly advocate joint and several liability.

Third, channeling liability indicates ex ante who among the potential tortfeasors is fully responsible and most of times without any claim against the others. Channeling liability is implemented by some international conventions. For example, the *Paris Convention on Nuclear Liability* has decided that nuclear liability is channeled to operators. The relative properties of the three rules have been extensively discussed in the law and economics literature which focus on four main topics: litigation costs, incentives, insolvency, and insurance (Faure 2016). We could add that the choice of one of the three rules (several, joint and several, or channeling liability) may be influenced by pressure groups without any global efficiency concerns.

Regarding litigation costs, the three rules are different. Channeling liability makes recovery easier for the victim: the identity of the party to be sued is certain, and only one lawsuit has to be filed. On the contrary, several liability seems costly for the victim who is required to pay for as much suits as the number of tortfeasors. Joint and several liability is intermediate: the litigation costs for the victim to get compensation back is relatively low but increase for the deep pocket tortfeasor which is required to pay for lawsuits against the other tortfeasors. And it could be expected that enforcement of the law will be more effective if litigation costs are low for the victim.

Regarding insolvency, the three rules have also different properties. The insolvency issue arises when at least one tortfeasor is not able to pay his share of damage back. Full solvency of tortfeasors seems to be a strong hypothesis which does not hold in many cases, and insolvency should be expected due to the high amounts that firms have to pay. First is the distributive issue: who should bear the risk of insolvency, the victim or the other tortfeasors? Most of legal systems consider that it is fair enough to avoid the undercompensation of the victim due to insolvency. The risk of non-recovery is shifted to the solvable tortfeasor.

This point has consequences on incentives. Under joint and several liability with insolvent actors and channeling liability, a tortfeasor may be asked to pay beyond his own share without any possibility of recovery from other defenders. The consequences on the level of care and the level of activity taken by the parties are not clear. First, the insolvent tortfeasor does not take into account the risks he created above the total value of its assets. Up to a certain point, he is underdeterred. Second, the solvent tortfeasor may be aware of this insolvency and internalizes the fact that, in case of several and joint liability, he should pay for the other's share. The main effect is on the activity level which may be reduced or eliminated to avoid any liability. This effect is called the crushing effect of liability. Third, the opposite effect may also happen. Intuitively, the reasoning insists on the fact that, in case of limited solvency, liability rule has no deterrent effect above a certain level

anymore (it has the same effect as statutory caps on liability). Up to a certain point, a domino effect occurs: the share of the insolvent tortfeasor to be paid by the remaining solvable actor is so huge that liability has no deterrent effect anymore on the solvent agent (Kornhauser and Revesz 1989). Lastly, a monitoring policy may be expected from parties: a tortfeasor who knows that he is likely to pay for others has an incentive to control and monitor other parties to be sure that they decide optimal levels of care and/or activity. However, in some cases, a defender has no way to monitor the level of care taken by other.

The three rules have strong consequences on insurance markets. The fact that one tortfeasor pays more than his share means that its insurance company is not able to calculate a risk premium anymore. Virtually, the insurance company will guarantee all the losses caused by all the participants of the market. Here is the risk of a crushing effect on the insurance market. The argument could be mitigated by the diversification of risks: It is unlikely that the same insurance company be concerned by all the cases implying multiple tortfeasors (Faure 2016).

Paradoxes

We would like to conclude by several paradoxes raised by multiple tortfeasor cases which illustrate significant challenges: overdetermined causation, aggregation issues, and offsetting benefits.

A first series of paradoxes is well-known by legal scholars: the overdetermined cases (Hart and Honoré 1985). It is the case of the two fires merging and destroying a property. In that cases, the black letters of the classical criterion of causation leads to irrational results. That is why other views on causation are proposed by legal scholars like the NESS test – necessary element of a sufficient test – (Wright 1985) or the causal contributions (Stapleton 2013). These criteria of causation may be modeled in an economic framework (Ferey and Dehez 2016b) and explain the reasons of the paradoxes.

Implying several people, multiple tortfeasor cases lead also to aggregation puzzles. In a recent paper, Porat and Posner have provided a general framework to capture aggregation issues. Regarding torts, they insist on the general consequences

of the burden of proof. As many scholars, they consider that preponderance of evidences rule may be translated in terms of a probability superior to 50%. Imagine that a victim suffers two harms: the first one is due to an error of the surgeon of an hospital with a probability of 0,6 (event A) and a second one by a nurse of the same hospital (event B) with a probability of 0,6 (Porat and Posner 2012). Then, the victim files a lawsuit against the hospital for the two harms. Without aggregation, liability of the hospital will be held for the two harms because the probability of event A and of event B is superior to 0,5. However, it could be said that the probability that the hospital be responsible jointly for the two harms is only of 0,36 ($0,6 \times 0,6$) which is the probability of the conjunction of the two events. The paradox raises from the fact that the preponderance of evidence leads to consider that the hospital should not be held responsible for the two harms. Such paradoxes are particularly interesting regarding multiple tortfeasors. A general typology of all the different types of aggregation issues is provided in Porat and Posner (2012).

The third type of paradox is implied by what is called “offsetting benefit”. Offsetting benefit issues are broader than multiple tortfeasor issues, but they are particularly concerned with. Porat and Posner have extensively dealt with these cases (Porat and Posner 2014). A simple example of the problem is the following. Imagine a victim driving his car to reach the airport where he has to take a flight. On the road to the airport, he is harmed by a wrongdoer and cannot take his plane. It then appears that the plane crashes. Following strict causation reasoning, it could be said that the wrongdoer has caused an injury but has also saved the life of the victim. Should we consider that the benefits due to the accident has to be taken account in the calculus of the compensatory damage? In this example, the victim could even compensate the tortfeasor due to the behavior of a potential second tortfeasor. There are different ways to solve these types of cases, and Porat and Posner provide some elementary principles to guide courts.

To conclude, it is clear that a lot needs to be done to better understand all the aspects of

multiple tortfeasors issues. We have shown how the difficulties of this topic are both philosophical, legal, and technical, implying different aspects in terms of efficiency and fairness and let a room open for political choices.

Cross-References

- [Causation](#)
- [Strict Liability Versus Negligence](#)

References

- Aivazian VA, Callen JL (1981) The Coase theorem and the empty core. *J Law Econ* 24:175–181
- American Law Institute (2012) Restatement (Third) of the law of torts: liability for physical and emotional harm. Executive Office, American Law Institute, Saint Paul
- Calabresi G (1970) The costs of accidents. Yale University Press, New Haven
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- European Group on Tort Law (2005) Principles of European tort law. Text and commentary. Springer, Vienna
- Faure M (2016) Attribution of liability: an economic analysis of various cases. *Chic-Kent Law Rev* 91:603–636
- Ferey S, Dehez P (2016a) Multiple causation, apportionment and the Shapley value. *J Leg Stud* 45:143–171
- Ferey S, Dehez P (2016b) Overdetermined Causation, Contribution and the Shapley Value. *Chic-Kent Law Rev* 91:637–658
- Hart HLA, Honoré T (1985) Causation in the law, 2nd edn. Oxford University Press, Oxford
- Kornhauser LA, Revesz RL (1989) Sharing damages among several tortfeasors. *Yale Law J* 98:831–884
- Landes WM, Posner RA (1980) Joint and multiple tortfeasors: an economic analysis. *J Leg Stud* 9: 517–555
- Landes WM, Posner RA (1983) Causation in tort law: An economic approach. *J Leg Stud* 12:109–134
- Miceli TJ (1997) Economics of the law. Oxford University Press, Oxford
- Miceli TJ, Segerson K (1991) Joint liability in torts: marginal and infra-marginal efficiency. *Int Rev Law Econ* 11:235–249
- Parisi F, Singh R (2010) The efficiency of comparative causation. *Rev Law Econ* 6:219–245
- Porat A, Posner EA (2012) Aggregation and law. *Yale Law J* 122:2–69
- Porat A, Posner EA (2014) Offsetting benefits. *Va Law Rev* 100:1165–1209
- Posner RA (1973) Economic analysis of law. Little Brown, Boston
- Shavell S (2004) Foundations of economic analysis of law. Harvard University Press, Cambridge
- Shavell S (2007) Liability for accidents. In: Shavell S, Polinsky M (eds) Handbook of law and economics, vol 1. Elsevier, Oxford, pp 142–182
- Stapleton J (2013) Unnecessary causes. *Law Quart Rev* 129:39–65
- Wright RW (1985) Causation in tort law. *Calif Law Rev* 73:1735–1828

Music Royalty Rates for Different Business Models: Lindahl Pricing and Nash Bargaining

Marcel Boyer^{1,2} and Anne Catherine Faye³

¹Department of Economics, Université de Montréal, Montréal, QC, Canada

²Toulouse School of Economics, Toulouse, France

³Analysis Group, Montréal, Canada

Definition

The Lindahl Equilibrium and Nash Bargaining Solution serve as useful and complementary analytical tools that provide guiding principles for the determination of appropriate music royalty rates in the current digital era, which poses challenges and pitfalls for the pricing of information goods, most notably musical works and sound recordings.

Introduction

We discuss the economic concepts of Lindahl Equilibrium and Nash Bargaining Solution as useful and complementary analytical tools in the context of the current digital era, which poses challenges and pitfalls for the pricing of information goods, most notably musical works and sound recordings. These concepts of modern economic theory are more than ever useful, even required, to provide the guiding principles underlying the determination of appropriate music royalty rates.

There is a general agreement between rightsholders, copyright users, and tariff-setting organizations that tariffs should be “fair and equitable” to both rightsholders and users and reflect the value or benefits the users derive from copyrighted works.

From an economic perspective, a “fair and equitable” level of transactions and compensation is equivalent to the level of transactions and compensation that emerges from a competitive market where willing buyers and willing sellers freely negotiate and settle transactions. Such willing buyers and sellers are assumed to be price-takers or devoid of market power. They would agree on transactions up to the point where the marginal value of an additional transaction for buyers is equal to the marginal cost of that additional transaction for sellers. From that perspective, the tariff to be paid for the use of copyrighted musical works and sound recordings should be based on the amount that users would willingly pay if they were transacting in a well-functioning competitive market.

The Copyright Board of Canada, the US Copyright Royalty Board, and other tariff-setting institutions represent a surrogate for a competitive market where the price that would prevail for copyrighted works is determined, if such a market existed and operated efficiently. To determine a competitive price for copyrighted works, these institutional bodies typically consider relevant proxies or indicators on buyers, sellers, prices of substitute products and services, industry characteristics, and virtual or simulated competitive processes such as auctions, etc. The challenges encountered in emulating a competitive market are numerous, some of the most salient being the fact that musical works and sound recordings are information goods and that digital technologies are profoundly changing the copyright landscape.

Information Goods and Digitization: Lindahl Pricing and Nash Bargaining

From an economic perspective, musical works and sound recordings are “information goods.” An information good is a good whose

consumption by one consumer does not prevent its consumption by others, a characteristic of “information” under different forms. In other words, once a musical work or sound recording is created, it can be “consumed” by all, as one person’s use of a musical work or sound recording does not prevent its simultaneous or subsequent use or consumption by someone else. See for example, Varian (1999) and Bakos et al. (1999).

Although related, the “information good” nature of musical works and sound recordings and the digitization of music are two different but equally challenging factors in the current copyright setting. The first one relates to the indefinite sustainability of a product, as one’s consumption does not destroy the consumed unit, which remains fully and unabatedly available for everyone else now and in the future, hence making musical works and sound recordings akin to non-decaying assets. The second one relates to the much-reduced dissemination cost of that product or asset in a digital context. See Boyer (2018).

In such a context, striking the right balance between creators’ rights and users’ interests is a difficult and multifaceted endeavor for different reasons: (a) musical as well as other cultural products such as literary works are costly to create and (b) digital technologies have significantly reduced the marginal cost of dissemination of those copyrighted works.

Due to these factors, pricing copyrighted musical works and sound recordings, with the objective of achieving proper compensation for creators, i.e., a creators’ right, *and* maximal dissemination of such goods, i.e., a users’ right, requires a move away from the usual analysis aimed at setting a product’s price equal to its marginal cost. Setting a price equal to marginal cost would clearly not enable the proper, fair, or competitive compensation of sellers, producers, and creators. Considering that musical works and sound recordings are permanent assets rather than perishable consumption goods raises questions regarding the proper price concepts to use: the same unit can be sold and resold infinitely many times.

The determination of relevant copyright tariffs rests not so much on the cost of creation,

which underlies the supply function of new musical works and sound recordings, but rather on the value of such goods for the users. Indeed, the Copyright Board of Canada (2002) acknowledged that “the important notion in information industries [is that] pricing tends to be based on the value to the buyer, not on cost to produce.”

The supply function of musical works and sound recordings in the rightsholders’ collective repertoire (stock) is horizontally constant at a price p just above 0 (infinite price elasticity at $p = \varepsilon$). That is because the marginal cost, which underlies the supply function, is quasi-zero for the rightsholders: zero marginal cost of reproduction and dissemination and zero marginal opportunity cost, those musical works and sound recordings being resalable an indefinite number of times. Setting the tariff equal to this marginal cost would generate no revenue for the seller, here the rightsholders, thereby failing to meet a central objective of the copyright institution. Insofar as the compensation of rightsholders must come through transactions, not from public subsidies or grants, the setting of tariffs is cast in a second best economic framework.

In this context, the proper price equilibrium is the Lindahl (1958) equilibrium. It is a generalization of the competitive equilibrium for public or information goods, according to which, given a stock of musical works and sound recordings, different users pay for access to a relevant repertoire, prices equal, or proportional to their respective derived marginal values.

The notion of Lindahl pricing was developed to characterize both the optimal or efficient quantity of a public good and a way to finance its production cost. It requires that the price paid by a given buyer or user of a public (or information) good be positively linked to the value or the amount of satisfaction derived from the consumption of that public good, as everyone is assumed to consume, at least virtually, the whole good. The user deriving a greater (marginal) value from the same marginal unit would pay a higher price. As long as the sum of those prices is above the marginal cost, additional units should be produced. The efficient quantity and quality level are

reached when the sum of users’ marginal values (prices) is equal to the marginal cost. The Supreme Court of Canada has asked the Copyright Board of Canada to consider such analytic framework in setting tariffs.

Technological Neutrality

The Supreme Court of Canada, in its 2015 landmark decision introducing the principles of technological neutrality and balance, stated:

One element of just compensation is an appropriate share of the benefit that the user obtains by using reproductions of their copyright-protected work in the operation of the user’s technology. That just compensation must be valued, however, in accordance with the principle of technological neutrality. While highly unlikely, where users are deriving the same value from the use of reproductions of copyright-protected works using different technologies, technological neutrality implies that it would be improper to impose higher copyright-licensing costs on the user of one technology than would be imposed on the user of a different technology. To do so would privilege the interests of the rights holder to a greater degree in one technology over the other where there is no difference between the two in terms of the value each user derives from the reproductions.

The converse is also true. Where the user of one technology derives greater value from the use of reproductions of copyright-protected work than another user using reproductions of the copyright-protected work in a different technology, technological neutrality will imply that the copyright holder should be entitled to a larger royalty from the user who obtains such greater value. Simply put, it would not be technologically neutral to treat these two technologies as if they were deriving the same value from the reproductions. (Supreme Court of Canada 2015, paragraphs 70, 71)

The Supreme Court asked the Copyright Board to consider that if a user with a given technology derives more value from a product, say a repertoire of musical works and/or sound recordings assets, than another user with a different technology, the former user should pay more than the latter, even if both use the same repertoire. In other words, the two users would access the same asset at different prices.

The Supreme Court thus recognized that, for efficiency and fairness reasons, users with different business models may be subject to different

tariffs for access to the same stock of musical works and sound recordings. There is a direct link between technological neutrality and Lindahl principles for pricing information goods.

Applying Lindahl principles is arduous and full of potential pitfalls especially in the context of copyright royalties. First, determining the (marginal) values different users assign to an asset is challenging. Second, many tariffs are expressed as a proportion or percentage of an accounting base rather than as a price per unit.

Finding the (marginal) values different users attach to or are willing to pay for an asset could be achieved through simulated auctions. The Copyright Board has alluded to such auctions in its 2002 Digital Pay Audio Decision as follows: “The Board finds it useful to comment on the ‘simulated auction’ approach which was discussed at the hearing. This scenario calls for setting the price of music at what one would be willing to pay to acquire a monopoly over DPA. That approach must be set aside because it focuses again on profitability at the expense of all else. This being said, the exercise is not without merit, if only because it highlights the important notion that in information industries, pricing tends to be based on the value to the buyer, not on cost to produce” (p. 8).

Many tariffs are expressed as a proportion or percentage of an accounting base due to the challenging factors involved in pricing copyrighted works on a per unit basis. Expressing royalties as a percentage of an accounting base, such as the user’s revenues, serves different purposes. It allows for the following: (a) savings in transaction costs, including assessing and verifying copyright payments; (b) immunity to accounting manipulations if the rate base is well chosen and valid; and (c) risk-sharing between rightsholders and users as a whole because it may be difficult or impossible to know *ex ante* which user will be successful in turning profitable access to a stock of musical works and sound recordings. However, expressing royalty payments as a percentage of an accounting base does not provide any indication about the price of copyrighted work. This follows because a percentage is not a price.

A Percentage Is Not a Price

A major source of confusion in royalty rate setting is the trap of “considering a percentage as a price.” There is no reason to believe that the proper price to be paid for the same inputs, namely, the right to access a stock of musical works and sound recordings, would correspond to the same percentage of revenues irrespective of the characteristics of the underlying industries under consideration. In fact, using the same percentage will usually lead to subsidizing one industry at the expense of another because, under the economic law of one price, the same price for the same inputs used in two different industries will, in general, represent quite different percentages of the value of the industry outputs (or revenues). Moreover, Lindahl equilibrium calls for different prices for different users or consumers of the same good.

The following example can help illustrate why a percentage is not a price. Suppose an apartment in a low income housing project is valued at \$100,000. Suppose also an apartment of the same size and configuration in a high-income housing project is valued at \$1,000,000, i.e., ten times more. Then suppose that the same quantity and quality of paint is used for both apartments. As expected, the cost of painting the two apartments would be the same as the law of one price applies to the input market (i.e., paint). In economics, the law of one price states that similar products in the same market should sell at similar prices. Let us suppose that the actual cost of paint for the high value apartment is \$1,000, i.e., 0.1% of the value of the apartment. If one takes the 0.1% as the “price” of the paint to be paid for the low-value apartment, one would get a price in dollar terms of $0.1\% \times \$100,000 = \100 or one-tenth of the cost of the same amount and quality of paint used for painting the high-value apartment. In fact, the cost of paint expressed as a percentage of the value of the two apartments would be quite different: 1.0% for the low-value apartment and 0.1% for the high-value apartment, a difference by a factor of ten for the same quantity and quality of paint.

As we further discuss below, regulatory bodies responsible for setting royalties on the use of copyrighted works tend to avoid transferring

percentages from one industry to another without proper adjustments. In doing so, they aim to ensure that percentages reflect similar prices for the same rights.

Making those different percentages correspond to Lindahl prices and percentages requires additional adjustments.

Balance Between Rightsholders and Users

The Supreme Court of Canada has stated that achieving balance between the rights of creators and users requires the Copyright Board take both rights into account in the determination of royalty rates by considering “respective contributions of, on the one hand, the risks taken by the user and the investment made by the user, and on the other hand, the reproductions of the copyright-protected works to the value enjoyed by the user” (Supreme Court of Canada 2015, paragraph 75). The principle of balance is related to the economic concept of a bargaining game, which in turn is related to the concepts of competitive equilibrium and negotiated price.

A bargaining game is an analytical framework of game theory that seeks to model a situation in which there is a conflict of interest between different agents who have the opportunity to reach a mutually beneficial agreement but may nevertheless veto any agreement. When “there is more than one course of action more desirable than disagreement for all individuals but conflicting views over which course of action to pursue, then negotiations to resolve the conflict will take place” (Osborne and Rubinstein 1994, p. 117).

In this context, the framework of a bargaining game appears as a tool to model the negotiation process and to determine how different agents who contribute to the creation of a given surplus or value added can share such value among them. A solution to the bargaining game is a sharing formula that specifies what percentage of the total value added each agent receives at the end of the bargaining game. The total value to be shared, the agents’ outside options (i.e., alternative options in case they do not reach an

agreement, including the no-investment option), as well as their bargaining power all play a role in the determination of the solution.

John F. Nash Jr., the 1994 laureate of the Nobel Memorial Prize in Economic Sciences, proposed a solution to such a bargaining problem, known as the Nash Bargaining Solution (NBS). The NBS is derived from four axioms defined as desired or reasonable properties. Namely, the solution should be (Pareto) efficient, symmetric, immune to equivalent reformulations of players’ objectives, and immune to irrelevant alternatives. The efficiency property simply states that the solution should leave no money on the table. The symmetry property states that if two agents are in similar positions or have similar capacities to negotiate, they should be treated equally, that is, obtain “similar” shares of the pie. Given that each party is rational, well advised, and controls an essential input, each holds similar power to negotiate and veto any solution and therefore can be considered to be equally capable of negotiating and affecting the solution. The other two axioms are more technical and not developed here.

Nash shows that there is only one solution to the above bargaining problem (Osborne and Rubinstein 1994, p. 307). In other words, there is a unique sharing formula that satisfies the four axioms: once each agent is properly compensated for the cost it incurs to sit at the negotiation table, the residual monetary value would or should be shared 50-50 between the agents. Given the reasonableness of its axioms, the Nash bargaining solution (NBS) is a powerful result. It says that whatever the negotiations conduct and/or process, i.e., no matter how the negotiation is conducted, the expected ultimate end-point or result can only be the NBS. As mentioned above, the NBS is closely related to the concepts of competitive equilibrium (willing buyer, willing seller) and negotiated price, two concepts regularly referred to in hearings before tariff-setting organizations.

The balance required by the Canadian Supreme Court to be considered in rate-setting must be reached in a context where standard perfectly competitive conditions do not prevail. Regarding copyrighted musical works and sound recordings, a standard competitive market with

individual buyers and sellers with no market power and symmetric information does not generally exist. However, implicit negotiations can take place between representatives of the parties before a rate-setting body emulating a competitive framework and solution. A negotiated price would then be considered the solution of the bargaining game involving the different parties. Such negotiated prices are grounded in economics as accounting for (and therefore balancing) the relative contributions and alternate options of the parties involved in the negotiations. The Nash Bargaining Solutions of negotiations between different users and the rightsholders call, under the 50-50 rule, for different royalty payments for different users with different business models for packaging and transmitting or distributing music.

Hence, the principles of technological neutrality (Lindahl pricing) and balance of rights (Nash Bargaining) in the context of information goods require that different business models be charged different royalty rates for the same access to musical works and sound recordings. In other words, efficiency and fairness in royalty setting run against business model neutrality.

The Case of Digital Pay Audio and Commercial Radio in Canada

In Canada, payments for communication rights of musical works and sound recordings in the digital pay audio industry (DPA) as well as in the commercial radio industry (CR) are measured as percentages of a rate base (the total revenue of the user). In fact, the Copyright Board (CB) has at numerous times compared and used as proxies royalty rates expressed as percentages of revenues by carefully applying relevant adjustments to account for differences among the industries considered.

Applying the above reasoning to the CR and DPA industries, similar prices for the same access rights to the same repertoire of musical works and sound recordings, when expressed in percentage terms, will represent a higher percentage in the lower value added DPA industry, where the main input is music, than in the higher value added CR

industry, where musical works and sound recordings are one input among many including news, weather, or traffic reports and on-air personalities. In other words, similar prices would yield different percentages.

In its decision, the Supreme Court of Canada instructs the Board to consider that if a user, such as DPA services, derives more value from the product, say the repertoire of SOCAN Collective of authors and composers (musical works) and Re:Sound Collective of performers and makers (sound recordings) than another user, such as a CR operator, the former user should be asked “to pay more” than the latter.

In Canada, the commercial radio royalty rates for musical works and sound recordings are both 4.2% (before adjustments for repertoire) of the user’s revenues. Suggesting to increase the combined CR royalty rate of 8.4% to 10.6% for DPA on the basis that the latter uses 25% more music per given period, would clearly not satisfy the technological neutrality principle insofar as the DPA industry derives greater value from musical works and sound recordings relative to CR, which generates revenues from other programs.

One possibility would be to, first, derive the level of royalties generated by the 8.4% rule and, second, express it as a percentage of revenues generated by music only; indeed, the commercial radio industry depends less on music than the digital pay audio services industry. On average, music format radio stations broadcast music content during 80% of programming time and other content (news, survival programming, on-air personalities, or talk) during the remaining 20%, while DPA services play music 100% of the time, that is, 25% more. However, the CB has repeatedly expressed the view that on-air talent is more important than music to radio stations. In its 2002 DPA decision, it stated: “Radio may be designed around the use of music and musical genres but as a cost, and (probably) as a drawing card, on-air talent is far more important” (Copyright Board of Canada 2002, p. 8). If so, one may estimate that on-air talent generates far more and music far less than 50% of revenues, that is, say 60% and 40%.

As such, the resulting percentage rate of music royalties expressed on the basis of music-generated revenue would be 21.0 (= 8.4/0.4)%. Increasing that percentage rate by 25% to account for the larger use of music in DPA would yield an equivalent rate for DPA, i.e., 26.3%.

Hence, the percentages, which would satisfy both the principles of technological neutrality and balance, assuming that the commercial radio rates are properly set at their competitive market value level, would be, 8.4% of revenues for Commercial Radio and 26.3% of revenues for Digital Pay Audio services, before any adjustments for repertoire or other reasons.

Cross-References

- [Copyright](#)
- [Economic Efficiency](#)

References

- Bakos Y, Brynjolfsson E, Lichtman D (1999) Shared information goods. *J Law Econ* 42(1):117–156
- Boyer M (2018, forthcoming) The competitive market value of copyright in music: a digital gordian knot. *Can Public Policy*
- Copyright Board of Canada (2002) Pay audio decision. <http://www.cb-cda.gc.ca/decisions/2002/20020315-m-b.pdf>
- Lindahl E (1958) Just taxation – a positive solution. In: Musgrave R, Peacock A (eds) *Classics in the theory of public finance*. Macmillan, London, pp 98–123 (the original appeared as Chap. 4 Part I of *Die Gerechtigkeit der Besteuerung*, Lund 1919; translated by Elizabeth Henderson)
- Osborne MJ, Rubinstein A (1994) *A course in game theory*. MIT Press, Cambridge
- Supreme Court of Canada (2015) *Canadian Broadcasting Corp. v. SODRAC 2003 Inc.*, 2015 SCC 57, File No. 35918
- Varian HR (1999) *Markets for information goods*. Institute for monetary and economic studies, vol 99. Bank of Japan, Tokyo

Naïve Consumers: Contract Economics

Elena D'Agostino
Department of Economics, University of
Messina, Messina, Italy

Abstract

Consumers are viewed as a weak contract party by both lawyers and economists, although with some distinctions. The unconscionability theory addresses consumers' naïvety, to be intended as partial or total incapacity of understanding contract terms, as a reason for public intervention. Economists as well agree that law must protect consumers against sellers' abuses, especially when contracts contain add-ons or are preprinted and no bargaining is allowed.

How law should intervene is still an open question. On the one hand, lawyers point out that courts should not enforce contract clauses literally but rather should replace them with terms that consumers could have reasonably expected and approved. On the other hand, economists focus on how sellers exploit naïve consumers and warn that regulation may turn out pejorative if it is not able to educate naïve consumers or if it allows the seller to raise the price up when forced to offer good terms.

The Notion of Naïvety in Law

Consumers can be viewed as a special category of buyers that has magnetized lawyers and economists' attention in the last decades. A distinction is, however, necessary to highlight what the two disciplines have in common and nevertheless why they differ.

Starting from a legal point of view, when lawyers refer to consumers, they have in mind a weak and nonprofessional contract party with limited or no contract power, and usually uninformed of every clause or consequence of the contract signed (see Korobkin 2003). This sort of considerations has driven the legal literature to formalize a general theory of unconscionability: accordingly, it is unconscionable a contract that mentally competent people would not sign and/or that no fair and honest person would accept. Not surprisingly, consumers' naïvety, to be intended as partial or total incapacity of understanding the contract content in each of its clauses, has been viewed as one of the main factors to support the theory of unconscionability.

Accordingly, since consumers may sign the contract without reading or understanding its content, their signature on the contract may not correspond to a meaningful consent to all its clauses. It comes out that courts should not enforce contract clauses literally, especially those so one sided

to turn out vexatious, but rather should replace them with terms that consumers could have reasonably expected and approved.

The Notion of Naïvety in Economics

On the other hand, economists make a second-order distinction between two categories of uninformed consumers: those who are simply uninformed but conscious of the potential risks hidden into a contract that has been prewritten by the counterparty and those who are not only uninformed but also unconscious of such risks. Consumers in the former category are labeled “rational” or “sophisticated” and are usually assumed to be able to fill their knowledge gap by paying a positive cost to read contract terms. Consumers in the latter category are labeled “boundedly rational” or “naïve” and are assumed to never read the contract simply because they do not realize the risk of signing without reading.

Ellison and Ellison (2009) analyze how firms can exploit consumers’ in Internet transactions where price search engines make a price search more difficult and sometimes not convenient.

As pointed out by Armstrong and Chen (2009), it does not imply that naïve consumers do not care of contract clauses at all or less than sophisticated consumers. There might be also consumers who are naïvely pessimistic, that is, they do not trust the seller and believe that contracts always include unfriendly clauses. More generally, we can say that naïve consumers simply do not realize the risk involved in signing a contract without reading or they never trust the seller and never sign.

Similarly to the legal literature, there is a general consensus in the economic literature as well that court intervention is necessary to protect naïve consumers. To be more precise, economists have focused on how unregulated seller(s) exploit such consumers. The answer clearly depends on the beliefs that naïve consumers hold about the terms in preprinted contracts. Despite in some

contexts a consumer might expect complex contracts to contain default terms (e.g., he might believe that the non-price terms in a complex contract address contingencies which are irrelevant to the buyer and are therefore regulated by default rules or uses), most of the literature prefers focusing on the case in which consumers are weaker and easier to exploit for a contract-drafting seller: it happens when consumers are naïvely optimistic, believing that preprinted contracts contain favorable terms.

The Effects of Regulation

A recent behavioral literature treats the infrequency of reading as indicative of buyer naïvety and argues that sellers lack an incentive to educate such buyers: cf. Gabaix and Laibson (2006) for the special case of add-on goods and D’Agostino and Seidmann (2016) in respect of contracts of adhesion. The intuition is the following: if all buyers believe that complex contracts contain favorable terms, then an unregulated seller would include the worst possible terms in preprinted contracts charging the price that buyers would be happy to pay for friendly terms according to those constraints imposed by the market structure in which they operate. Regulations may therefore benefit such naïve buyers.

Despite the general principle that parties are free to negotiate contracts, courts, legislatures, and regulators have sometimes overruled onerous terms in consumer contracts. There are two underlying and potentially conflicting rationales: Sect. 218 of the Uniform Commercial Code states that a clause is unenforceable if a buyer would not have traded had he known its contents, which suggests that naïve buyers should be protected. By contrast some courts, notably in *Henningsen v. Bloomfield Motors* (NJ 1960), have cited market share as an aggravating factor, which suggests that all buyers should be protected against sellers who exploit market power to offer onerous contracts.

It is also unclear whether regulations are designed to protect buyers who have already accepted onerous terms (and must necessarily gain) or to protect buyers who have yet to enter the market.

Suppose all consumers are, economically speaking, naïve in the sense that they believe that contract terms are favorable and do not take into account the risk of accepting a preprinted contract without reading its content. Sellers in a free market would therefore include onerous terms in their contracts charging the highest price that such consumers are willing to pay for favorable terms. Consumers therefore risk to get a negative payoff, to be intended as the difference between their evaluation for an onerous contract and the price they pay believing that it is rather favorable, under the supposition that they value more a favorable contract than an onerous contract.

Trivially, regulations which do not educate buyers have no effect on play: if naïve consumers remain unaware of the effect of regulation, they will not be able to call the seller in front of a court of law if terms are different from those legally acceptable. Conversely, effective regulations which either mandate terms turning out favorable to consumers or prohibit rather onerous terms induce both a monopolist and each competitive seller to offer a favorable contract, but the effect on price crucially depends on the market structure. Precisely, a monopolist will charge the highest price that consumers are willing to pay for those terms, whereas competition will lead price down to the production cost. As an effect, consumers get zero from buying and may be better off compared to a free market in which they would have been offered onerous contracts charging a price close or equal to their reservation price for a favorable contract.

Suppose now that some consumers are naïve and some others are rational. If rational consumers must pay a cost to read and understand the contract, a commitment problem arises and an unregulated seller has again no incentive to include favorable terms. Precisely, consumers cannot commit to read the contract and seller (s) cannot commit to include favorable terms. As

a consequence, an unregulated seller cannot charge a price equal to consumers' reservation price for favorable terms if he plans to trade with rationals, but under some conditions he may find it profitable to include favorable terms with positive probability if rational consumers read with positive probability. Naïve consumers are therefore protected by rational consumers if the latter category is sufficiently large in the market to force seller(s) to reduce the price and possibly to include favorable terms with positive probability. Conversely, if naïves represent a large percentage of consumers into the market, then seller(s) will ignore rationals and equilibrium conditions correspond to those found when all consumers are naïve. Focusing on the former and more interesting case, regulation will force sellers to offer favorable terms with market structure dictating the equilibrium price. The effect on naïve consumers' payoffs is therefore ambiguous depending on what price they were charged in a free market given the probability of finding favorable terms.

References

- Armstrong M, Chen Y (2009) Inattentive consumers and product quality. *J Eur Econ Assoc* 7(2–3):411–422
- D'Agostino E, Seidmann DJ (2016) Protecting buyers from fine print. *Eur Econ Rev* 89:42–54
- Ellison G, Ellison SF (2009) Search, obfuscation and price elasticities on the internet. *Econometrica* 77(2):427–452
- Gabaix X, Laibson G (2006) Shrouded attributes, consumer myopia, and information suppression in competitive markets. *Q J Econ* 121:505–540
- Korobkin R (2003) Bounded rationality, standard form contracts, and unconscionability. *Univ Chicago Law Rev* 70:1203–1295

National Identity

► [National Locus](#)

National Law

► [National Locus](#)

National Locus

Giuseppe Eusepi

Department of Law and Economics of Productive Activities, Sapienza University of Rome, Faculty of Economics, Rome, Italy

Abstract

The idea of nation grew out of the ancient idea of law and subsequently evolved in the form of national states as territorial boundaries. The unitary national state turned what was an internal conflict between a statesman and a merchant into an external conflict among nation states. As the USA and the EU demonstrates, the concept of nation state as a public good is either ambiguous or tautological. In a more or less distant future, the national locus is apt to move from a territorial to a procedural dimension.

Synonyms

[National identity](#); [National law](#); [Nation-state](#)

Definition

National consciousness as opposed to nationalism.

The idea of nation has been worked out by the Romantic movement, and it may be regarded as the force that set in motion movements engaged in asserting national identities in Europe such as the Italian Risorgimento (Salomone 1970) and the German Nationsbildung (Hroch 2005), which began during the first half of the nineteenth century and extended until the second half of the century and beyond.

The progenitors of the idea of nation are to be found in the idea of state, which marks the end of the Middle Ages and the beginning of the Modern Age, and much earlier in the idea of law that had its roots in ancient Greece and, above all, in ancient Rome. In ancient Rome, law was basically

a private law: the ideas of law and order guaranteed private property and the related freedom of exchange.

Nineteenth-century idea of nation capsizes the private-public law relationship, which shaped the ancient Roman law. The conflict between public law and private law is best represented by the perennial conflict between a statesman and a merchant: the former operates for the purpose of bringing about adjustments of relations among individuals, and the latter is concerned with individuals' desire to get rich. The idea of nation emerged exactly to nuance this conflict by turning conflicts within the nation into conflicts outside the nation. The concept of national locus separates the nation as a unit from the rest of the world. In symbolic terms, the national anthem and the flag symbolize the national locus by definition, but from a more rational point of view, the clearest definition of national locus is the Constitutional Charter and constitutional political economy *sensu lato* (Buchanan and Tullock 1962). The clue to constitutional political economy is to be found in the consensual-contractual dimension, which is the outward form of the inward notion of metarule; thus, the notion of national locus becomes constitutional locus. In this context, the centrality of the constitutional law requires that control on application of the law be operated by a judge acting objectively and impartially just as a judge of the Constitutional Court.

Standard or mainstream economic thinking has been much concerned with a substitute for the procedural component, which was viewed as irrelevant. One such substitute has been seen in the objective dimension with the consequence that the concept of national locus loses its process connotation (including the democratic one) and becomes a measure of an optimal national quantity commonly lumped in the label GDP or better public goods. In its widest abstract sense, the concept of national locus can be pictured as a defense or protection line against external invasions by other nations as well as by foreign goods. Although the concept of national locus seems to have strengthened itself in public economics, especially in the form of Samuelsonian national public goods, nevertheless it is not free from

conflicts. And, in fact, the conflict arises, not only because the distinction between national public goods and local public goods is not so sharp but because the theoretical foundations of the national and local dimensions may be ambiguous. National public goods, if we are to understand what they really are, should be viewed in their territorial dimension – and in this case the distinction is tangible but tautological. There is in fact a difficulty in drawing a line between “local locus” and “national locus” because they shade into one another and give rise to a dimensionally ambiguous concept, which is not clearly determined. The same may be said of the national locus vis-à-vis the supranational locus. The United States and the EU are a flagrant example of this indeterminacy. While in the United States the federal government is the emblem of the national community (national locus/national constitution), in the EU, member states (national loci) are a subset of a supranational body.

Despite its manifest negativity, this ambiguity is tonic to the advancement of the very idea of national locus. No doubt it is too soon to foreshadow an era in which the concept of national locus is conceived of as separated from the concept of nation as territory. But there seems to be good ground for asserting that the national locus will increasingly move over from a spatial dimension to a procedural dimension (supranational alliances). Moving from a territorial dimension to a procedural dimension involves the abandonment of the concept of nation and a stricter role for law and rules.

In a nutshell, the national locus is a notion moving a great distance away from a territorial symbolic connotation of a nation clustering around its hymn and flag. Against this background, the national locus is apt to transform itself into the notional locus of the law (Eusepi 2008) built up over the two foundations of the law: civil law and common law (Pound 2000).

Cross-References

- [Becker, Gary S.](#)
- [Constitutional Political Economy](#)
- [Independent Judiciary](#)

References

- Buchanan JM, Tullock G (1962) *The calculus of consent*. University of Michigan Press, Ann Arbor
- Eusepi G (2008) *Dura lex, sed lex? Insights from the subjective theory of opportunity cost*. *Eur J Law Econ* 26:253–265
- Hroch M (2005) *Das Europa der Nationen. Die moderne Nationsbildung im europäischen Vergleich*. Vandenhoeck & Ruprecht, Göttingen
- Pound R (2000) *The ideal element in law*. Liberty Fund, Indianapolis
- Salomone AW (ed) (1970) *Italy from the Risorgimento to fascism: an inquiry into the origins of the totalitarian state*. Anchor Books, Garden City

Nation-State

- [National Locus](#)

Negotiated Procedures in EU Competition Law

Frédéric Marty¹ and Mehdi Mezaguer²

¹CNRS - UMR 7321 GREDEG, Université Côte d’Azur, Université Nice Sophia Antipolis, Valbonne, France

²CNRS – GREDEG, Université Côte d’Azur, Université Nice Sophia Antipolis, Valbonne, France

Definition

The enforcement of EU competition law in the field of antitrust, e.g., the sanction of abuse of dominant position and collusive agreements, increasingly uses negotiated procedures. Negotiating remedies with incriminated undertakings is a well-known practice in the field of merger control. The practice of settlements is also significantly developed in the United States. However, it remains a relative new approach under the EU competition law enforcement. This chapter presents the three main tools at the disposal of the EU Commission: the leniency program, the direct settlement, and the commitment procedures. It

analyses their main challenges and issues in both legal and economic fields.

Negotiated Procedures Under EU Competition Law: An Introduction

Negotiated procedures under EU competition law mainly encompass three procedures: leniency, commitments, and direct settlements. We mainly consider the negotiated procedures implemented for the application of Articles 101 and 102 of the Treaty on the functioning of the European Union. We draw some parallels with remedies in merger and acquisition control.

These procedures are provided under European Union law using a specific framework. Firstly, they were organized around both soft law and hard law after a period of insecure practice on the shadow of law: Sunlight is said to be the best of disinfectants and the twilight zone in which settlements currently are negotiated, desperately needs light (Van Bael 1986). Secondly, these procedures stemmed from the practice of the European Commission. The Commission increasingly tries to organize its intervention on the market using cooperation with companies for the application of EU competition law (in our case, it is mainly antitrust law). Thirdly, these procedures are organized under the notion of mutual concessions which allows an approach based on rationality. Between main legal instruments and the peripheral ones, a particular dynamic has been installed to the advantage of the public authority. Under EU law, the European Commission is clearly at the center of the proceedings. From that perspective, negotiated procedures have to be discussed above all around the question of their essential nature: Do these procedures pertain to a logic of cooperation with the public agency or to a real negotiation between two partners?

Understanding the reality of these procedures implies to put into light the interests for companies and the public authority, appreciate both legal and economic approach, and consider the problems of the procedures under fundamental rights.

An Overview of EU Negotiated Procedures for Articles 101 and 102 TFEU

The Negotiated Procedures Before the Negotiated Procedures

Understanding the specificities of EU competition law-related negotiated procedures supposes to consider the historical dynamic that led to them, especially the experience of the ad hoc agreements that were implemented in the field anticompetitive practices before the EU 1/2003 Regulation, and to make some connections with the case of remedies in merger controls.

The informal practice of the negotiated practices is a specific situation under EU competition law. Indeed, the European Commission has started in the 1960s to relay on informal arrangements following the fundamental principle that no situation is ever lawful or unlawful forever (see, e.g., cases *Nicholas frères* (Commission Decision 64/502/CEE, 30 July 1964) and *Henkel-Colgate* (Commission Decision 72/41/CEE, 23 December 1971). The European Commission was at this time in charge of a very innovative field, and this type of *cooperation* could have been seen as a will to implement softly antitrust law. The first decisions applying such an approach were mentioned in the reports on competition law by the European Commission and have mainly concerned agreement cases (abuses of domination cases appeared in the beginning of the 1980s). For example, in its first report in 1971, the Commission explained such situation by the need of saving resources. In its fifth report, issued in 1975, the Commission mentioned the “friendly” nature of its intervention. Anyway, in that period, the informal intervention of the European Commission was not well known by public or academics, and the Commission itself was not really able to explain the nature of such procedures (when questioned, e.g., by a resolution of the European Parliament in 1987).

As a consequence, the informal development of negotiated procedures has created some uncertainties, particularly since it was not possible to isolate each procedure according to its specificities. From this perspective, the creation of a formal architecture allowed an individual and global

understanding of negotiated procedures. The formal scheme also had to respond to the need for legal certainty and clarification regarding practices involving cooperation through reciprocal concessions. The resulting procedures, without claiming similarity, converge on several points but need at the same time to be presented over their singularity.

When working on the proposal of a new framework for the application of Articles 101 and 102 (former 81 and 82), the Commission provided that: “The second way in which the proposal will increase the protection of competition is by allowing the Commission to concentrate on the detection of the most serious infringements” (EU Commission 2000). Negotiated procedures aimed indeed to increase efficiency in law application by the liberation of available resources. This has to be considered as a real philosophy at the European level which led to the creation of a formal framework of three procedures qualified as negotiated (Mezaguer 2015).

Leniency Procedures

The first procedure corresponds to leniency programs. These procedures were created only under soft law. Indeed, in 1996, 2002, and 2006, the European Commission has adopted three communications for the organization of the proceedings. The first communication provided that these procedures were intended to apply competition rules with efficiency. The 1996 communication on leniency introduced the first formal instrument providing a legal framework for negotiated procedures under EU competition law. Nevertheless, in the first times of the regulation 1/2003, which constitutes the main instrument of EU antitrust law, no mention of leniency procedures was made. Finally, Regulation 2015/1348 has mentioned such procedures but only related to Regulation 773/2004 which provides only procedural rules. The legal situation of leniency is rather paradoxical.

However, leniency has become a major instrument of cartel resolution even in member states of the EU. Moreover, the European model is a reference for member states even if, in a 20 January 2016, *DHL Express*, C-428/14 case, the European Court of Justice has provided that the creation of

leniency program was not an obligation for member states.

Leniency gives several ways of cooperation to the competition authority and to the incriminated undertakings from different perspectives. Firstly, the company denouncing the cartel runs for immunity if it shows a “true spirit of cooperation.” This situation was generalized under the 2002 communication and differs from the US situation while even the first company is recognized guilty (but without any fine). The condemnation of the first company to a zero euro fine will probably have a specific importance for follow-on civil actions (even if the present framework is questionable, see, for instance, Cauffman (2011) and Mezaguer (2015)). Nevertheless, the first company to denounce must obtain a reward far superior to those who follow it so that the leniency program can be efficient. A leniency program should initiate a denunciation race among the cartelists to be efficient.

After this first company, leniency will allow other cartelists to receive rewards (between 5 and 50% depending on the period) for not disputing the facts, for giving valuable material for the proof, and for going through useful cooperation with the Commission. From this perspective, cooperation implies express, clear, and precise recognition of the facts (*Tokai Carbon*, Court of First Instance, 29 April 2004, joint cases T-236, 239, 244, 246, 251, and 252/01). Lower-rank leniency maintains incentives for the companies to engage a race for cooperation. From this perspective, companies are no longer supposed to help in the mere initiation of the investigation, as it is the case for the first-rank leniency, but must reinforce the Commission in its work of qualifying the infringement. Indeed, it is well known that “a reduction in the fine will be granted for a contribution during the administrative procedure only if that contribution enabled the Commission to establish an infringement with less difficulty and, where appropriate, to put an end to that infringement” (Conclusions of the Advocate General Geelhoed, *Commission v SGL Carbon*, 19 January 2006, case C-301/04 P). Finally, the Commission benefits from a wide margin of interpretation to reward companies. This range of

reductions distinguishes the leniency from the direct settlement procedure.

Direct Settlement Procedures

The transaction procedure corresponds to direct settlements. This procedure has been created through both a Communication of 2 July 2008 (2008/C-167/01) and a Regulation of 30 June 2008 (n° 622/2008) of the European Commission. This Regulation has modified the specific procedural regulation 773/2004, which entirely relies on the European Commission, and not the general one. Even if the use of regulations contrasts with the soft law, the European Commission is still at the center of decisional power. Moreover, the direct settlement procedure is also seen as a piece of a general transactional scheme for cartel resolution with the leniency program. Under the direct settlement procedure, companies who plead guilty before the Commission can expect a 10% fine reduction. At the creation of the procedure, former competition commissioner Neelie Kroes (2005) suggested that “we may need to look at how some form of plea bargaining procedure could bring advantages.” Under the actual framework for direct settlement, companies that “plead guilty” and take a commitment not to contest the conclusions of the Commission receive a 10% reduction of the fine as a reward (the reward is combinable with leniency ones). Consequently, direct settlement is a hybrid procedure between American “plea-bargaining” and French-style “non-contestation des griefs.” It is more a simplified procedure than a negotiated one.

Commitment Procedures

Commitment procedures were the first ones which were included in the general Regulation 1/2003 since their creation (Article 9). Commitment procedure has to be distinguished from leniency and direct settlement first of all because it does not apply to cartels. Indeed, Regulation 1/2003 clearly provides that such a procedure is not possible when the Commission intend to impose a fine. As for other “negotiated” procedures, commitment procedure relies on an efficiency-based objective. Concerning the question of its application under Article 101 or 102 of the TFEU, The

Alrosa/De Beers case (Court of Justice, *Commission v Alrosa*, case C-441/07, 29 June 2010) has shown that Article 9 could be used in both situations (Mezaguer 2015).

Under the commitment procedure, things are clearer. Indeed, when the Commission intends to adopt a decision (requiring that an infringement be brought to an end), the companies can offer commitments meeting the concerns of the public authority, which are expressed in a preliminary assessment. From that perspective, the Commission adopts a decision making those commitments binding on the companies. Moreover, Article 9 of Regulation 1/2003 provides that “Such a decision may be adopted for a specified period and shall conclude that there are no longer grounds for action by the Commission.” Before issuing a decision making the commitments bidding the Commission must submit them to a market test procedure to invite all the interested parties to make comments. These ones aim at helping the Commission to discuss these initial proposals and to require if necessary the proposal of modified ones.

What Are the Reasons to Commit into These Procedures?

We analyze the determinants of the choice to opt for a negotiated procedure by taking successively the point of view of the competition authority and the one of the undertakings.

The Theoretical Advantages of Negotiated Procedures for a Competition Agency

We first present the common advantages of all the negotiated procedures in competition laws for an enforcement agency before considering specifically some advantages related to leniency programs and to commitment procedures (Wils 2008).

Firstly, negotiated procedures aim at reducing administrative costs for the enforcement agency. Indeed, the negotiated option may lead to a shorter duration and to reduce the burden of proof. Secondly, a negotiated settlement allows the competition authority to obtain more rapidly the end of the prejudicial practice. Such a result is also

obtained possibly more surely, as the firm chooses and implements itself the remedy that it had proposed on a voluntary basis. Thirdly, the negotiated procedure in itself limits the information asymmetries between the undertaking and the authority. The last one has not demonstrated the existence of an abuse or of an agreement and has not assessed its effects. In addition, for the commitment procedures, the uncertainties related to the adequacy and to the proportionality of the remedies proposed by the undertaking are limited through the market test procedure, which allows the authority to benefit from the opinions of the different stakeholders (consumers, competitors, etc.). Fourthly, it limits the risk induced by an adversarial procedure. To the extent that the competition authority has neither to assess the net effect of the market practice at stake on the consumer welfare nor to consider the defense of the incriminated undertaking, it reduces at the minimum the risk of being disavowed by the appeal court. Fifthly, negotiated procedures, especially the commitment ones, lead to implementation of compliance programs and by doing so favor the dissemination of a competition culture within the organization. Sixthly, the guarantees about the effectiveness of the remedies are enhanced by the fact that a failure in their implementation would lead to an automatic fine, irrespective of any assessment of the effect of this nonfulfillment. For instance, Microsoft was fined on this basis for its negligence in its implementation of the remedies negotiated in the MS Explorer case, as we will see below.

In a more specific way, we can consider that negotiated procedures have two main types of advantages for competition authorities. It avoids the burden of the characterization of competition law provisions infringement through material or economic evidences. For instance, in the case of leniency programs, the material evidences are brought by the undertaking applying for leniency. The second main advantages consist in the effect of the decision on market structure or on the dominant undertakings' future behavior. A fine mainly produces a deterrence effect. It does not lead to correct the effects of the previous market

practices. In a very opportunistic spirit, the fine can be seen by a dominant firm or by the cartelists as the *cost of doing business*. On the contrary, a negotiated decision leads to behavioral or even structural measures that may remedy to an unsatisfactory situation on markets and to some extent to reestablish a level playing field on the market. It may ensure for the future a fairer and more effective competition.

An adversarial decision, as the *Google Shopping* one (June 2017) or the *Android* one (July 2018), for instance, might lead to equivalent corrective measures through the injunctions to provide an equal treatment to rival comparison shopping services and its own services in the first case or to end its tying practices and to allow the development of *fork Android* operating systems, in the second case. However it remains that these two Commission decisions are both appealed before the General Court. Even if, the General Court and the Court of Justice will uphold these injunctions, their implementation will must be strictly monitored. Such remedies will be challenged and at best implemented in a noncooperative way. At the opposite, a commitment procedure participates to a jointly constructed competition regulatory approach and not only to an ex post legal enforcement of competition law provisions. It remains that the negotiation at stake is not really a horizontal bargaining between two equal partners but rather a transactional agreement with a public authority within a vertical relationship. The EU Commission *Google Shopping* case demonstrated that the two "partners" were not in equivalent position.

The Advantages Considered on the Incriminated Firms' Side

For the incriminated undertaking, opting for a negotiated procedure may reduce significantly the procedure duration. Indeed, the shorter the case, the less expensive its costs in terms of financial resources, management attention, and reputational effect (especially on the stock market). In addition, avoiding any recognition of guilt may induce two main benefits: Firstly, it makes successful follow-on actions for damages less probable, and, secondly, it allows to stay away from any

fine increase in a future case grounded of the aggravating circumstance of repeated offense.

In the specific case of leniency programs, the undertaking applying for the leniency may hope exiting from a cartel agreement without any legal (defense) costs or administrative sanction, as fines. In the case of a direct settlement, the undertaking benefits from a fine reduction. In other words, the fine may be reduced outside the scope of leniency since the undertaking does not challenge the theory of damage presented by the competition agency or its assessment of effects. In addition, the commitment procedure allows the undertaking to avoid to be fined and to have to recognize to have breach the law. In addition, the undertaking still benefit from informational asymmetries allowing to propose the less costly remedies possible.

Considering all these advantages on both sides, negotiated procedures seem to pertain to a *win-win strategy* (Bellis 2013).

The US Origins of Negotiated Procedures in Competition Law: What Can We Learn for EU Ones?

The Antitrust Division head, Thurman Arnold, has played an essential role in this resolved use of these settlements rather than opting for adversarial procedures before courts (Waller 2004). In the competition law-related field, the negotiated procedures constitute a US transplant. It is worthwhile to consider the US antitrust history to put into relief the main features and the underlying logic of these procedures. Although the real start of the large implementation of settlement procedures by the Antitrust Division of the Department of Justice (DoJ) can be dated in the late 1930s and in the late 1940s, the first Antitrust Division's consent decrees was the *US v Otis Elevator Company* in 1906.

How to explain this specific choice at this moment of American antitrust history? The F.D. Roosevelt administration was disappointed about the results of the cooperation with dominant firms during the *National Industrial Recovery Act* (NIRA) implementation before its invalidation by the Supreme Court. Big firms were suspected of

having not fairly play the game of an economic coordination based on government support to promote investment and employment. The coordination among firms effectively led to stop price drops, but it mainly allowed firms to increase their markups without effective counterparts. This disappointment has pleaded for a public antitrust enforcement renewal, especially because, the government benefited from a large portfolio of easy-to-win cases. The NIRA agreements supervision had led government agencies to accumulate data and evidence about collusive practices. The government position was also comforted by the TNEC report conclusions. The Temporary National Economic Committee established in 1938 aimed at analyzing the monopoly powers in the US economy. While government accumulated evidence about collusions and monopolization practices, why did Th. Arnold opted for settlements and not for conventional antitrust procedures? The reasons were twofold. The first one was to obtain quickly effective results by avoiding endless legal procedures. The second one was also related to a procedural concern: averting the legal risk induced by a still conservative Supreme Court case law.

This historical experience is all the more relevant for us that the EU Commission itself has used these kinds of procedure after sector-specific enquiries in order to avoid General Court and Court of Justice legal control and to circumvent member states' reluctances to accept legal proposal aiming at liberalizing some economic sectors as utilities. We may provide the example of the energy sector enquiry in 2005–2007. A large number of formal procedures were opened after this enquiry against gas and electricity dominant operators across the EU (Hancher and de Hauteclocque 2011). For all except one, the final decision was not an infringement one but a negotiated one. Even nowadays, far-reaching and broad scope competition law remedies are obtained through commitment procedures in the EU energy sector as testified in the May 2018 *Gazprom* decision (case 39.816 Upstream gas supplies in Central and Eastern Europe, 24 May 2018).

However, US and EU procedures and practices remain specific. In the US case, for instance, for

the Antitrust Division of the DoJ, an antitrust settlement is a horizontal contract between the firm and the agency that requires judicial supervision. Indeed, the 1974 Tunney Act requires an *ex ante* judicial review of the DoJ consent decrees. Whatever the supervision of these agreements, it remains that the US antitrust enforcement realized a shift from a litigation-oriented model toward a more regulatory-based regime. The balance between courts and agencies has shifted dramatically toward the last ones. Ginsburg and Wright (2012) showed that by the 1980s, already 97% of the DoJ antitrust cases were settled. From 2004 to nowadays, almost all its cases were resolved through these negotiated procedures. The tendency is the same considered on the FTC side.

In cartel-related cases, the rise of negotiated procedures was later but equally important. The US corporate leniency program was first introduced in 1978. It was revised in 1993 and completed by an individual leniency program 1 year later. Its scope of application covers agreements aiming at setting prices, market or consumer sharing devices, and bid-rigging practices. From the initial 1978 procedure to the 1993, three major revisions were implemented: (1) leniency is now automatic for qualifying companies if there is no preexisting investigation; (2) leniency is still available even if cooperation begins after the investigation is underway; and (3) all employees who come forward with their company and cooperate are individually protected from criminal prosecution. The real start of the US leniency program can be dated 1993: from 1993 to 2010, US enforcement data reveal a 20-fold increase. Nowadays, in the USA 90% of the penalties imposed by the DoJ were linked to leniency-related cases (OECD 2018).

The institutional specificities of the US enforcement regime may explain this precocious and fast development. Opting for a negotiated procedure might be a rational choice for an enforcement authority while considering the judicial and political risks associated to unfruitful lawsuits. In addition, the burden of the rule of reason had grown earlier in the USA as the shift toward an effects-based approach in antitrust laws enforcement started at the late 1970s. This

tendency also raises a theoretical question: are the settlements more efficient in terms of administrative performance (deciding cases quicker and easier) or in terms of obtaining concessions from the firms for *not-so-easy* cases for which the enforcement agency has no certainty about its own chances to be successful before a court? Such a game relies on the fact that the competition authority does not disclose its evidence and remains vague on the theory of damage or on the assessment of the damage to competition. As we have already mentioned for the EU competition law case, the enforcement agency does not issue a statement of objections but only a communication of competition concerns. The second player – the incriminated firm – does not know what the competition agency exactly knows and what its real chance of being convicted are. Because of the asymmetry of information that benefits to the authority, a risk-averse undertaking may prefer enter in a commitment procedure. Indeed, the trade-off between prohibition and commitment decisions cannot be analyzed without taking into account uncertainties related dimensions (Gautier and Petit 2018).

Is the tendency observed for the EU competition law enforcement as significant as it is for the US Case? From 2007 to 2017, we count 19 Article 7-based decisions (antitrust prohibitions) and 32 Article 9-based ones (antitrust commitments). If we consider the case of cartels, we may count 40 cartel prohibition decisions (based on an adversarial procedure) and 25 based on settlements. However considering on the period from 2013 to 2017, the balance is rather different with 7 adversarial procedures and 18 negotiated ones (EU Commission 2018). The same constraints tend to produce the same results. In addition, we have to keep in mind that competition law enforcers are engaged in convergence process.

Economic and Legal Concerns Related to the Recourse to These Procedures

This conclusion illustrates some of the economic and legal issues raised by the recourse to negotiated procedures. We first consider the case of

leniency programs before considering the one of commitments. We finally consider the nature of these procedures in order to draw a dividing line between negotiated ones and transactional ones.

A Law- and Economic-Based Discussion of the Possible issues Raised by Leniency Programs and Commitment Procedures

The case of leniency programs raises several issues within the economic field. For instance, what about the incentives to collude while the expected cost of the sanction is reduced while a firm anticipates that she will be the first to betray its partners in crime? Do these procedures only lead to dismantling only poorly efficient cartels? Should it be necessary to grant a monetary reward to the cartel member who reports a cartel agreement before any investigation (Brisset and Thomas 2004)? Considered at the legal point of view, how to consider the capacity of a cartel, possibly the initiator of the cartel, to escape at any sanction after a breaching of competition laws? How to articulate public and private enforcement, especially if we consider the follow-on actions aiming at obtaining damages? The near predominant part of leniency procedures in the cartel-related competition law enforcement undoubtedly raises issues in terms of restorative justice. At the same time, controversies over differences in the treatment of unilateral practices in the USA and the European Union can be read through the prism of relative weight differences between public and private enforcement (Cosnita-Langlais and Tropeano 2018). The rise of leniency programs in the US case and its consequences in terms of private enforcement do not raise the same issues as in the EU where follow-on actions have to be encouraged (see, for instance, Bueren and Smuda 2018).

The development of commitment procedures also leads to raise several issues in the economic field. For instance, how taking into account information asymmetries to assess the adequate, effective, and proportionate character of remedies? It raises significant adverse selection-related concerns. Do the behavioral or structural remedies proposed sufficient to address competitive concerns? For instance, in case of asset divestitures,

are the values of these last ones properly selected? Will the future buyers able to exert an effective competitive pressure? The EU procedure of the *up-front buyer* aims at limiting the risk to transfer the asset to a market player who has not the financial or technical capacities to compete or has excessive incentives to collude or to avoid a too fierce competition with to dominant firm. For instance, a divestiture benefiting to a major competitor may create a symmetry among the main competitors within the relevant market and by doing so enhance the risk of collective dominance.

These concerns are not the only one induced by the information asymmetries at stake in commitment procedures. The supervision of the proper implementation of remedies also raises moral hazard-related issues. Will the undertaking implement properly and efficiently the remedies it has proposed? According to EU regulations, a failure to comply with commitments that a Commission's decision made binding leads to an automatic sanction. Microsoft experienced the situation with its €561 million fine imposed to its failure to provide European users a screen choice of web browsers from May 2011 to July 2012 despite its 2009 commitments (see the EU Commission decision, case AT.39530, *Microsoft*, 06/03/2013).

Symmetrically, information imperfections may lead the antitrust enforcement authority to require excessive remedies. Farrell (2003) illustrated this case for mergers remedies with the scalp, overfixing, and broad scope phenomena. The first one corresponds to an application of disproportionate remedy from a dominant firm in order to sanction its past behaviors or to address its structural dominance. The second one consists in requiring remedies going beyond what is necessary by taking into account the information asymmetries that the undertaking does benefit. It plays the role of a security margin. The third phenomenon corresponds to remedies unfitted to the damage theory presented in the communication of the competition concerns. It might correspond to a situation in which the competition authority takes advantage of the negotiated procedure to address several issues even if all of these ones are not closely related to the case.

Such phenomena may be difficult to observe in adversarial procedures mainly because of the judicial control exerted on them. However, this control is by far relaxed under the EU competition law by the Court of Justice judgment in *Alrosa* (EU Court of Justice 2010). To the extent that the remedies are voluntarily proposed by the undertaking, the EU Commission has not to check if less demanding or less intrusive remedies might be proposed. If the EU Commission has also to perform a proportionality test, this last one is limited to the remedies effectively proposed by the undertaking. In other words, if the dominant undertaking only proposes structural remedies, the authority has not to verify if behavioral ones may allow to obtain similar effects. These specificities may lead to concerns about competition authorities' capacity to obtain "excessive" remedies in these negotiated procedures. It may be illustrated, for instance, by the case of divestitures. These last ones are seldom used in adversarial decisions (Article 7) but can be observed in Article 9 ones. The case of the European energy sector is emblematic of such difference. The EU case may be all the more specific that the incumbent has special duties regarding the effectiveness of competition. This duty may increase the probability of being fined under a conventional procedure. It enhances the incentives to opt for negotiated procedures, and it simultaneously increases the potential cost of a negotiation failure if the Commission will decide to go back to a conventional procedure. The Google case is striking example of such a risk. Taking into account this one may induce some bias in the "negotiations" by leading firms to propose "disproportionate" remedies in order to have a greater probability to see these ones accepted.

Indeed, the use of negotiated procedures raises issues in terms of remedies proportionality. If we insist on the importance of information asymmetries, we may fear that a dominant and better-informed dominant undertaking may propose insufficient remedies or may behave strategically during their implementation in order to reduce their effect. In the same way, we might take into consideration the risk that the competition agency tends to opt for settlements while its

expectations to win before courts are unfavorable. The risk in such a case is to negotiate unsatisfactory remedies to close the procedure. However, as we have underlined, the symmetrical risk cannot be minimized, especially if the dominant undertaking does prefer avoiding a prohibition decision, taking into account its reputational impacts, its induced risks in terms of follow-on actions, and possibly the possible fine increases in future decisions on the basis of the aggravating circumstances pronounced in case of recidivism.

This competition authority's strong position in the bargaining may lead to costly remedies for the incriminated dominant undertaking. We can provide an example in the domain of mergers with the *Bayer/Monsanto* case (decision of the European Commission; 11 April 2018, case M8084). Bayer took the commitment to sell to BASF a part of its activities in order to prevent a possibly non contestable position in the market of seed and pesticides. The Commission has required an up-front buyer condition. It led to limit the number of potential acquiring firm and possibly to reduce the transfer price. The up-front buyer requirement aims preventing to transfer the assets to a non-efficient market operator who cannot exert a long-term credible competitive threat on the dominant firm. Such a requirement makes sense in order to protect the competitive process, while it has an adverse effect on the dominant firm's interests.

At the legal point of view, commitments procedures may also raise several concerns. Firstly, what is the proper scope of competition law remedies? Should these ones concern prices? On the principle, it has not to. The competition authority has not the play the role of a price regulator. Nevertheless, as soon as an essential facility is at stake, commitments deal with access price-related issues. Compulsory licensing-based remedies also raise the same concern. An even more far-reaching question could be put into relief for remedies aiming at reinforcing the competitive situation of a given market.

Again, the *Gazprom* case can be relevant to illustrate this situation. In 2015, the Commission sent to Gazprom a statement of objections. According to the Commission's preliminary view, this company breached EU competition

rules by pursuing an overall strategy to partition gas markets along national borders in several member states in Central and Eastern Europe. This strategy may also have enabled Gazprom to charge higher gas prices. The remedies proposed by Gazprom and negotiated with the EU Commission both address these competitive concerns and deepen the gas internal market. They do not sanction an anticompetitive behavior by pronouncing a fine in a deterrence purpose as a prohibition decision would do. These remedies pertain to the building of competitive markets. The competition field will be not the same before and after the remedies. It is not an issue to reestablish the condition of a free and undistorted competition but an issue to create a competitive market.

What are the remedies in this specific case? A first one undoubtedly pertains to a logic of guaranteeing the end of the alleged anticompetitive practices. Gazprom commits to remove all the contractual provisions that imposed geographical restrictions or that limit the customer's capacity to resell Russian gas. A second remedy concerns gas prices. A structured process to ensure competitive gas price will be put in place in order to guarantee that Russian gas price will be aligned with the prices observed on Western European market places. The third remedy may contribute to change the competitive structure of the gas market itself: "Gazprom will enable gas flows to and from parts of Central and Eastern Europe that are still isolated from other Member States due to the lack of interconnectors, namely the Baltic States and Bulgaria." In other words, this commitment will oblige Gazprom to open its network for intra-EU gas exchanges. This remedy is a quasi-structural one. It will play as an alternative to new investment in new gas infrastructures within the EU.

A Legal Perspective on the Commitment Procedure

The rise of commitment procedures may raise other concerns in the legal field. Firstly, the negotiated procedures under the EU competition regulations lead to limit the scope of judicial control. It may question the effectiveness of the guarantees that protect the fundamental rights of the undertakings. Remedies affect their freedom of contract

and their property rights. In the same vein, their entitlement to a fair trial may be altered. Secondly, we may wonder what could be the undesirable collective consequences of a near from generalized recourse to these procedures. It may reduce the quality of the jurisprudence itself. Case law is a public good at the economic sense of the term. A commitment decision reveals a poor information compared to a prohibition decision. The disappearance of the adversarial stage of the procedure hinders discussions about the theory of damage, about the balance of the effects, and about the adequacy and the proportionate character of the remedies. Altering *the struggle for law* deteriorates the quality of the signal produced by the case law. It may have several consequences. A first one is to reduce the capacity of the different stakeholders to anticipate the decisions. It impairs the legal certainty attached to the legal rule definition and its implementation. By doing so it increases the dominant undertaking's propensity to opt for such procedures. Eventually, the higher the number of cases settled, the higher the possibility to observe the development of parallel case law, specific to negotiated procedures and all the more robust that it is not balanced by any judicial control exerted by the General Court and by the Court of Justice.

We might finally wonder if negotiated procedures under EU competition regulation do really imply an effective negotiation. We have noted that a commitment procedure is not a private contract that must be validated by a court as it is the case in the USA. The bargaining between the incriminated undertaking and the authority cannot be only conceived in a horizontal way. Indeed, there is a significant verticality at stake. The EU Commission is not a player as another one. The Commission benefits from a large margin of discretion in favoring this kind of procedure for a given case (through communicating competition concerns and not issuing a statement of objections). The Commission also decides unilaterally to accept or not the commitments proposed by the undertaking and to come back to an infraction decision.

The *Google Shopping*, *Android*, and *AdWords* cases are particularly striking. Considering that the initial negotiated procedure was unsuccessful

(e.g., that the commitments proposed by Google were not sufficient to address its competitive concerns), the Commission unilaterally decided to go back to an Article 7 procedure and to send in 2015 a statement of objections. Three procedures were launched. A first one led to the *Google Shopping* decision of June 2018 (with a €2.42 million fine). A second one came to the *Android* decision of July 2018 (with a €4.34 million fine). A third one, corresponding to *AdWords* related practices is still expected.

On the one hand, it tends to enhance the credibility of the Commission. Commitments are not suggested to dominant undertakings for weak cases, and the Commission demonstrates that it may refuse insufficient proposals. The Commission's behavior may be analyzed as a reputational investment. On the other hand, these decisions contrast with the Recital n°13 of Regulation 1/2003 according to which: "Commitment decisions are not appropriate in cases where the Commission intends to impose a fine." Negotiated procedures in antitrust appear more as a competition policy tool (with all the characteristics associated to public policy in terms of verticality) than a "contractualization" of the competition law enforcement in a pure horizontal logic.

References

- Bellis J-F (2013) Article 102 TFEU: the case for a remedial enforcement model along the lines of section 5 of the FTC Act. *Concurrences* 1:54–61
- Brisset K, Thomas L (2004) Leniency program: a new tool in competition policy to deter cartel activity in procurement auctions. *Eur J Law Econ* 17(1):5–19
- Bueren E, Smuda F (2018) Suppliers to a sellers' cartel and the boundaries of the right to damages in U.S. versus EU competition law. *Eur J Law Econ* 45(3):397–437
- Caffman C (2011) The interaction of leniency programmes and actions for damages. *Comp Law Rev* 7:181–220
- Cosnita-Langlais A, Tropeano J-P (2018) How procedures shape substance: institutional design and antitrust evidentiary standards. *Eur J Law Econ* 46(1):143–164
- E.U. Commission (2000) Proposal for a council regulation 2000/0243, COM (2000) 582 final, 27 Sep 2000, p 6
- E.U. Commission (2018) Report on competition policy, COM(2018) 482 final, 18 June
- EU Court of Justice (2010) case C-441/07P, *Commission v Alrosa*, judgment of 29 June
- Farrell J (2003) Negotiation and merger remedies: some problems. In: Lévêque F, Shelanski H (eds) *Merger remedies in American and European Union competition law*. Edward Elgar, Cheltenham, pp 95–105
- Gautier A, Petit N (2018) Optimal enforcement of competition policy: the commitments procedure under uncertainty. *Eur J Law Econ* 45(2):195–224
- Ginsburg DH, Wright JD (2012) Antitrust settlements: the culture of consent. In: William E. Kovacic: an antitrust tribute, *liber amicorum*. Concurrences, Paris, ed. 407p
- Hancher L, de Hauteclouque A (2011) Manufacturing the EU energy markets: the current dynamics of regulatory practice. *Comp Regulat Netw Ind* 11(3):307–334
- Kroes N (2005) 'The first hundred days', Speech at the International forum on EC competition law, Brussels, 7 April 2005
- Mezaguer M (2015) Les procédures transactionnelles en droit antitrust de l'Union européenne – Un exercice transactionnel de l'autorité publique, Bruylant, Bruxelles, 582p
- OECD (2018) Challenges and co-ordination of leniency programmes, DAF/COMP/WP3 (2018)1
- Van Bael I (1986) The antitrust settlement practice of the EC commission. *Common Market Law Rev* 23:61–90
- Waller SW (2004) The antitrust legacy of Thurman Arnold. *Saint John's Law Rev* 78:569–614
- Wils W (2008) The use of settlements in public antitrust enforcement: objectives and principles. *World Comp* 31(3):335–352

Neuro Law and Economics

Emanuele Lo Gerfo¹ and Ferruccio Ponzano²

¹Economics, Management and Statistics
Department, University of Milano Bicocca,
Milan, Italy

²Department of Law and Political, Economic and
Social Sciences, University of Piemonte
Orientale, Alessandria, Italy

Abstract

Neuro law and economics is a very young discipline that aims to study law and economics phenomena towards the help of neuroscientific techniques. These studies are the result of bringing together two different approaches that start from different disciplines in order to analyze, in particular, some aspects of the punishment. These two approaches are behavioral and experimental economics and the analyses deriving from the so-called neuro-law. These

two disciplines have been developed independently in the last decades and only in the last years they are jointly used to implement scientific analyses of punishment.

Introduction

Neuro law and economics can be considered as a very young discipline that studies law and economics phenomena towards the help of neurosciences. These studies can be seen as the result of bringing together two different approaches that start from different disciplines in order to analyze, in particular, some aspects of punishment. They implement some experimental and neuroscientific techniques. These two approaches are the studies of behavioral and experimental economics and the analyses deriving from the so-called neuro-law. These two disciplines – that incorporate interdisciplinary approaches – have been developed independently in the last decades and only in the last years they are jointly used to study some scientific phenomena. Then, we have to present these different approaches to understand what neuro (law) and economics is.

Behavioral and Experimental Economics

First of all, we present how the development of Behavioural and Experimental Economics allowed to shed light, among other topics, on the study of punishment in a very new perspective for economists. Let us consider a simple environment as the one represented in the ultimatum game, where we have the presence of two subjects that face a simple economic dilemma. The first player has to choose how to divide a given amount of money between himself and the second player. Then, the second one may or may not accept the choice made by the first one. If the second one accepts the monetary endowment is divided as the first player decided. If the second player does not accept the choice of the first one both the players earn an amount of money that is equal to zero. The classical economic theory predicts that the second player will accept each positive monetary offer from the first one, even if the offer is very

unfair. Moreover, the second player is indifferent between the acceptance or the rejection when the first player offers zero. Güth et al. (1982) show – using an the ultimatum game in an economic experiment – that people – the second players – are ready to reject a positive monetary offer if they consider it unfair, even if this choice reduces to zero their monetary endowment. Then, people are ready to spend money to sanction other people if they behave in an unfair way. In particular, Güth et al. (1982) show that people are ready to spend money to sanction an unfair action that damages them. This pioneering result has been confirmed by many other studies in the following decades, also with the implementation of different experimental protocol as the public good game with punishment (Fehr and Gächter 2000, 2002) and the trust – or the investment – game (Berg et al. 1995).

An important ancillary result has been reached after that a new experimental protocol using the third party punishment game has been developed. In his simplest version, in this game we have the presence of three players. A first player can decide how to divide a given amount of money between himself and a second player. After this choice, a third player must decide whether or not to sanction the first one if she/he thinks that the division between the first two players has been unfair. Obviously, the sanction has a positive monetary cost for the third player. Also in this case, the classical economic theory never predicts that a third-party will be ready to spend money to sanction an unfair behavior. As Fehr and Fischbacher (2004) show for the first time, people are ready to sanction an unfair behavior in this case also, even if it damages a person that is not the punisher. Then, they show that people are ready to spend money to sanction an unfair action even if they are not directly damaged by that. Also in this case, other studies confirm this result (see, for instance Henrich et al. (2006) and Marlowe and Berbesque (2007)) also in environments with the presence of a jury that acts as a third party (Ottone et al. 2015).

Obviously, the study of behavior related to sanction is really wide. For a survey on this topic and some suggestions for future research, the readers can refer the paper by Mulder (2016).

Neurolaw

Secondly, we have to present the development of neurolaw. In fact, law and science over time had to meet and to confront each other. Scientific discoveries have always raised a strong international juridical debate. From this meeting and clash between law and science, more or less suddenly, innovations and changes arise into legal system. Suffice it to think to legal issues about the declaration of human death or when we have before us a brain death. Other examples could be permission to use new informatics technologies inside the courtrooms or to use biological evidences such as blood or DNA trails.

Neurolaw stems from this continuous debate between law and science, particularly from Neuroscience.

Neuroscience is a discipline that studies the nervous system. It arises from the relationship of biology with medicine, psychology, mathematics, engineering, chemistry, thus evolving as an interdisciplinary science. Aim of neuroscience is to understand how the interconnections among single nervous cells (neurons) produce different perceptive and motors acts and different cognitive processes such as decision making or problem solving.

In the last years of 1900 different disciplines apparently unrelated to neuroscience begin to incorporate into their researches neuroscientific data and methodologies giving rise to new disciplines such as Neuroeconomic, Neuromarketing, or Neuroesthetics.

Neurolaw explores the effects of neuroscientific discoveries on legal rules (Petoft 2015) investigating law-relevant mental states and decision-making processes in defendants, witnesses, jurors, and judges.

Aim of law is to regulate individual's conducts, to respect human dignity, and to create a just and fair legal system. Moreover, law judges evidences about the causes of human behaviors, and neuroscientific study of human brain functions can make a very important contribution, so that the interaction/integration of neuroscientific discoveries in Law is inevitable.

The term “neurolaw” first appeared in 1991 in a scientific paper by Taylor et al. entitled “Neuropsychologists and Neurolawyers.” In this paper, Taylor and his co-authors analyzed how neuroscience, and in particular Neuropsychology, could provide probative data during a trial with defendants with cerebral lesions.

In subsequent years, thanks to the advent of new technique of neuroimaging such as functional magnetic resonance imaging (fMRI) or the not invasive technique of brain stimulation such as transcranial magnetic stimulation (TMS) and Transcranial Direct Current Stimulation (TDCS), the attention to neurolaw increased. In 2007, Law and Neuroscience Project was born from The John D. and Catherine T. MacArthur Foundation. The interdisciplinary project has the goals to help the legal system in order to avoid an improper use of neuroscientific evidence in some law contexts. Furthermore, it tries to deploy neuroscientific insights to improve the fairness and effectiveness of the criminal justice system.

The neuroscientific techniques usually adopted in neurolaw are electroencephalography (EEG), fMRI and TMS and TDCS. EEG is a not invasive technique which measures brain activity during the rest or during some cognitive tasks towards the information provided by superficial electrodes placed on various regions of the scalp.

fMRI is a correlational methodology which enables to detect the so called BOLD (blood oxygenation level dependent) signal. It is an indirect measure of neural activity and represents a small change in the levels of oxygenated red blood cell when a muscular or mental activity is performed. This technique is very useful because it is characterized by a high spatial resolution (the identification of volumetric area of 3 mm).

TMS and TDCS are causal methodologies, which enable to interfere with the cerebral activity and to observe functional changes. The first uses the electric field elicited by a magnetic field, and the second uses low intensity of electric current. Both have a moderate spatial resolution and TMS has an optimal temporal resolution.

It is possible to identify several topical areas on which the research of neurolaw scientists is focused.

Mind Reading or Lie Detection

To detect deception and truth in an individual is the dream of many judges. The use of lie detector tests has become a popular and legendary symbol from crime dramas to comedies to advertisements. The lie detector test provides for the use of three physiological signals recorded with a Polygraph: heart rate/blood pressure, respiration, and skin conductivity. Courts, including the United States Supreme Court, had repeatedly rejected the use of Polygraph because of its inherent unreliability.

Recently EEG method was employed in two forensic techniques: Brain Fingerprinting (BF) and Brain Electrical Oscillations Signature (BESOS) (for a description of these technique see Chauhan (2016) and Puranik et al. (2009)). The first was admitted in 2003 in a legal proceeding of Terry Harrington in Iowa. BF detects particular brain waves called p300. According to its inventor Lawrence Farwall (neuroscientist at Harvard university), these waves are correlated to memories of the past. Terry Harrington was sentenced to life for a murder, but defense subjected him to BF which resulted in a negative outcome, as if memories of the murder were not in his memory. Iowa District Court admitted BF test as scientific evidence and the murder case was re-opened. Later Terry Harrington was acquitted after the confession of a key witness. The case of Terry Harrington is the first and only one where a Court admitted the BF test. This is relevant because, although the BF has not been exposed to a peer review, its admission challenges previous judgment. This case is the proof that a neuroscientific evidence can influence the judge's decision making.

Different neuroscientific researches investigated the brain networks that underlie lie production (Meijer et al. 2016 for review). fMRI is the most used method and it is employed by several companies such as No Lie MRI. Some studies conducted by Daniel Langleben showed how lies are distinguished from truth by increased prefrontal and parietal activity (Langleben et al. 2005). Although numerous fMRI neurolaw researches on lie detection are continuing, the Courts still exclude fMRI lie detection evidence from trials. As an example, Judge Eric M. Johnson of the

Maryland Sixth Judicial Circuit had refused to admit potential exculpatory fMRI evidence during the murder trial of State v Gary Smith, claiming that "the use of fMRI to detect deception has not achieved general acceptance in the scientific community" (MacArthur Foundation 2016).

In fact, fMRI studies have different general limitations: because of its correlational nature, this methodology cannot be used to prove causation. When an fMRI study shows that a brain region is active when a person is trying to lie, one cannot exclude that the same brain region might be active because it is involved and activated during a state of anxiety or in any other cognitive processes. One meta-analysis study showed that the regions usually related to deception were the ventrolateral and dorsolateral prefrontal cortex, inferior parietal lobe, anterior insula, and medial superior frontal cortex (Christ et al. 2009). These brain regions are also activated during general executive processes as planning, working memory, and emotional process. Therefore, a limitation is to discern whether the neural activity measured by fMRI is associated with lying or with other neural processes. Moreover, many scientists argue that lie detection data recorded with fMRI are valid only within the context of controlled laboratory experiments, which often poorly approximate the real word.

Neuroscientists agree that more studies are needed before starting to use fMRI lie detection methodology.

Criminal Responsibility

Differently from fMRI, MRI was frequently admitted in courtrooms when it is relevant to obtain behavioral information correlated to specific brain pathologies. In 2007 a court of New York have authorized the scan of the defendant's brain, the popular journalist Peter Braunstein indicted of kidnapping and sexual assault. The defense intended to show that the brain illness "Schizophrenia" might have made it difficult for its client to control violent impulses. At the end of the trial, Braustein was utterly convicted: this case demonstrates how a neuroscientific evidence can be useful but not necessarily

lead to acquittal. Neuroscientific evidence could be able to indirectly address questions of intent in defendants with neurological or psychological illness that might reduce the ability of judgment during the crime. Another important contribution of neurolaw regards the brain development in adolescence. In fact, neuroscientific evidence has demonstrated that the adolescent brain has a slow development of the cerebral regions necessary for cognitive control as Prefrontal cortex (Blakemore and Robbins 2012). For this reason, adolescence is characterized by low risk aversion and impulsiveness. The contribution of neuroscience to law about the behavior of adolescents brought in 2005 the US supreme court to the abolition of the death penalty.

Another important contribution of neuroscience to law is represented by studies on memory focused on memory limitations. These limitations may affect not only witnesses but also juries.

Bringing Together Behavioral and Experimental Economics and Neurolaw

Now neurolaw researches also deal with neuroscience evidence to understand and to improve decision making of the judge. Human societies universally expect that criminals will be punished, usually by impartial third-party decision makers. Then, the integration of Neurolaw together with behavioral economics can use various experimental protocols to investigate the punishment. Obviously, paradigms commonly used by neuroscientist are those used by experimental economists such as the third-party punishment or the hypothetical crime scenarios. The first is the economic game presented before where a potential punisher, the third-party, observes how a dictator share an amount of money with a receiver. In this case he may choose to punish the dictator spending his money or he can simply observe the scene. In this case, neuroscientific tools can be used to observe what happens in the brain of the three players. In the second paradigm, the subjects make various decisions about some different scenarios where a crime can be committed. These paradigms are used together with neuroscientific

methodologies as fMRI or not invasive brain stimulation (TMS, TDCS) in order to investigate the neuronal underpinning of punishment. The data showed different brain networks, with an involvement of anterior Insula and dorsal and anterior cingulate cortex and Amigdala. This first network seems to be involved in the detection of norm violation (Sanfey et al. 2003; Krueger and Hoffman 2016). A second network is formed by medial prefrontal cortex, ventromedial and dorsomedial prefrontal cortex, posterior cingulate cortex, and tempo parietal junction. The second brain network connected to the first one is generally associated to self-monitoring, mentalizing (Bressler and Menon 2010), emotional processes related to the harm to the victim. Both networks are connected to a central executive network engaging posterior parietal cortex and dorsolateral prefrontal cortex. The central executive network underlies the final decision making and then selects a specific punishment, integrating the information processed in previous networks with the contextual facts as circumstances surrounding the crime. The neuroscientific evidence about punishment could elucidate the neuronal and contextual variables that can influence the decision making.

Moreover, Neuroeconomics, another neuroscientific discipline, helps the juridical system and several times neuroeconomic studies are borderline with neurolaw studies. Pisoni et al. (2014) in a TMS study assessed that fair or unfair economic decisions, and context, in terms of different interaction styles, may modulate the corticospinal excitability. Authors show that economic exchange, indeed, especially when imposed by only one part (proposer), can elicit positive or negative emotions according to the context in which it occurs, implying different moral evaluation and reactions. These results are very important for a legislator and for policy, because explore the people's reactions to injustices.

Future Developments

Although there are the a number of studies about legal decision making, lie detection, mind reading, and criminal responsibility, legal experts and

lawyers still try to understand the best way to integrate neuroscience, behavioral economics, and law. Today, the best neuroscientific data that neurolaw can deliver to law are the evidence about criminal responsibility of individuals with neurological diseases. Some neurological pathologies indeed can elicit different criminal behaviors, especially diseases caused by damage to frontal brain areas involved in control of instinctive behaviors and decision making. The other topics of neurolaw so far have not gained scientific validations external to laboratories. The law is good to be prudent. Anyway interdisciplinarity will lead us to the law of future.

Cross-References

- Behavioral Law and Economics
- Cognitive Law and Economics
- Crime: Economics of, the Standard Approach
- Crime and Punishment (Becker 1968)
- Experimental Law and Economics
- Prosocial Behaviors

References

- Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity and social history. *Games Econ Behav* 10:122–142
- Blakemore SJ, Robbins TW (2012) Decision-making in the adolescent brain. *Nat Neurosci* 15(9):1184–1191
- Bressler SL, Menon V (2010) Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci* 14(6):277–290
- Chauhan JSD (2016) Brain fingerprinting. *Int Res J Eng Technol* 3(07):494–497
- Christ SE, Van Essen DC, Watson JM, Brubaker LE, McDermott KB (2009) The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cereb Cortex* 19(7):1557–1566
- Fehr E, Fischbacher U (2004) Third-party punishment and social norms. *Evol Hum Behav* 25:63–87
- Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. *Am Econ Rev* 90(4): 980–994
- Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415:137–140
- Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. *J Econ Behav Organ* 3(4):367–388
- Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Henrich N, Lesorogol C, Marlowe F, Tracer D, Ziker J (2006) Costly punishment across human societies. *Science*. New Series 312(5781):1767–1770
- Krueger F, Hoffman M (2016) The emerging neuroscience of third-party punishment. *Trends Neurosci* 39(8): 499–501
- Langleben DD, Loughhead JW, Bilker WB, Ruparel K, Childress AR, Busch SI, Gur RC (2005) Telling truth from lie in individual subjects with fast event-related fMRI. *Hum Brain Mapp* 26(4):262–272
- MacArthur Foundation research network on Law and Neuroscience (2016) www.lawneuro.org/LieDetect.pdf
- Marlowe FW, Berbesque JC (2007) More ‘altruistic’ punishment in larger societies. *Proc R Soc B Biol Sci* 275(1634):587–590
- Meijer EH, Verschuere B, Gamer M, Merckelbach H, Ben-Shakhar G (2016) Deception detection with behavioral, autonomic, and neural measures: conceptual and methodological considerations that warrant modesty. *Psychophysiology* 53:593
- Mulder LB (2016) When sanction convey moral norms. *Eur J Law Econ*. <https://doi.org/10.1007/s10657-016-9532-5>
- Ottone S, Ponzano F, Zarri L (2015) Power to the people? An experimental analysis of bottom-up accountability of third-party institutions. *J Law Econ Org* 31(2): 347–382
- Petoft A (2015) Neurolaw: a brief introduction. *Iran J Neurol* 14(1):53
- Pisoni A, Lo Gerfo E, Ottone S, Ponzano F, Zarri L, Vergallito A, Lauro LJR (2014) Fair play doesn’t matter: MEP modulation as a neurophysiological signature of status quo bias in economic interactions. *Neuroimage* 101:150–158
- Puranik DA, Jospeh SK, Daundkar BB, Garad MV (2009, November) Brain signature profiling in India: it’s status as an aid in investigation and as corroborative evidence-as seen from judgements. In: Proceedings of XX all India forensic science conference, pp 815–822
- Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the ultimatum game. *Science* 300(5626):1755–1758

Next-Generation Access Networks

Claudio Feijóo
Sino-Spanish Campus, Tongji University -
Technical University of Madrid, Shanghai, China

Abstract

Next-generation access networks refer to telecommunications infrastructures – and the

technologies that support them – able to provide final customers with data rates above 30–50 Mbps.

Synonyms

Ultrafast broadband networks; Ultra-broadband networks

Sometimes the “access” is missing and the term is used simply as “next-generation networks (NGN),” although strictly speaking NGN would refer to the whole infrastructure and not only to the part closer – the access – to the final user.

Acronym

NGAN

Definition

Next-generation access networks refer to telecommunications infrastructures – and the technologies that support them – able to provide final customers with data rates above 30–50 Mbps.

Reference Framework: Broadband and Economic Development

The rise of the knowledge economy has reinforced the role of telecommunications as a strategic investment:

... the ability to communicate information at high speeds and through various platforms is key to the development of new goods and services. Broadband enables new applications and enhances the capacity of existing ones. It stimulates economic growth through the creation of new services and the opening up of new investment and jobs opportunities. But broadband also enhances the productivity of many existing processes, leading to better wages and better returns on investment. (EC 2006)

Next-Generation Networks and Next-Generation Access Networks

Next-generation networks (NGN) are the supporting infrastructure of ubiquitous broadband.

They are defined as networks based on the Internet Protocol able to deliver multiple data applications – whether originally based on voice, data, and video – to multiple devices, whether fixed or mobile. In addition, the provision of applications is decoupled from networks facilitating the introduction of innovations (De-Antonio et al. 2006).

An NGN can be divided into two main parts (Knightson et al. 2005):

- A backbone transport network that interconnects the local nodes where data traffic from the final users is gathered to be switched and further transported. The backhaul from distant nodes to the core network is typically included as part of the backbone, although it is convenient at times to consider it separately (the so-called middle-mile).
- An access network that links final users with local nodes, the next-generation access network (NGAN), colloquially called the last-mile.

NGAN Requirements

There is nothing like a strict definition of the minimum access speeds provided by an NGAN or any other defining parameter. A tacit agreement at the industry level seems to put this figure at 50 Mbps or beyond, but to prove the vagueness of the case, there are no indications whether this number refers to both the upstream and downstream parts or it should be applied just to the downstream channel. Even less is mentioned about the quality of service-guaranteed data rates per customer. Also, while the above figure may represent as of May 2014 some consensus, there are a number of regulatory decisions and digital strategy plans that implicitly address figures from 30 Mbps to 100 Mbps; see the Digital Agenda Europe for 2020 as a main example. In any case, these figures are beyond conventional broadband capabilities and therefore the name of ultra-broadband networks.

Some sources provide a narrower definition in relation to NGN to include exclusively wired access networks which consist wholly or partly in optical elements and are capable of providing enhanced broadband access compared to services provided over existing copper networks.

According to this type of definitions, wireless networks are not part of NGAN; see below.

NGAN Technologies

In general, broadband access technologies can be classified by the physical medium into two major groups: wired (or fixed line) technologies and wireless technologies. The main wired technologies are based on fiber, coaxial, copper wire (or any combination of them), and power line. Wireless technologies can be either satellite based or terrestrial. Terrestrial wireless solutions can be either fixed or mobile. Due to their niche market prospects and limitations, power line communications, satellite solutions – or other recently proposed airborne solutions such as balloons – and fixed wireless are not usually accounted among the NGAN technologies.

Therefore, the list of NGAN technologies as of 2014 reduces to:

- *Fiber to the home (FTTH)*. In this technology fiber runs all the way to the customer premises from the local node. FTTH technology is the ultimate fixed solution, supplying the highest data rates possible per household (Kramer et al. 2012).
- *Fiber to the basement/building/cabinet/curb/node/premises (FTTx)*. FTTx is a generic term for those technologies which bring fiber from the central office closer to the subscriber. They come in many varieties depending on the termination point of the optical network. In all of these architectures, the fiber from the central office is brought down to a node where equipment is housed in a cabinet to convert signals from the optical network (fiber) into electronic (copper wire, wireless connection, or even coaxial cable). The main advantage of these solutions is reusing part of the existing legacy network.
- *Digital Solutions on subscriber loop (xDSL)*. From the point of view of the architecture of the technology, xDSL solutions are equivalent to FTTx solutions mentioned above, their only difference being the perspective adopted: from the copper wire or from the fiber optics side.

Among them VDSL is the most widely used technology over copper wire. With new technological developments able to increase data rates, copper lines will continue to be a strategic asset well into the midterm. Not only are they able to provide data rates that would fall into the NGAN category, but, in addition, they also allow for a smoother and more scalable path in the transition from existing broadband to FTTH.

- *Upgrade of cable television networks (DOCSIS)*. Typically cable networks use HFC (“hybrid fiber-coaxial”) technology. DOCSIS (“Data Over Cable Service Interface Specification”) is the name of the series of standards developed for high-speed data transmission over cable television networks with release 3.0, the most widespread as of 2014. In general it could be said that the role of cable networks is particularly relevant in the NGAN competition scenario as the only different infrastructure from those of historical telecoms operators.
- *4th generation mobile communications (4G)*. The case of mobile wireless is controversial regarding its inclusion into NGAN. Mobile technologies are approximately 3–5 years behind fixed technologies in terms of sustained data rates per user. However, they already deliver peak data rates well above 100 Mbps, and they are not far from reaching the 10 Mbps level per user with some consistency. Therefore, mobile broadband connections are considered here as a suitable (stand-alone or complementary) technological alternative to fixed access technologies. The most relevant technology is LTE, labeled as 4G by the International Telecommunications Union in 2010 (Ghosh et al. 2010). 4G plays a fundamental role: not only is it the cheapest solution for rural areas, but it can also complement or even replace fixed broadband in urban and suburban areas, especially as wireless technologies fit mobile lifestyles better.

NGAN Deployment

The choice of access technology is simply a matter of deployment costs (which in turn depend

basically on socio-demographics and geographic variables and possible reuse of existing infrastructures) and the user's requirements (and expectations).

Regarding the status of deployment, in general terms it can be said that as of 2014 NGAN is still in a relatively early stage of deployment – particularly out of main urban areas. In any case, the transition from copper to fiber access networks is underway, and it is expected to result in the replacement of most copper access networks over the next two decades.

Future Directions

From a technical perspective, future prospects for NGAN include some form of fixed-mobile convergence, where fiber networks will be complemented by heterogeneous wireless networks (Raychaudhuri and Mandayam 2012). The rationale is that no access technology enjoys the optimal characteristics for satisfying all the requirements demanded by users in every circumstance. Therefore, this case is leading operators to create platforms capable of integrating different access technologies over the same backbone network. The future market of the ICT sector, characterized by “comprehensive” operators, would be quite different from the current one, where there is clear separation between technologies.

From an economic perspective, the conditions for the deployment of NGAN are currently on the forefront of the debate about the role of telecommunication markets, the best regulation for them, the conditions for the return on investments, the type and level of competition, the requirements for sustained innovation, and the level and modes of potential public involvement (Bauer 2010).

Cross-References

- [Public Investments: Broadband](#)
- [Telecommunications](#)

References

- Bauer JM (2010) Regulation, public policy, and investment in communications infrastructure. *Telecommun Policy* 34(1):65–79
- De-Antonio J, Feijóo C, Gómez-Barroso JL, Rojo D, Marín A (2006) A European perspective on the deployment of next generation networks. *J Commun Netw* 5:47–55
- EC (2006) Bridging the broadband gap. European Commission, Brussels. Retrieved from <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2006:0129:FIN:EN:PDF>
- Ghosh A, Ratasuk R, Mondal B, Mangalvedhe N, Thomas T (2010) LTE-advanced: next-generation wireless broadband technology [Invited paper]. *Wirel Commun IEEE* 17(3):10–22
- Knightson K, Morita N, Towle T (2005) NGN architecture: generic principles, functional architecture, and implementation. *Commun Mag IEEE* 43(10):49–56
- Kramer G, De Andrade M, Roy R, Chowdhury P (2012) Evolution of optical access networks: architectures and capacity upgrades. *Proc IEEE*. <https://doi.org/10.1109/JPROC.2011.2176690>
- Raychaudhuri D, Mandayam NB (2012) Frontiers of wireless and mobile communications. *Proc IEEE*. <https://doi.org/10.1109/JPROC.2011.2182095>
- ### Further Reading
- Astely D, Dahlman E, Furuskar A, Jading Y, Lindstrom M, Parkvall S (2009) LTE: the evolution of mobile broadband. *IEEE Commun Mag* 47(4):44–51
- Bourreau M, Cambini C, Hoernig S (2012) Ex ante regulation and co-investment in the transition to next generation access. *Telecommun Policy* 36(5):399–406
- Brito D, Pereira P, Vareda J (2010) Can two-part tariffs promote efficient investment on next generation networks? *Int J Ind Organ* 28(3):323–333
- Cave M (2010) Snakes and ladders: unbundling in a next generation world. *Telecommun Policy* 34(1):80–85
- Cave M, Martin I (2010) Motives and means for public investment in nationwide next generation networks. *Telecommun Policy* 34(9):505–512
- Charalampopoulos G, Katsianis D, Varoutas D (2011) The option to expand to a next generation access network infrastructure and the role of regulation in a discrete time setting: a real options approach. *Telecommun Policy* 35(9–10):895–906
- Coomonte R, Feijóo C, Ramos S, Gómez-Barroso J-L (2013) How much energy will your NGN consume? A model for energy consumption in next generation access networks: the case of Spain. *Telecommun Policy* 37(10):981–1003. <https://doi.org/10.1016/j.telpol.2013.09.002>
- Elixmann D, Ilic D, Neumann KH, Plückerbaum T (2008) The economics of next generation access-final report.

- WIK-Consult Report for the European Competitive Telecommunication Association (ECTA)
- Essiambre R, Tkach RW (2012) Capacity trends and limits of optical communication networks. *Proc IEEE*. <https://doi.org/10.1109/JPROC.2012.2182970>
- Fredebeul-Krein M, Knoben W (2010) Long term risk sharing contracts as an approach to establish public-private partnerships for investment into next generation access networks. *Telecommun Policy* 34(9): 528–539
- Ghosh A, Mangalvedhe N, Ratasuk R, Mondal B, Cudak M, Visotsky E, . . . Novlan TD (2012) Heterogeneous cellular networks: from theory to practice. *Commun Mag IEEE*. <https://doi.org/10.1109/MCOM.2012.6211486>
- Given J (2010) Take your partners: public private interplay in Australian and New Zealand plans for next generation broadband. *Telecommun Policy* 34(9):540–549
- Gomez-Barroso JL, Feijóo C (2010) A conceptual framework for public-private interplay in the telecommunications sector. *Telecommun Policy* 34(9):487–495. <https://doi.org/10.1016/j.telpol.2010.01.001>
- Hoernig S, Jay S, Neumann K-H, Peitz M, Plückerbaum T, Vogelsang I (2012) The impact of different fibre access network technologies on cost, competition and welfare. *Telecommun Policy* 36(2):96–112. <https://doi.org/10.1016/j.telpol.2011.12.003>
- Huigen J, Cave M (2008) Regulation and the promotion of investment in next generation networks – a European dilemma. *Telecommun Policy* 32(11):713–721
- Inderst R, Peitz M (2014) Investment under uncertainty and regulation of new access networks. *Inf Econ Policy* 26:28–41
- Kazovsky LG, Shaw WT, Gutierrez D, Cheng N, Wong SW (2007) Next-generation optical access networks. *J Lightwave Technol* 25(11):3428–3442
- Kazovsky L, Wong S-W, Ayhan T, Albeyoglu KM, Ribeiro MRN, Shastri A (2012) Hybrid optical-wireless access networks. *Proc IEEE*. <https://doi.org/10.1109/JPROC.2012.2185769>
- Marsden C (2010) Net neutrality. Bloomsbury Academic, London. Retrieved from <http://www.bloomsburyacademic.com/pdf/NetNeutrality.pdf>
- Nitsche R, Wiethaus L (2011) Access regulation and investment in next generation networks – a ranking of regulatory regimes. *Int J Ind Organ* 29(2): 263–272
- Noam E (2010) Regulation 3.0 for telecom 3.0. *Telecommun Policy* 34(1–2):4–10
- Ruhle E-O, Brusici I, Kittl J, Ehrler M (2011) Next Generation Access (NGA) supply side interventions – an international comparison. *Telecommun Policy* 35(9–10):794–803. <https://doi.org/10.1016/j.telpol.2011.06.001>
- Wallsten S, Hausladen S (2009) Net neutrality, unbundling, and their effects on international investment in next-generation networks. *Review of Network Economics* 8(1):90–112
- Wong E (2012) Next-generation broadband access networks and technologies. *J Lightwave Technol* 30(4): 597–608

No-Fault Revolution and Divorce Rate

Bruno Jeandidier

Bureau d'Economie Théorique et Appliquée (BETA), CNRS and University of Lorraine, Nancy, France

Abstract

Concomitantly with the no-fault divorce revolution in the United States, there has been an increase in the divorce rate. From this observation, the question emerged of whether divorce law had a neutral effect on divorce behavior. Economists have conceptualized the issue by using an approach based on the “Coaseian” negotiation process, comparing unilateral divorce to divorce by mutual consent. The theoretical model predicts the neutrality of the law. But in the case of divorce, these assumptions are questionable. A determination as to the neutrality or non-neutrality of divorce law can thus be made based on empirical analysis. Early empirical studies came to conclusions favoring neutrality, but little by little, as methodological controversies accumulated, those empirical conclusions became more refined. Finally, a certain consensus emerged that the revolution in divorce would seem to have had a positive short-term impact on the divorce rate on the one hand, due more to the transition to unilateralism than to the abandonment of any reference to fault, and to have had a negative effect on the longer-term divorce rate on the other, because divorce law reform would seem to have prompted better-quality marriages.

Introduction

The term “no-fault divorce revolution” comes from the United States. It originated in the debates that came about in the 1980s as one after another the various American states transitioned from at-fault divorce regimes to no-fault divorce regimes (about economic origins of no-fault

divorce revolution; see Leeson and Pierson 2017). This change in American civil law stirred fierce conflict between the advocates of traditional family law and advocates of more progressive and feminist positions. The origins of this polemic are doubtless to be found in the stir created in the media and the political sphere by the alarmist book by Weitzman (1985). Based on data drawn from California State divorce rulings, the author showed that the introduction of no-fault divorce in that state would seem to have resulted in declines in the standard of living for the average female divorcee that were far more significant than had been the case prior to divorce law reform. Subsequent work (including Peterson 1996) showed Weitzman's (1985) estimates to have been largely exaggerated; nevertheless the debate had been launched around the negative consequences of the no-fault divorce revolution. Parallel to questions about the estimation of consequences relative to the standard of living of ex-spouses following a divorce, the debate also focused heavily on whether or not this new divorce regime would in fact encourage divorce. In other words, was it the abandonment of fault as a condition of divorce that explained the increase in the divorce rate observed during that same period? Economists have seized upon this question, seeing it as an interesting application of the more general issue of the neutrality of law. Starting from the pioneering article by Peters (1986), the no-fault divorce revolution thus spread to scientific journals in economics and sociology as well, giving rise to a fairly extensive series of articles in which more recent articles challenged the preceding ones and so on.

Peters' Pioneering Article Establishes the Theoretical Framework for Economic Analyses of the No-Fault Divorce Revolution

For Peters (1986), and subsequently for other economists, the discussion does not center on the concept of fault in itself but on the decision-making process: unilateral decision versus decision by mutual consent. In a no-fault divorce

regime, either spouse may apply for divorce unilaterally on the grounds that he or she no longer wishes to live as a couple, even if their partner does not agree. In a fault-based divorce regime, an injured spouse cannot apply for a divorce if the offending spouse does not agree to it; mutual consent is required unless fault is proven in court. It was in particular the observation of frequent difficulties in proving fault (with negative effects) that led American legislators to abandon the at-fault divorce regime.

Peters (1986) then uses Coase's theoretical model as an analytical framework to express the hypothesis of the neutrality of the law in regard to divorce: under certain conditions, negotiations can be arranged between the spouses, resulting only in efficient divorces, regardless of the divorce regime in force. The divorce regime in force would therefore be considered neutral.

Broadly, and making a simplified use of the notation proposed by Peters (1986), the gain derived from marriage R can be expressed as follows:

$$R = M(1 - p) + E(A_f + A_m)p \quad (1)$$

where p is the probability of divorce, M is the benefit derived from living as a couple, and A_f and A_m are the estimated postdivorce opportunities for the wife and husband, respectively, with the following assumptions:

- M is fixed, known, and perfectly divisible.
- Although A_f and A_m are not known with certainty, each spouse is fully aware of the opportunities available to their partner (informational symmetry), and these postdivorce gains are perfectly divisible.
- A division of M was agreed upon at the time of marriage, attributing X to the wife and $(M - X)$ to the husband.
- The negotiations between the couple take place without transaction costs (because people in couple relationships know one another very well).

The divorce is thus a fairly simple matter: it will be an efficient divorce if $M < (A_f + A_m)$; the

husband will want a divorce if $(M - A_m) < X$, and the wife will want a divorce if $A_f > X$.

If both spouses want a divorce and the divorce is efficient, or, conversely, if both spouses wish to remain married and the divorce is not efficient, the question of the neutrality of the law does not arise. In the other two cases, the following reasoning will apply. The situation we will study will be one in which the husband wants a divorce but not the wife; the same reasoning may however be applied symmetrically, with the wife desiring the divorce and not the husband.

We will first look at a situation where the divorce would be efficient. If the divorce law in force is unilateral, the husband will ask for a divorce without taking his partner's negative opinion into consideration; this behavior is efficient. If, on the other hand, mutual consent is required, the husband will have to negotiate his wife's acceptance of the divorce: in an "at-fault" divorce regime, the spouse seeking the divorce thus negotiates the "right to remain married" held by the other party. The object of this negotiation is to grant compensation (C_f) to the wife in order at least to cancel out the loss that the divorce represents for her ($X - A_f$):

$$C_f = (X - A_f) + \alpha(A_f + A_m - M) \quad (2)$$

The negotiation therefore concerns α , a coefficient that represents the extent to which the husband is willing to share the gains to be obtained from the divorce with his former spouse.

Secondly, in the case of divorce by mutual consent, the husband will not be able to obtain the divorce if the divorce is not already efficient prior to negotiation, because the gains he will derive from the divorce will be insufficient to compensate for the loss the divorce will represent for his partner, and it will thus be efficient to remain married. In the case of unilateral divorce, the husband will request a divorce a priori, with no need for negotiation. However, the wife is still able to negotiate in regard to her husband's "right to divorce" and may in fact renegotiate the division of marriage gains X so that it is more in the husband's interest to remain married than to separate:

$$(M - X) = A_m + \beta[M - (A_m + A_f)] \quad (3)$$

Ultimately, whatever the divorce regime in place (unilateral or mutual consent), if negotiations are held under the assumed conditions (symmetry of information, no transaction cost, divisibility), the Coase model shows that only efficient divorce situations ($M < (A_f + A_h)$) will arise. Thus the negotiated compensation system makes it possible, on the one hand, to avoid the inefficient divorces that might occur under a unilateral divorce regime and, on the other hand, to successfully conclude the efficient divorces that in a mutual consent divorce system could be blocked by one of the two spouses. The particular type of divorce law in force would thus be neutral with respect to divorce decisions, and so, empirically, there should be no increase in the divorce rate associated with the change in divorce regime.

Are the Hypotheses of the Coase Model Suited to an Analysis of the No-Fault Divorce Revolution?

Peters (1986) acknowledges that the hypotheses of the Coase negotiation model may be strong in the case of divorce. On the one hand, it is quite possible that symmetry of information regarding postdivorce opportunities for each of the spouses ($A_m + A_f$) may not be respected (one strategy may consist in concealing this information, and such information may be costly to obtain). In this case, negotiation in the event of a unilateral divorce may result in the non-avoidance of inefficient divorces (insufficient negotiation of the share-out of marriage gains). And, in case of divorce by mutual consent, it is possible that efficient divorces may not be implemented (insufficient negotiation of compensation). If this is the case, an imperfect symmetry of information would lead to an increase in the divorce rate.

On the other hand, it is possible that marriage gains M may not initially be well known (at the time of marriage). In this case, in a unilateral divorce regime, since the wife cannot oppose the divorce and cannot expect to obtain compensation, it is possible that she will self-insure

(by making less of an investment in the domestic sphere), which is likely to reduce marriage gains M and hence increase divorce probability p . In this scenario, the type of divorce is thus not neutral with regard to the divorce rate.

Zelder (1993a, b) analyzes two other limitations of the thesis of the neutrality of divorce law in a discussion of two other hypotheses. On the one hand, the author shows that in the presence of transaction costs, the divorce law is not neutral. The author first studies the situation of unilateral divorce regimes, where transaction costs are much lower (even zero) than those in mutual consent divorce regimes. This is the hypothesis most often retained in sociological studies based on the fact that unilateral divorce is simple and inexpensive. In the context of efficient divorce, in the first case (unilateral), the wife is not able to fully induce the husband to remain in the marriage for lack of sufficient means (which is in line with the assumption of legal neutrality). In the second case (mutual consent), the husband cannot induce the wife to divorce because the transaction costs would absorb her divorce surplus. There would thus be more divorces under unilateral divorce. The author then proceeds to study situations where transaction costs are prohibitive regardless of the divorce regime. In a unilateral divorce regime, if divorce is inefficient, the wife cannot induce the husband to remain in the marriage (whereas she could do so in the absence of transaction costs), and in a mutual consent divorce regime, the husband no longer has the means to induce his wife to divorce (in spite of his divorce surplus). There again, the divorce rate should be higher in unilateral divorce.

On the other hand, Zelder (1993a, b) also looks at the fact that marriage gains M may not be divisible and/or transferable (at least partially) between spouses, when said gains consist in a public good. Now, the primary marriage gain is often the child, who is, in fact, a public good. Under a unilateral divorce regime, when a child is present, even if the divorce is inefficient, the spouse who wishes to remain married will not be capable of inducing his or her partner to give up the divorce because their asset (the “consumption” of the child) is not transferable because it is

not divisible. Again, under this assumption of indivisibility, the divorce rate should therefore be higher in unilateral divorce regimes.

These studies show that the neutrality of divorce law, as conceptualized through the Coase negotiation model, is highly dependent on the assumptions applied in the theoretical model. Thus empirical observation must ultimately be used to attempt to make a determination in regard to the neutrality of divorce law, in an investigation of whether or not the divorce rate observed does depend on the type of divorce regime in force.

Empirical Controversies Regarding the Impact of the No-Fault Divorce Revolution on the Divorce Rate

Because the various American states adopted the no-fault regime on different dates, the United States has provided an exceptional natural experimental site to study the relationship between the divorce rate and the type of divorce regime. Again in his pioneering article, Peters (1986) empirically demonstrates that divorce law is neutral, based on data from the *Current Population Survey* for 1979. His demonstration is based first of all on the observation that, all matters being otherwise equal, the fact of residence in a state that had opted for no-fault divorce was not significantly related to the probability of having had a divorce in the 4 years prior to the investigation. Secondly, his demonstration is supported by the observation that, on the other hand, residence in a State that had opted for no-fault divorce was positively and significantly correlated with the amount of compensation obtained (child support, alimony, etc.), which would thus support the negotiation hypothesis of the theoretical model.

Allen (1992), using the same data as Peters (1986), then challenged this demonstration by advancing three methodological criticisms. On the one hand, the status of states that made changes to divorce legislation during the 4-year observation period would appear to be ambiguous. On the other hand, the categorization of States according to the classification “at-fault versus no-fault” would appear highly debatable,

insofar as there is a whole continuum of situations between strict at-fault regimes and strict no-fault regimes; for example, certain states had barred the attribution of fault for the submission of a divorce application but retained the notion of fault based on negotiations regarding the distribution of assets and compensation. This critique was also made by Brinig and Buckley (1998). Lastly, it would be a mistake to introduce State-based indicators into the specifications because of the correlation with the indicator “at-fault versus no-fault.” Taking these criticisms into account, Allen (1992) then shows that, contrary to the results obtained by Peters (1986), residence in a state that had opted for no-fault divorce is in fact positively and significantly related to the probability of divorce: the divorce regime would therefore not be considered neutral.

Peters (1992) responds to Allen (1992), accepting the first two criticisms and thus adopting the corrections proposed by Allen (1992) but rejecting the third criticism on the grounds that this specification allows consideration of an unobserved heterogeneity according to which it is possible that it was the States with the highest divorce rate that were the first to opt for a no-fault divorce regime (reversal of causality). In reassessing his model with two of the three corrections proposed by Allen (1992), Peters (1992) then reaches the same conclusion obtained in his original 1986 work, namely, that the type of divorce law is neutral with regard to the probability of divorce.

A year later, Zelder (1993a) published results based on different data (*Panel Study of Income Dynamics*). In the first estimate, the author confirms the conclusions reached by Peters (1992), by showing the absence of significant correlation between the probability of divorce and residence in a state that had opted for no-fault divorce. Then, in the second estimation, the author demonstrates the relevance of his public good hypothesis. To do so, the author crosses the indicator “at-fault versus no-fault” with an estimation of the public good (the share falling to the children – estimated based on expenditures for the children – in the total assets of the couple). The fact that this cross-variable has a regression coefficient that is

significantly positive thus shows that residing in a State which has opted for no-fault divorce would encourage divorce when there are children present: divorce law would therefore not be neutral for married parents.

Another controversy then arose among researchers no longer using individual data but divorce rate time series by state. Nakonezny et al. (1995), using average 3-year divorce rates, show that divorce law is not neutral, insofar as these rates, observed before and after the transition to the no-fault divorce regime, are significantly different. Glenn (1997) criticizes this approach on the grounds that these estimations need to be corrected taking into account the general trend in divorce rates; the author shows that the differences highlighted by Nakonezny et al. (1995) are actually four times lower. Rodgers et al. (1997) then respond to Glenn (1997), accepting his criticism and introducing a corrective to reflect the trend but asserting that the new version is not adequate to reverse their initial conclusion of the non-neutrality of divorce law. Brinig and Buckley (1998) arrive at an identical conclusion, with different specifications. Without going directly into the controversy opposing Glenn with Rodgers et al., Ellman and Lohr (1998) propose a more elaborate methodology for correcting the trend, by taking into account peaks occurring before and after the legislative reform (the wait-and-see effect before, the surge effect after); yet their conclusions are mixed, suggesting that each State has a certain specificity. Glenn (1999) then once again revived the controversy with Rodgers et al. by questioning the method they used to calculate the trend (linear, over 10 years), insofar as the general progression of the divorce rate was not linear; it in fact accelerated at the end of the period, which with a linear specification would result in the attribution of part of that acceleration to the divorce law reform. Finally, Rodgers et al. (1999) responded to this criticism by means of a rather lengthy graphical analysis carried out on a state-by-state basis, an analysis which does not, however, lead to a definitive conclusion.

The third wave of controversy primarily involved Friedberg and Wolfers. Friedberg (1998) gathered all prior criticisms so as to better

integrate them into his analysis of individual data: the issue of endogeneity, the issue of classifying states based on the types of divorce legislation in force, the issue of trend correction, etc. His conclusions are as follows: whereas on the one hand there would indeed seem to be an effect due to divorce law (the divorce rate was 6% lower in the States that did not make a change to their legislation), this impact would appear to have been due primarily to the transition to a completely no-fault divorce regime (when the notion of fault is partially preserved, the impact is weak), and the effect would seem to be fairly permanent (it was strong just after the reform and then strengthened over the next 2 years). Wolfers (2006) then criticizes Friedberg's (1998) specification, primarily from the perspective considering the fixed "state-year" cross-effect (which mixes the fixed per-state effect and the effect of the reform) and the observation window being too short to take the trend into account. From the perspective of the results, his primary difference with Friedberg (1998) is that while the effect of the reform would indeed appear to be significant immediately after its entry into force, this effect nevertheless stabilizes after 8 years and then declines. This conclusion is consistent with that of Gruber (2004), using aggregate data drawn from several successive population censuses. Divorce law would thus appear to be non-neutral in the short term but neutral in the longer term. Drevianka (2008) then adds nuance to this conclusion. By separating no-fault divorce reforms from unilateral divorce reforms (since the two concepts are not strictly identical), the author shows that in fact the latter type of reform has an impact on the probability of divorce, and not the former.

The studies conducted by Wolfers (2006) are significant, because they open up a new research question: why does the impact of the no-fault divorce revolution diminish over time? The studies conducted by Rasul (2005) and Mechoulam (2006) then demonstrate the double impact of divorce law: in the short term, a transition to no-fault (easier) divorce leads to an increase in the divorce rate, and in the longer term this effect is offset by a negative impact on the divorce rate. Indeed, divorce becomes less probable because

new marriages are of better quality (better, but later matches) in anticipation of the possibility of unilateral divorce easier to obtain and thus reducing marriage gains. In sum, the impact of the no-fault divorce revolution would appear fairly limited. In a certain way, 20 years later we came "full circle," since Peters (1986), in his pioneering article, had suggested that the no-fault divorce revolution could ultimately lead to a drop in the divorce rate if, anticipating an easier divorce and thus a conduct of self-insurance on the part of wives (synonymous with decreased marriage gains), behavior on the marriage market (marriage and remarriage) tended toward greater selectivity in matching.

The debate over the no-fault divorce revolution has not really crossed the Atlantic. Few studies have examined the neutrality of divorce law in Europe. The few European works were late in coming and therefore have benefited from the methodological advances made in North American studies. Taking up the methodologies used by Friedberg (1998) and Wolfers (2006) and applying them to 14 European countries experiencing divorce reforms between 1950 and 2003 (excluding countries that legalized divorce during this period), Gonzales and Viitanen (2009) conclude that in Europe, a transition to no-fault divorce would seem to have had a positive and permanent effect on the divorce rate, whereas a transition to unilateral divorce would seem to have had a short-term positive effect, fading after 5 years (results quite similar to those obtained by Wolfers (2006) for the United States). Kneip and Bauer (2009) confirm the latter results by conducting a similar analysis of 18 European countries for the period 1960–2003. These authors, however, contribute an original clarification, by distinguishing *de jure* unilateral divorce from *de facto* unilateral divorce. Divorce is *de facto* unilateral when mutual consent is no longer required at the end of a (usually short) separation period undertaken by the couple. They then proceed to show that, without calling into question the very short-term positive effect (on the divorce rate) of *de jure* unilateral divorce reform, the move to *de facto* unilateral divorce would also seem to have had a positive effect on divorce rates but a

permanent one. In sum, for the period 1970–1990, divorce law reforms in Europe would seem to account for one-third of the growth in the divorce rate (which grew by 2% in the same period), but would not at all explain the growth in the divorce rate before 1970 and after 1990. Bracke and Mulier (2017) discuss the classification of Gonzalez and Viitanen (2009) between no-fault divorce and unilateral divorce and propose to introduce in addition the reforms of procedure that make divorce easier. Reforms that reduce the duration of divorce proceedings (unilateral divorce being conditional on a period of de facto separation) would be also important. On the basis of data relating to Belgium and using a method of cointegration, they show that the reduction in the duration associated with the simplifications of 1994 had the same impact on the divorce trend as the introduction of unilateral divorce in 1974. They also estimate that a reduction of 1 month of legal divorce process would increase the divorce trend by 1.4%.

Conclusion

Economic analysis of the impact of the no-fault divorce revolution on divorce behavior has been a very interesting application, combining theory and empirical analysis, of the issue – very central to the economic analysis of law – of the neutrality of the law relative to individual behaviors. The refinement of econometric methodologies and the historical hindsight provided by the gradual accumulation of rich databases have led researchers to obtain nuanced conclusions, in particular by showing that the apparent neutrality of the divorce law relative to the divorce rate resulted, in fact, from the combination of a positive short-term effect and a negative longer effect (which is indirect, through marriage behavior). But as this revolution began to mature and the debate around the growth in the divorce rate was no longer really topical, other research questions emerged. How might divorce law impact negotiations with regard to specialization within couples, regarding premarital cohabitation, fertility, conjugal violence, female suicide, etc.?

Cross-References

- ▶ [Alimony](#)
- ▶ [Becker, Gary S.](#)
- ▶ [Coase, Ronald](#)

References

- Allen DW (1992) Marriage and divorce: comment. *Am Econ Rev* 82(3):679–685
- Bracke S, Mulier K (2017) Making divorce easier: the role of no-fault and unilateral revisited. *Eur J Law Econ* 43(2):239–254
- Brinig M, Buckley FH (1998) No-fault laws and at-fault people. *Int Rev Law Econ* 18(3):325–340
- Drewianka S (2008) Divorce law and family formation. *J Popul Econ* 21:485–503
- Ellman IM, Lohr SL (1998) Dissolving the relationship between divorce laws and divorce rates. *Int Rev Law Econ* 18(3):341–359
- Friedberg L (1998) Did unilateral divorce raise divorce rates? Evidence from panel data. *Am Econ Rev* 88(4):608–627
- Glenn ND (1997) A reconsideration of the effect of no-fault divorce on divorce rates. *J Marriage Fam* 59(4):1023–1025
- Glenn ND (1999) Further discussion of the effects of no-fault divorce on divorce rates. *J Marriage Fam* 61(3):800–802
- Gonzalez L, Viitanen TK (2009) The effect of divorce laws on divorce rates in Europe. *Eur Econ Rev* 53(2):127–138
- Gruber J (2004) Is making divorce easier bad for children? The long-run implications of unilateral divorce. *J Labor Econ* 22(4):799–833
- Kneip T, Bauer G (2009) Did unilateral divorce laws raise divorce rates in Western Europe? *J Marriage Fam* 71:592–607
- Leeson PT, Pierson J (2017) Economic origins of no-fault divorce revolution. *Eur J Law Econ* 43(3):419–439
- Mechoulan S (2006) Divorce laws and the structure of the American family. *J Leg Stud* 35:143–174
- Nakonezny PA, Schull RD, Rodgers JL (1995) The effect of no-fault divorce law on the divorce rate across the 50 states and its relation to income, education, and religiosity. *J Marriage Fam* 57(2):477–788
- Peters EH (1986) Marriage and divorce: informational constraints and private contracting. *Am Econ Rev* 76(3):437–448
- Peters EH (1992) Marriage and divorce, reply. *Am Econ Rev* 82(3):687–693
- Peterson R (1996) A re-evaluation of the economic consequences of divorce. *Am Sociol Rev* 61(3):528–536
- Rasul I (2005) Marriage markets and divorce laws. *J Law Econ Org* 22(1):30–69
- Rodgers JL, Nakonezny PA, Schull RD (1997) The effect of no-fault divorce legislation: a response to a reconsideration. *J Marriage Fam* 59(4):1026–1030

- Rodgers JL, Nakonezny PA, Schull RD (1999) Did no-fault divorce legislation matter? Definitely yes and sometimes no. *J Marriage Fam* 61(3):803–809
- Weitzman LJ (1985) *The divorce revolution: the unexpected social and economic consequences for women and children in America*. The Free Press, New York
- Wolfers J (2006) Did unilateral divorce raise divorce rates? A reconciliation and new results. *Am Econ Rev* 96(5):1802–1820
- Zelder M (1993a) The economic analysis of the effect of no-fault divorce law on the divorce rate. *Harv J Law Public Policy* 16(1):240–268
- Zelder M (1993b) Inefficient dissolution as a consequence of public goods: the case of non-fault divorce. *J Leg Stud* 22:503–520

Non-controlling Minority Interests

► Passive Minority Interests

Non-market Valuation

Sébastien Roussel¹ and Léa Tardieu²

¹Department of Economics, CEE-M, Université Montpellier, CNRS, INRA, SupAgro, Université Paul Valéry Montpellier 3, Montpellier, France

²BETA, Université Lorraine, INRA, AgroParisTech, Nancy, France

Definition

Non-Market valuation is a set of techniques that aims at reflecting the economic value of changes, in the availability or quality, of goods and services that are not intended to be traded in the market (e.g., health care, education, environment). The objective is to estimate the impacts of these changes on one's utility and by extension on the social welfare, in order to manage these goods and services by considering their true value to society.

Economic Values

In a neoclassical perspective, goods and services are valued in a utilitarian framework, i.e.,

individuals are rational, have various categories of desires and wishes (e.g., food, housing clothes...), and are able to classify them according to their preferences. They aim at reaching a maximum level of individual welfare according to their income constraint (utility maximization under budget constraint). Moreover, individuals are able to value the impact of an additional unit of good on their own welfare, which is decreasing as the units aggregate (decreasing marginal utility law). In the end, the economic value of any good increases with its usefulness and scarcity.

Markets generate the value of all traded goods and services as relative prices. Prices are therefore very useful in comparing different effects, as they give some indications of goods and services relative scarcity. It is assumed that values may be reflected in individuals' willingness to pay (WTP) or willingness to accept (WTA) associated with a change in the availability or quality of the goods (see Table 1). The social value of goods is then the sum of the marginal utilities of each unit for all of its users.

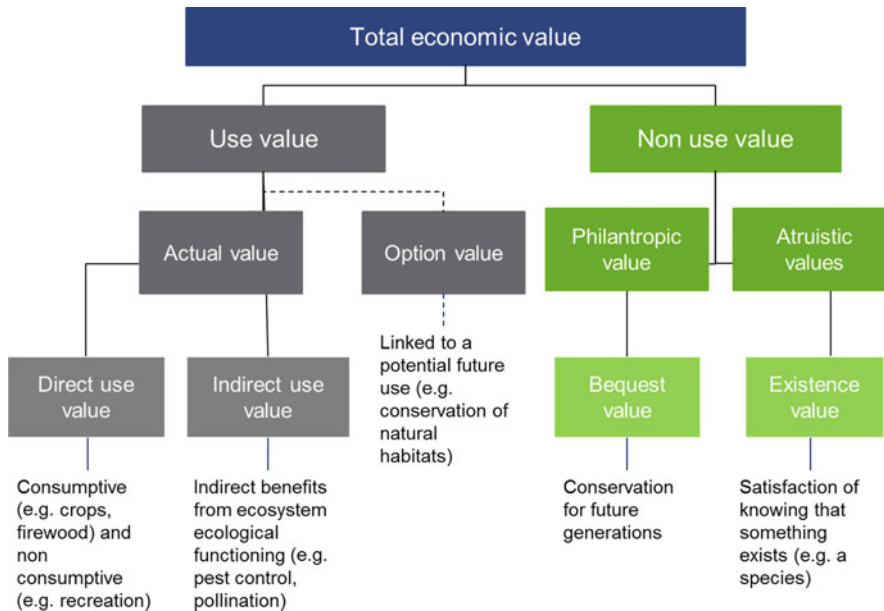
In a nutshell, economic values are mainly:

- Anthropocentric, since only the contribution to human well-being is taken into account
- Instrumental, as it is based on the utilitarian setting of the consequences of choices and actions and not on intrinsic or “deontological” values
- Subjective, because the value depends on individual's preferences

Economic values are classically divided into two types of values that make up the total economic value (TEV), in particular for environmental goods and services: use values, which include the benefits derived by an agent from a direct or

Non-market Valuation, Table 1 Willingness to pay (WTP) and willingness to accept (WTA)

	Improvement	Damage
Willingness To Pay (WTP) to...	Benefit from	Avoid a
Willingness To Accept (WTA) to...	Give up	Offset a



Non-market Valuation, Fig. 1 Use and nonuse values. (Adapted from TEEB 2010)

an indirect use of the good or service at stake, and also a potential option value; nonuse values, which reflect ethical or altruistic preferences (see Fig. 1).

Why Using Non-Market Valuation?

When the market is distorted or when goods and services have no market, market prices are poor indicators of the true economic value. As a consequence, how can one assign a monetary value to changes of goods and services that are not intended to be traded in the market (e.g., health care, education, environment)? The conventional view would be that nonmarketed goods and services are priceless, and a purely market rationale would act as if they had no value, especially while considering externality mechanisms. To tackle this issue, economists have developed a variety of methods that allow the construction of monetary indicators to value loss or gain associated with changes in the availability or the quality of these goods or services (see section “[Nonmarket Valuation Techniques](#)”).

In terms of law and economics, nonmarket valuation plays a crucial role to provide information to support public and private decisions in a

cost-benefit analysis (CBA) perspective (Hanley and Barbier 2009), in particular regarding high probabilities of harm and environmental justice concerns (Hsu 2005). For the effects at stake to be co-measurable, they are usually monetarized. Nevertheless, we may note that as argued by Hanley and Spash (1993), this is merely a device of convenience rather than an implicit statement that only money is meaningful.

In addition, nonmarket valuation can also help in environmental and resource economics at:

- Fixing the level of conservation effort
- Highlighting priorities in cost-effectiveness analyses by answering the question: “what is the best use of one euro invested in conservation?” in *ex-ante* analysis
- Communicating on the magnitude of global issues to promote awareness to policy makers and the general public
- Evaluating offset amounts in *ex-post* analysis

Non-Market Valuation Techniques

Nonmarket valuation techniques are used to re-build individuals WTP/WTa towards changes

Non-market Valuation, Table 2 Nonmarket valuation techniques. (Adapted from Tardieu 2017)

Economic techniques	Description
Market costs approaches	
Avoided damage costs	Uses the costs associated with the mitigation of a damage as the proxy for the value
Replacement costs	Uses costs of replacing a nonmarket service as a proxy for the value
Opportunity costs	Explicitly considers the value that is lost in order to protect, enhance or create a nonmarketed asset
Production function	Focuses on the (indirect) input costs of a particular good or service for the production of a market good corresponding to the nonmarket one
Revealed preferences	
Travel cost method	Uses data on people's actual behavior in real markets that are related to an environmental good. For example, the behavior studied is the number and distribution of trips that people make to outdoor recreation sites as a function of the cost of a trip. The travel cost is the weak complement (a complementary marketed good) of the outdoor recreation value
Hedonic pricing	Weak complementarity is assumed between the price of a property and the quality of the surrounding environment. For example, the nonmarket value is revealed through observations on the demand of residential properties
Stated preferences	
Contingent valuation	Estimates values by constructing a hypothetical market and asking survey respondents to directly report their willingness to pay to obtain a specified good or their willingness to accept giving up a good
Choice modeling	Based on a hypothetical market where respondents have a series of choice tasks in which they are asked to choose their preferred option (including <i>status quo</i>). Each option is described in terms of a set of attributes describing the good (including a price attribute) presented at various levels according to an experimental design. The analysis of the respondents' choices is based on Random Utility Maximisation (RUM) theory
Secondary valuation technique	
Benefit transfer	Uses economic information collected at a given area (study site), at a given time to make inference on nonmarket goods or services in another location (application site)

in the availability or quality of nonmarketed goods and services (see Table 2).

These techniques can be either direct or indirect. Direct methods are based on observable costs (avoided damage costs, replacement costs, opportunity costs) or allow deriving values from existing markets of close substitutes or inputs (production function). Indirect methods are based on individuals' revealed preferences through a substitute market (hedonic pricing, travel cost method), on individuals' stated preferences on an hypothetical market (contingent valuation, choice modeling), or on value transfer regarding nonmarkets goods and services from locations with similar characteristics (benefit transfer). By definition, each method has its flaws, as they either rely on individuals which may be misinformed, on proxies that may poorly represent agents' behavior, or finally on markets in which prices are poor indicators of the economic

value. Moreover, all presented methods are not indicated to assess all types of values. For instance, nonuse values can be exclusively assessed by stated preferences methods. Last, revealed preferences methods are indicated for *ex-post* analysis, while stated preferences methods are suitable for *ex-ante* analysis.

Nonmarket Valuation Limits and Criticisms

Although nonmarket valuation plays a critical role to provide information to support decisions, this gives rise to conceptual and ethical issues especially with regards to environmental goods and services. These issues lead to limits and criticisms with regards to, respectively, anthropocentrism, which is at the core of value building, and how

heterogenous components of the total economic value (TEV) can be effectively aggregated; value discounting regarding future uses; and the spatial dimension of environmental goods and services valuation, including ecosystem services (Tardieu 2017; Roussel 2018). Furthermore, issues arise also in law and economics regarding how the information from nonmarket valuation can be used to value environmental damages and how suitable financial offsets can be defined.

With regards to anthropocentrism, some limits do appear linked to possible confusion between an intrinsic value of nature, independent of humans, and the one derived from marginal changes linked to individual preferences. Moreover, there is also a reluctance to add benefits and costs without reference to their social distribution. In terms of values aggregation, one of the main difficulties is related to the heterogeneity of the methods used to derive these values (see section “[Nonmarket Valuation Techniques](#)”), which can be considered as not directly comparable. For example, the aggregation of use values measured by revealed preferences and nonuse values measured by stated preferences is relatively tricky. Connected to these issues arise technical difficulties as double-counting or the simultaneous economic valuation of uses that are socially or technically mutually exclusive (e.g., wood provisioning service and regulation service regarding carbon sequestration from forests ecosystems), as well as the degree of substitutability between so-called manufactured capital and of natural capital.

If we now turn to value discounting regarding future uses, there is a choice uncertainty when taking time into account, particularly when we consider environmental issues with far-reaching consequences in the future. This has a twofold consequence on valuation (Chevassus-au-Louis et al. 2009): on the one hand, the costs and benefits linked to far-reaching consequences tend to see their weight in decisions reduced; on the other hand, there is an uncertainty on these consequences as well as the behavior and expectations of future generations. Accordingly, the selection of a discount rate can be justified following two main reasons: present preference and the decrease in marginal utility with higher incomes for future

generations. Discounting can also be ethically criticized because discarding far-reaching consequences may lead to neglect critical issues with threshold (e.g., biodiversity).

In addition, the spatial dimension has often been excluded from economic valuation, leading to results far removed from reality. For example, taking into account the spatial dimension of ecosystem services in environmental economics implies measuring these services (supply and demand) according to their spatial context. The ecosystem services supply is indeed influenced by spatial variables such as climate variables, topography, soil type, hydrological conditions (Nelson et al. 2009). Simultaneously, the ecosystem services demand changes depending on the location of the ecosystem providing this service as well as the location of the potential recipients, the number of substitutes, and the accessibility (Bateman et al. 2006). In brief, the value will be influenced by three variables: supply (quantity and quality), demand (number of beneficiaries, uses, socio-economic characteristics), and the spatial context (complementary or substitutes goods).

Last, one can wonder about the reliability of the information from nonmarket valuation and how this can be used to value environmental damages and to define suitable financial offsets (Thompson 2002; Carson et al. 2003). The famous example of the Exxon Valdez oil spill in 1989 then led to nonuse value loss ranging from \$2.8 billion to \$7.19 billion (Carson et al. 2003), whereas use values loss were evaluated \$5 million (Hausman et al. 1995). The ensuing debate with the Panel chaired by Kenneth Arrow and Robert Solow (Arrow et al. 1993) is thus an iconic illustration of controversies over the past decades and on-going debates on nonmarket valuation and Nature monetization (Gómez-Baggethun et al. 2010; Costanza et al. 2014).

Cross-References

- [Cost–Benefit Analysis](#)
- [Ecosystem Services](#)
- [Externalities](#)
- [Public Goods](#)

References

- Arrow K, Solow R, Portney P, Leamer E, Radner R, Schuman H (1993) Report of the NOAA panel on contingent valuation. *Fed Regist* 58(10):1–67
- Bateman JJ, Day BH, Georgiou S, Lake I (2006) The aggregation of environmental benefit values: welfare measures, distance decay and total WTP. *Ecol Econ* 60(2):450–460
- Carson RT, Mitchell RC, Hanemann M, Kopp RJ, Presser S, Ruup PA (2003) Contingent valuation and lost passive use: damages from the Exxon Valdez oil spill. *Environ Resour Econ* 25:257–286
- Chevassus-au-Louis B, Salles J-M, Pujol J-L (2009) Approche économique de la biodiversité et des services liés aux écosystèmes. Contribution à la décision publique. La Documentation Française, Paris (Rapports et documents. Centre d'analyse stratégique)
- Costanza R, de Groot R, Sutton P, van der Ploeg S, Anderson SJ, Kubiszewski I, Farber S, Turner RK (2014) Change in the global value of ecosystem services. *Glob Environ Chang* 26:152–158
- Gómez-Baggethun E, de Groot R, Lomas PL, Montes C (2010) The history of ecosystem services in economic theory and practice: from early notions to markets and payment schemes. *Ecol Econ* 69(6):1209–1218
- Hanley N, Barbier EB (2009) Pricing nature: cost-benefit analysis and environmental policy. Edward Elgar, Cheltenham
- Hanley N, Spash CL (1993) Cost benefit analysis and the environment. Edward Elgar, Cheltenham
- Hausman JA, Leonard GK, McFadden D (1995) Utility-consistent, combined discrete choice and count data model assessing recreational use losses due to natural resource damage. *J Public Econ* 56(1):1–30
- Hsu S (2005) On the role of cost-benefit analysis in environmental law. *Environ Law* 35(1):135–174
- Nelson E, Mendoza G, Regetz J, Polasky S, Tallis H, Cameron D, Chan KMA, Daily GC, Goldstein J, Kareiva PM, Lonsdorf E, Naidoo R, Ricketts TH, Shaw M (2009) Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. *Front Ecol Environ* 7:4–11
- Roussel S (2018) Ecosystem services. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York. https://doi.org/10.1007/978-1-4614-7883-6_711 (in press)
- Tardieu L (2017) The need for integrated spatial assessments in ecosystem service mapping. *Rev Agric Food Environ Stud* 98:173–200
- The Economics of Ecosystems and Biodiversity (TEEB) (2010) In: Kumar P (ed) *The economics of ecosystems and biodiversity: ecological and economic foundations*. Earthscan, London/Washington, DC
- Thompson DB (2002) Valuing the environment: courts' struggles with natural resource damages. *Environ Law* 32(1):57–89

Non-observability

► [Information Deficiencies in Contract Enforcement](#)

Non-price Predation

► [Raising Rivals' Costs](#)

Non-verifiability

► [Information Deficiencies in Contract Enforcement](#)

Norms and Standardization

Mitja Kovac

Department of Economic Theory and Policy,
Faculty of Economics, University of Ljubljana,
Ljubljana, Slovenia

Abstract

The purpose of this chapter is to survey the academic literature on the economics of norms and standardization in the domain of administrative law and to synthesize their main themes. The chapter begins by introducing basic economic framework and continues addressing the issues of norms and tax compliance. Next, the chapter surveys some of the most important areas in the administrative law and economics literature. Topics include norms and environmental compliance, standards in traditional law and economics scholarship, the issues of norms, standards and regulatory functions of the state and the occupational health and safety standards and norms.

Definition

Standard is a written definition, limit, or rule, approved and monitored for compliance by an authoritative agency as a minimum acceptable benchmark. Standardization is a framework of agreements to which all relevant parties in an industry or organization must adhere to ensure that all processes associated with the creation of a good or performance of a service are performed within set guidelines. Standards occupy a middle position between information measures, representing low intervention and prior approval, representing high intervention. Standard is also the legal or social criterion that adjudicators employ to judge actions under particular circumstances. Standards provide a greater degree of flexibility to judges and allow them to consider fact-specific circumstantial evidence. Norms are behavioral regularities associated with a feeling of obligation supported by normative attitudes.

Introduction

One of the pathbreaking insights offered by the law and economics scholarship is the notion that contracts, laws, and all human relationships are, due to bounded rationality, positive transaction costs, and asymmetric information problem, necessarily incomplete. However, the legislator may choose the level of identified incompleteness of the enacted laws by formulating legislation with different degree of specificity. The law and economics literature refers to this choice as a choice between standards and rules (Luppi and Parisi 2011). Kaplow (1995) argues that legislator is actually, while choosing between rules and standards, balancing between *ex ante* or *ex post* content and should calculate costs associated with each option. Whereas Schäfer (2001) suggests that the use of rules rather than standards has the advantage of reducing corruption, concentrating human capital, and cutting down court delays caused by complex decisions. Obviously, his suggestion rests on an assumption that legislators are inherently less corrupt and better informed than judiciary. Moreover, one may argue that norms

and standards should be analytically seen as an economic institution developed to address the notorious positive transaction costs problem (Coase 1961; Marneffe et al. 2015) and the overwhelming market failures problem (Gómez-Barroso 2016).

According to the dominant law and economics scholarship quality standards which subject suppliers of goods and services to behavioral controls and which penalize those who fail to conform are the prevailing form of social regulation (Ogus 2004). Ogus in his encyclopedical work on economics of regulation stresses that standards have been applied, while pursuing different goals and employing different techniques, to a vast variety of commercial, industrial, and social activities (Ogus 2004). Standards can be subdivided into three categories which represent different degree of intervention: (a) target standards, (b) specification standard, and (c) performance standard. Extensive law and economics literature investigates different aspects of standards (more than 400 references) in relation to public interest justification, cost-effectiveness models for standard setting, legal formulation, and even, for example, on private interest considerations (Ogus 2004). Careful examination, however, reveals the lack of rigorous law and economics analysis of standards in narrowly defined administrative legal setting.

However, while one may argue that the economic role of standards in the domain of administrative law calls for further law and economics treatment, an overview of law and economics scholarship devoted to the study of norms reveals a vast amount of literature. McAdams and Rasmusen (2007) report that since the early 1990s, considerable scholarship in law and economics has turned its attention to norms. Numerous articles have addressed the power of social norms and their relevance for law (Ellickson 1991, 1998; Cooter 1996; Posner 2000a), have investigated the nature of norm's definition (Ellickson 1998; Picker 1997; Mahoney and Sanchirico 2003; Kaplow and Shavell 2002; Kahan 2001; Scott 2000), have shed the light on how norms work (Hirschelifer and Rasmusen 1989; Ellickson 1998, McAdams 1996; Buckley

2003), and have stressed the importance of norms to legal analysis (Posner and Rasmusen 1999; Epstein 2001). Moreover, there is an extensive scholarship on how norms affect welfare (Sunstein et al. 2000; Shavell 2002; Kaplow and Shavell 2002) and on the interaction between the law and norms (Posner 1996a, b; Shavell 2002). Furthermore, extensive law and economics scholarship addresses the specific application in the fields of tort law, contracts and commercial law, property, intellectual property, criminal law, family law, international law, and even constitutional law (McAdams and Rasmusen 2007).

In addition, norms have been fairly addressed also in the economic analysis of administrative law. For example, Hasen (1996) offers norms as a possible explanation for rational choice theories of voting behavior. Hasen argues that as with other types of socially beneficial behavior, individuals might receive small social rewards for voting or small sanctions for not voting, and hence voting rights might reflect the degree to which a community succeeds in socially beneficial behavior (Hasen 1996). Hansen suggests creation of a legal duty to vote and supplement informal, social incentives with legal sanctions (Hasen 1996).

Norms and Tax Compliance

Posner (2000b) started the discussion on whether strict enforcement of tax law is complementary with social norms and introduced a signaling as a compromise between the standard model and the approaches that try to make sense of social norms by complicating utility functions. His signaling model differs from the standard model only by introducing the plausible assumption that people have private information about their own tastes, including their discount rates (Posner 2000b). Posner's signaling model implies that if information were costless, social norms would not exist and consequently rejects the claim that social norms are internalized or that people feel guilty when they violate social norms (Posner 2000b). Lederman (2003) takes investigation further by exploring the relationship between enforcement

and compliance norms. She suggests that enforcement and compliance norms are complementary and that the enforcement can buttress norms-based appeals for compliance (Lederman 2003). Her study offers empirical evidence that there is a general societal norm of tax compliance in the United States but that, among certain groups, there may be a norm of noncompliance (Lederman 2003). She also suggests that enforcement will increase the number of people who obey the laws for prudential reasons (Lederman 2003). If one would create such a critical mass of taxpayers than the violation of the law will, according to Lederman (2003), it will create a disutility on the rest of population and hence create a tax payment incentive stream.

Norms and Environmental Compliance

The impact and interplay of norms have been also widely discussed in the relation to environmental compliance as part of general administrative procedure. Vandenberg (2003), for example, proposes a conceptual framework that accounts for the influence of norms on environmental decision-making and on corporate environmental compliance. Vandenberg assesses substantive norms of law compliance, human health protection, environmental protection, and autonomy and suggests that administrative enforcement policies should strive to harness those norms (Vandenberg 2003). Carlson (2001) analyzes types of the US local governments' strategies framed for inducing higher levels of individual environmental compliance. Local governments have actually, in order to achieve such targeted behavior, employed different novel forms of recycling. Carlson (2001), while assessing these new forms, argues that a lawmaker should focus on the policies that reduce the cost of recycling rather than those that try to change people's preferences. By employing such policies, lawmakers will, analytically speaking, make signaling more effective and will also introduce direct esteem toward those citizens that recycle their daily waste (Carlson 2001).

Standards in Traditional Law and Economics Scholarship

The traditional general economic rationale (outside contract law) for government involvement in standardization, as with several other government activities, has derived from the possibility of “market failure” and the public good character of standards. Left to its own devices, the market produces too little or too much standardization or standardization of the wrong sort (Swan 2000). Swan (2000) also points out that an important role of standardization is creation of a strong, open, and well-organized technological infrastructure that serves as a foundation for innovation-led growth. Standardization also increases competition and enable firms to reduce costs and increase quality. Standards might also spur the development of product and service markets that are based on the newest technologies (Swan 2000). Moreover, Swan (2000) also provides the most comprehensive source of estimates of the macroeconomic benefits of standardization and emphasizes the following ones as the most important: (a) competitiveness cannot be achieved by innovation alone, but requires efficient diffusion of innovation, and standardization plays a key role in that, (b) standards provide a positive stimulus to innovation, (c) standards contribute at least as much as patents to economic growth, (d) standards have a positive effect on trade and do not seem to act as barriers to trade, (e) international standards are more important than national standards in encouraging intra-industry trade, (f) standards enhance international competitiveness, and, finally, (h) the macroeconomic benefits of standardization exceed the benefits to companies alone.

Norms, Standards, and Regulatory Function of the State

It is well acknowledged that informal of formal norms, principles of conduct, tacit agreements, and standards play a pivotal role in the

coordination and harmonization of economic activity (Weigel 2006). Sometimes, as Weigel (2006) points out, norms and standards represent a framework for a more formalistic setup of legal rules, directives, and decrees. Such informal rules frequently delineate collective action which brings them into the domain of public law. As previously discussed standards and norms are instrumental in understanding of the most puzzling economic and social issues. Widespread technical standards and certifications which represent (or which should represent) a transaction costs minimizing mechanism could be, for example, listed as such examples, as such institutions that pursue allocative efficiency.

The most widely discussed regulatory regimes imposing standards are the ones of occupational health and safety, consumer protection/consumer products, and environmental pollution. In the domain of administrative law, Choi (1996) offers a two-period model (with merely two agents) and investigates the trade-off between ex post standardization and ex ante standardization. His model implies that in ex post period each user selects a technology from a random distribution of technologies (Choi 1996). Choi’s model actually identifies Nash equilibrium that indicates an excess of ex ante standardization and a deficit of ex post standardization if users choose to experiment in ex ante period.

Jeanneret and Verider (1996) in their subsequent investigation analyze conditions for compatibility in vertically differentiated demand where imported high-quality good compete with low-quality domestic good. They derive a range of tax rates which make it profitable for both suppliers to opt for compatibility. In such compatibility only and only the compatible goods will prevail. Jeanneret and Verider (1996) also note that these ranges depend on the size of the quality augmenting effect of compatibility. They also show that EU policies that subsidize switching costs are likely to fail if the subsidy scheme is not established before foreign and domestic firms decide on compatibility (Jeanneret and Verider 1996).

Moreover, such subsidy scheme has to be binding for a longer period so that policy is credible for the decisions horizon of the suppliers (Jeanneret and Verider 1996). Goerke and Holler (1998) interpret their insights and argue that EU standardization should show continuity. Goerke and Holler (1998) also argue that there might be a high potential for rent seeking in the organization of European standardization activities and standard setting. They stress the unwillingness of private agents to contribute to the financing of the standardization procedures as an obvious rent-seeking behavior's indicator. They also criticize the lack of financial arrangements on the side of EU commission and the lack of clearly assigned rights as the main sources of inefficient standard setting procedures (Goerke and Holler 1998). Hence, European Union, while providing an uncertain institutional framework (lack of exact duties, responsibilities, and influence in the standardization process) and while retaining the power to revert to detailed, interventionist standard setting, prevents companies, research institutions, chambers of commerce, consumer organizations, trade unions, and other stakeholders from making sufficient, standard-setting, commitments (Goerke and Holler 1998).

In addition, Weigel (2006) argues that the most crucial issue regarding norms and standardization in administrative law, that still awaits critical law and economics assessment, is the enforcement issue.

Occupational Health and Safety

The control of risks arising at the workplace occupies a special position among regulatory regimes and has a well-documented history (Thomas 1948). Ogas (2004) identifies the optimal loss abatement (optimal care) as the public interest economic goal of regulatory standards. He defines optimal care as the one where the marginal benefits equal marginal costs. However, as Ogas (2004) correctly observes, such an optimal care is rarely met. Employers have imperfect information of the risks involved in particular jobs and cannot

be assumed to make rational decisions. Hence, an unregulated market would rarely generate optimal safety standards of occupational health and safety. In order to mitigate identified market failures, administrative agencies developed several different regulatory strategies that should produce optimal occupational health and safety standards. The predominant strategy is the agency approach where an independent public agency is entrusted with the task of determining exclusively the standards which will correspond to the optimal level of care in the particular workplace circumstances. Agencies actually mimic the idealized labor market, taking into account also distributional considerations (Ogas 2004). Under the quasi-market approach the primary responsibility for setting the standards is entrusted to the employer. In such model, administrative agency plays merely a residual role and scrutinizes the standards emerging from such practices to verify that they are consistent with the goal of optimal safety (Ogas 2004). Such type of standard-setting attempts to foster market-type solutions and external, by administrative agency-imposed standards, is only enforced in the instances of failures of such a quasi-market approach.

Concluding Remarks

The traditional general economic rationale (outside contract law) for government involvement in standardization, as with several other government activities, has derived from the possibility of "market failure" and the public good character of standards. Standards have been actually applied, while pursuing different goals and employing different techniques, to a vast variety of commercial, industrial, and social activities. However, while standards and norms have been thoroughly addressed in more than 400 references in economic literature and also produced a vast amount of law and economics literature, the assessment of public/administrative law actually reveals surprising lack of systematic assessment. Identified gap in the law and economics of

standards and norms in the domain of administrative law opens doors for an extensive law and economics treatment that would offer sets of legal and economic arguments for an improved regulatory response.

References

- Buckley FH (2003) The morality of laughter. University of Michigan Press, Ann Arbor
- Carlson A (2001) Recycling norms. *Calif Law Rev* 89:1231–1300
- Choi JP (1996) Standardization and experimentation: Ex ante vs. ex post standardization. In: Holler MJ, Thisse JF (eds) *The economics of standardization*, Special issue of the European journal of political economy, vol 12. Elsevier, Amsterdam, pp 273–290
- Coase R (1961) The problem of social cost. *J Law Econ* 3:1–44
- Cooter RD (1996) Decentralized law for a complex economy: the structural approach to adjudicating the new law merchant. *Univ Pennsylvania Law Rev* 144:1643–1696
- Ellikson RC (1991) Order without Law: how neighbors settle disputes. Harvard University Press, Cambridge
- Ellikson RC (1998) Law and economics discovers social norms. *J Leg Stud* 27:537–552
- Epstein JM (2001) Learning to be thoughtless: social norms and individual computation. *Comput Econ* 18(1):9–24
- Goerke L, Holler MJ (1998) Strategic standardization in Europe: a public choice perspective. *Eur J Law Econ* 6:95–112
- Gómez-Barroso JB (2016) Market failure (analysis). In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York, pp 1–5
- Hasen RL (1996) Voting without law? *Univ Pennsylvania Law Rev* 144:2135–2179
- Hirschelifer D, Rasmusen E (1989) Cooperation in a repeated prisoner's dilemma with ostracism. *J Econ Behav Organ* 12:87–106
- Jeanneret MH, Verdier T (1996) Standardization and protection in a vertical differentiation model. In: Holler MJ, Thisse JF (eds) *The economics of standardization*, Special issue of the European journal of political economy, vol 12. Elsevier, Amsterdam, pp 273–290
- Kahan DM (2001) The limited significance of norms for corporate governance. *Univ Pennsylvania Law Rev* 149:1869–1900
- Kaplow L (1995) A note on subsidizing gifts. *J Public Econ* 58(3):469–477
- Kaplow L, Shavell S (2002) *Fairness versus welfare*. Harvard University Press, Cambridge
- Lederman L (2003) The interplay between norms and enforcement in tax compliance. *Ohio State Law J* 64:1453–1513
- Luppi B, Parisi F (2011) Rules versus standards. In: Parisi F (ed) *Production of legal rules*. *Encyclopedia of law and economics*, vol 7. Edward Elgar, pp 43–53
- Mahoney PG, Sanchirico C (2003) Norms, repeated games and the role of law. *Calif Law Rev* 91:1281–1329
- Marneffe W, Bielen S, Vereeck L (2015) Transaction costs. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York, pp 1–6
- McAdams RH (1996) Groups norms, gossip and blackmail. *Univ Pennsylvania Law Rev* 144:2237–2292
- McAdams RH, Rasmusen EB (2007) Norms and the law. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*, vol 2. North-Holland, Amsterdam
- Ogus AI (2004) *Regulation: legal form and economic theory*. Hart Publishing, Oxford
- Pickier RC (1997) Simple games in a complex world: a generative approach to the adoption of norms. *Univ Chic Law Rev* 64:1225–1287
- Posner EA (1996b) The regulation of groups: the influence of legal and non-legal sanctions on collective action. *Univ Chic Law Rev* 63:133–197
- Posner EA (1996) The legal regulation of religious groups. *Legal Theory* 2(1):33–62
- Posner EA (2000a) *Law and social norms*. Harvard University Press, Cambridge
- Posner EA (2000b) Law and social norms: the case of tax compliance. *Virginia Law Rev* 86:1781–1820
- Posner RA, Rasmusen E (1999) Creating and enforcing norms, with special reference to sanctions. *Int Rev Law Econ* 19:369–382
- Schäfer HB (2001) Enforcement of contracts. *DSE World Bank. Inst Found Mark Econ* 3:141–146
- Scott RE (2000) The limits of behavioral theories of law and social norms. *Virginia Law Rev* 86:1603–1645
- Shavel S (2002) Law versus morality as regulators of conduct. *Am Law Econ Rev* 4:227–257
- Sunstein CR, Schkade D, Kahneman D (2000) Do people want optimal deterrence? *J Leg Stud* 29:237–253
- Swan GMP (2000) *The economics of standardization*. Standards and Technical Regulations Directorate Department of Trade and Industry, United Kingdom
- Thomas MW (1948) *The early factory legislation: a study in legislative and administrative evolution*. Thames Bank Pub. Co, Leigh-on-Sea
- Vandenbergh M (2003) Beyond elegance: a testable typology of social norms in corporate environmental compliance. *Stanf Environ Law J* 22:55–144
- Weigel W (2006) Why promote the economic analysis of public law? *Homo Econ* 23(2):195–216

Novelty

► [Innovation](#)

Nudge

A Critical Perspective

Angela Ambrosino¹, Valeria Faralla² and Marco Novarese³

¹Department of Economics and Statistics Cognetti de Martiis, University of Turin, Torino, Italy

²Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università del Piemonte Orientale, Alessandria, Italy

³Department of Law and Economics, University of Eastern Piedmont, Centre for Cognitive Economics, Alessandria, Italy

Abstract

In recent years, the nudge approach and the choice architecture design have become popular in policy and academic circles. The aim of this entry is to present a general overview of the theory of nudges and a critical appraisal of its application to practice, in policy and program as well as in research design, also with respect to possible alternative approaches.

Following the publication of “Nudge: Improving Decisions about Health, Wealth, and Happiness” (Thaler and Sunstein 2008), the term *nudge* has entered the economic and juridical focus in relation to public policies. Nudge refers to the idea that people’s behavior can be mildly, or gently, pushed towards a certain course of action. As a result, nudge can be considered as a tool of political economics, using individual cognitive characteristics to stimulate people towards a positive action without restricting their freedom of choice. This is an alternative form of intervention to other government tools of public policies, such as incentives, rules, and constraints, or education and empowerment. The proponents of this theory consider nudge as a less coercive tool than regulatory efforts and incentive-based systems, although it is considered to be associated with better results if compared with free markets or traditional informative campaigns. The debate is

deeply connected with the debate on paternalistic policies and the regulation approach and is also tied with bounded rationality. However, the interplay between all these aspects makes the definition of nudge problematic. Thaler and Sunstein, indeed, propose a definition which is still not overwhelmingly accepted (Rebonato 2012).

Policies inspired by nudge appear to be cheap and have the advantage of being experimentally testable and measurable (even if these policies can be tied with cultural and political issues such as the effects of the electricity consumption). As a result, nudge represents a novelty in the field. On the other hand, however, the use of nudges in real-life situations may involve the need to lie (or at least hide something) about the aim of actions that people are asked to perform. This clearly represents a point against the use of nudges.

The nudge theory is built on behavioral economics and on the idea that people have a bounded rationality, often do not have well-defined preferences, and are subject to a number of biases which cause people to make choices not in their own interest, sometimes also against their voluntary will. As for willpower, several examples exist: the desire to save and study, to consume healthier food, or avoid gambling activities. The desired action can be frustrating when lack of willpower is in action (Elster 1979).

But nudge is also based on other aspects linked to the decision-making mechanism. In many situations, indeed, people do not have well-defined preferences and, therefore, decisions become more complex. When there is a lack of clear preferences or inability to perform statistical computations and follow logic rules, people seem to be guided by simple rules (i.e., heuristics; see Kahneman 2012) in order to save mental energies. Specifically, the rational system of decision-making appears to be challenged (and often won) by a more impulsive system, faster, which is guided by simplifying strategies, such as information availability (the ease with which the side of a problem can be visible), and would give rise to distortions and simple heuristics. These distortions and heuristics tend to lead to systematic errors. The rational system (i.e., system 2) is connected with conscious and deliberate

decisions, whereas the other system (i.e., system 1) is highly unconscious and automatic. As a result, aspects which characterize the context of choice such as the number of alternatives available to individuals, the order of presentation of these alternatives, the way in which the choice is presented and framed, the existence of a default option, or the presence of a status quo alternative, which are typically considered as apparently irrelevant features become significant.

Nudge is then added to the debate over human rationality and about the influences on the decision-making process. A rational agent, as described by standard neoclassical normative models (von Neumann and Morgenstern 1947), is not affected by nudge intervention. However, in real-life situations, people appear to be not so rational (Kahneman 2003). In particular, the way through which the choice is designed and presented, the architecture of choice, becomes relevant.

Accordingly, the same mechanisms connected with the systematic errors can be used to direct people towards more healthy behaviors.

Nudge, which in this view represents a development of behavioral economics, has been already anticipated by a series of papers on the framing effect which has been found to elicit certain types of choices, rather than others, that are not necessarily for the good of individuals (e.g., marketing purposes). In this case, knowledge about decision-making processes can be used to persuade people (see Cialdini 2001a, b). The relationship between behavioral economics, law, and public policies developed early, also thanks to Kahneman (e.g., Kahneman et al. 1998; McCaffery et al. 1995). More recently, Parisi and Smith (2005) review the various implications of behavioral and cognitive economics for the law, such as the possibility to use cognitive biases to develop an approach that could be more effective than using monetary and nonmonetary rewards and punishments.

The subtitle of Thaler and Sunstein's book, "Improving Decisions about Health, Wealth, and Happiness," mainly refers to the measures aimed at pushing the limits of willpower. In front of a plate of peanuts, people tend to not restrain and

eat, even when they do not want or do not need to. The act of pushing the plate in an unhandy place, which can be considered as a pre-commitment strategy, is an example of nudge. In their book, the authors suggest to take away the peanuts, but this appears to be a command-and-control form regulation, which is a mandate that cannot be avoided. This is in fact an example of the difficulty to define a nudge.

Moreover, the two authors propose simple examples of nudge that do not seem to improve individuals' well-being, at least in a direct and clear way. For example, the choice of an opt-in mechanism, rather than opt-out (i.e., explicit versus presumed consent), seems to increase organ donation rates because of the human tendency to remain at the status quo (Thaler and Sunstein 2008). The choice of a certain type of mechanism represents an example of choice architecture through which a public policy can be implemented and that can be considered as an alternative to public awareness campaign or strong enforcement. In fact, it seems to be true that nudging leaves margins of freedom to individuals and does not bind people to an action unless they have a clear opinion in one direction or another. People seem to be influenced by nudges when they have not a clear opinion and, rather, tend to conform to social norms and what they perceived as normality.

Considering the difference between various policies, this example shows the difficulty with the current definition of nudge but also the complexity of the discussion around the architecture of choice. A more general definition considers nudge as a tool for implementing public policies by means of behavioral economics.

Nudges' problems and critics come from different perspectives.

In their website, Thaler and Sunstein display the ballot used by Hitler as an example of bad nudge (Nudge Blog 2010). As stated before, nudge can indeed be used in order to persuade people to the right, as well as the wrong, direction. The problem is clearly that of understanding who decides what is the right/wrong direction. Thaler and Sunstein lead the main debate to the issue of strong paternalism and soft paternalism and to the

fairness of one approach over another. Strong paternalism assumes that the state knows and decides what is the best for people and then constrains them to have a coherent, consistent behavior. Strong paternalism is also connected with the idea of an omniscient legislator that acts rationally and is perfectly capable of using nudges and push people towards efficient behaviors. Soft paternalism, or libertarian, is instead based on the idea that we have to help people in getting their best interest in terms of well-being. Here is the heart of the debate between strong and soft paternalism and about the possibility to implement a soft, libertarian paternalism by means of a theory of nudges.

Meanwhile, the opponents of this theory consider nudge as a dangerous tool because, differently from more explicit and direct forms of regulation, it operates through unconscious channels (Kahneman 2012).

In line with other critical-oriented works, referring to different example of nudging, Rayner and Lang (2011) claim that nudge is not a new policy. In particular, they highlight how similar tools tend to deny the general idea of politics that requires informed choices and discussions, and deals with problems in an explicit and direct way. They also underline the fact that nudge can be applied to avoid taking actual choices (Rayner and Lang 2011).

Bovens (2009) precisely criticizes the lack of transparency of nudges and, particularly, the fact that choice architecture does not trigger real changes in individual preference or improve the use of willpower. More recently, Vallgård (2012) criticizes the libertarian paternalism and denies the possibility to find an ethical reason for nudge, at least from a libertarian perspective. Furthermore, nudge would be anything new.

The fact that the theory of nudges exploits the bounded rationality is also controversial. The definition of nudge given by Rebonato (2012) underlines this aspect, which is instead not corroborated by Sunstein (2014).

Is it really impossible to develop consciously accepted forms of nudge and then give legitimacy to this type of intervention? Loewenstein et al. (2015) warn their subjects of the presence of a nudge, and in particular of a default option, in

choosing between different medical treatments. The effect of the default option seems to persist, even when individuals are informed about its use. As a result, the effect was not necessarily connected with a lie.

Colander and Chong (2010) discuss instead about the benefits of giving directly to people, rather than to the government or an economist, the choice of being nudged to obtain the best result. This is the path to an actual libertarian paternalism, where individuals can freely choose their choice architecture.

A different criticism refers to the effectiveness and the strength of nudge. For instance, Loewenstein and Ubel (2010) address this question in an article published in the *New York Times*. The question was about the reduction of electricity consumption and the preference for a nudge, rather than an incentive-based intervention (i.e., by charging higher costs), to approach the problem. Does the fact that a particular nudge intervention has been found to have an effect in controlled experiments suggest that these effects can be observed in real-life situations? Even more important, does statistical significant results imply a real impact of nudge interventions?

This article also proves the importance of this issue, considering that it has been taken out of academic publishing and issued in one of the most important newspapers in the USA. In the UK, for instance, a nudge unit has been created. Furthermore, the Obama administration has pointed out the need for a wider use of behavioral economic techniques in the implementation of public policies.

But if people are wrong, is there a guarantee that the governments are not going to make similar errors? Other than ethical issues, with what legitimacy can they inspire paternalistic policies?

Grüne-Yanoff and Hertwig (2016) criticize the use of nudges in the perspective of an evolutionary rationality. After all, the bounded rationality is a critic to the economic mainstream which is considered as a reference point for a hypothetical rationality, one from which we go away. However, according to the evolutionary rationality, this critic should not imply that people make systematic mistakes and are basically unable to take

decisions. Moreover, people are not necessarily driven by nonmodifiable and unconscious mechanisms, as those represented by the alleged system 1. In certain circumstances, indeed, simple rules and heuristics can be smart and effective. In everyday life, we need rapid decisions and, differently from the theory, logic is not always applicable. Rather than rationality, we need the ability to survive and to be clever and also to interact with the environment. Errors can be made because the environment has changed over millennia, for human beings as well as other living creatures. Furthermore, errors can also depend from an incorrect way of framing problems. Specifically, the use of conditional probabilities in describing the Bayes' Theorem makes hard to handle it, even for experts. Considering the probabilities as relative frequencies can instead make people understand the actual risk they face, giving them an effective freedom of choice. Problems need to, and can, be formulated in a more clear and understandable way. The theory of nudge also moves from an idea of a mainstream behavior. That rationality is, however, not applicable to real life, which is instead characterized by different situations that cannot be simply described in terms of calculable risks and probabilities.

As a result, as an alternative to nudges, other authors propose the boost approach, whose aim is to extend people's decision-making skills and refine the decision-making environment rather than focus on distortions and biases. In particular, Grüne-Yanoff and Hertwig (2016) refer to nudge and boost policies as two different research programs. The first program is based on the analysis of systematic cognitive heuristics and biases, which are considered to result in poor choices, whereas the second is a simple heuristics program where bounded rationality is considered in its capacity, rather than inability, to produce good inference and choice.

Conclusion

Nudge is receiving an increasing and strong attention for both the development of the behavioral approach and the possibility (which the nudge

approach seems to offer) to realize low-cost public policies, whose effectiveness can be proven.

It is hard to believe that the political debate around the paternalistic approach can ultimately have a solution. Indeed, this is a long-term discussed issue and nudge tends to complicate, rather than simplify, its resolution.

The problem lies in the fact that different forms of nudge exist and each form is associated with different critical issues. At the same time, however, most discussions tend to simplify and take into account only specific examples.

Kahneman (2012) underlines how writing with a difficult-to-read font can focus attention and decrease the possibility of errors occurring in the solution of a problem. This is an example of nudge that increases the ability of people to influence for the better the relationship between individual well-being and choices.

Cross-References

- [Behavioral Law and Economics](#)
- [Financial Education](#)

References

- Bovens L (2009) The ethics of nudge. In: Grüne-Yanoff T, Hansson SO (eds) *Preference change: approaches from philosophy, economics and psychology*. Theory and decision library A(42). Springer, New York, pp 207–220
- Cialdini RB (2001a) The science of persuasion. *Sci Am* 284(2):76–81
- Cialdini RB (2001b) *Influence: science and practice*. Allyn & Bacon, Boston
- Colander D, Chong AQL (2010) The choice architecture of choice architecture: toward a Non-paternalistic nudge policy. Department of Economics Middlebury College Middlebury, Middlebury college economics discussion paper no. 10–36. Available at <http://sandcat.middlebury.edu/econ/repec/mdl/ancoec/1036.pdf>
- Elster J (1979) *Ulysses and the sirens*. Cambridge University Press, Cambridge
- Grüne-Yanoff T, Hertwig R (2016) Nudge versus boost: how coherent are policy and theory? *Minds Mach* 26(1):149–183
- Kahneman D (2003) Maps of bounded rationality: a perspective on intuitive judgement and choice. *Am Econ Rev* 93(2):162–168
- Kahneman D (2012) *Thinking, fast and slow*. Penguin Books, London

- Kahneman D, Schkade D, Sunstein C (1998) Shared outrage and erratic awards: the psychology of punitive damages. *J Risk Uncertain* 16(1):49–86
- Loewenstein G, Ubel P (2010) Economics behaving badly. *N Y Times*, July 14. Retrieved from <http://www.nytimes.com/2010/07/15/opinion/15loewenstein.html>
- Loewenstein G, Bryce C, Hagmann D, Rajpal S (2015) Warning: you are about to be nudged. *Behav Sci Policy* 1(1):35–42
- McCaffery EJ, Kahneman D, Spitzer ML (1995) Framing the jury: cognitive perspectives on pain and suffering awards. *Virginia Law Rev* 81(5):1341–1420
- Nudge Blog (2010) The Nazis nudged. [Blog post], August 10. Retrieved from <http://nudges.org/2010/08/11/the-nazis-nudged/>
- Parisi F, Smith VL (2005) The law and economics of irrational behavior. Stanford University Press, Stanford
- Rayner G, Lang T (2011) Is nudge an effective public health strategy to tackle obesity? *No. Br Med J* 342: d2177
- Rebonato R (2012) Taking liberties: a critical examination of libertarian paternalism. Palgrave Macmillan, New York
- Sunstein CR (2014) Why nudge? The politics of libertarian paternalism. Yale University Press, New Haven
- Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
- Vallgård S (2012) Nudge: a new and better way to improve health? *Health Policy* 104(2):200–203
- von Neumann J, Morgenstern O (1947) Theory of games and economic behaviour. Princeton University Press, Princeton

Nuisance

Keith N. Hylton

Boston University School of Law, Boston, MA, USA

Abstract

This entry sets out the law and the economic theory of nuisance. Nuisance law serves a regulatory function: it induces actors to choose the socially preferred level of an activity by imposing liability when the externalized costs of the activity are substantially greater than the externalized benefits or not reciprocal to other background external costs. Proximate cause doctrine plays a role in supplementing nuisance law.

Introduction

Nuisance law has suffered from the difficulty scholars have encountered in attempting to codify it in the form of simple rules and to understand the functions that the law serves. Prosser (1971) once described nuisance as an impenetrable jungle, and the dearth of efforts to state it in the form of black letter rules suggests that this opinion has been shared by many legal scholars. The process of scholarly codification, that is, of taking a mass of seemingly inconsistent court decisions and generating from them a set of clear legal propositions, has been slow in the area nuisance law.

Scholarly codification and an understanding of function are likely to occur contemporaneously. When courts and legal scholars have a firm understanding of the functions of a legal doctrine, it is relatively easy for them to summarize it in the form of simple rules. For example, the “Hand Formula” of *US v. Carroll Towing* is a summary of the negligence test that reflects a widely accepted understanding of the function of negligence law.

In recent years, research has focused on the function of nuisance law. I will set out the functional approach here, which uses economic analysis to understand nuisance doctrine at a high level of detail (Hylton 2011).

Earlier efforts have been made to provide an economic theory of nuisance law. Most of those early efforts, stemming from Coase (1960), have relied on the theory of transaction costs to explain the functional distinction between nuisance and trespass law (Calabresi and Melamed 1972; Merrill 1985; Smith 2004). But the core of nuisance law consists of balancing tests and limitations on the scope of liability that are not easily understood on the basis of transaction-cost theory. These tests and limitations are better understood using externality analysis.

Far from being an impenetrable jungle, nuisance law is a coherent body of rules that serves a socially desirable function. Nuisance law optimally regulates activity levels. The law induces actors to choose the socially optimal level of an activity by imposing liability when the externalized costs (of the activity) are substantially in

excess of externalized benefits or far in excess of background external costs. Proximate cause doctrine plays an important supplementary role to nuisance doctrine in regulating activity levels.

Nuisance Law

A nuisance is typically defined as an intentional, unreasonable, nontrespassory invasion of the quiet use and enjoyment of property. Each one of these terms has a special meaning in the law. Most of the terms are easily understood in terms of their general use in tort law. However, the term “unreasonable” is the most tricky concept, because there is equivocal use of the same term in other parts of tort law.

Take, for example, the word “intentional” in the definition of a nuisance. Intentional in nuisance law has a meaning that is not very different from its meaning in other areas of tort law. Typically the defendant is guilty of an intentional nuisance if he is aware of the invasion. The law does not require the defendant to have set out to harm the plaintiff.

Similarly, nontrespassory has a meaning that is readily ascertainable from the tort’s case law. A trespassory invasion is one that displaces the plaintiff from all or some portion of his property. For example, a large rock that is thrown over to the plaintiff’s property displaces the plaintiff from the space in which it travels and ultimately lands. This can be contrasted with a nontrespassory invasion, such as smoke or noise, which does not displace or oust the plaintiff from any space on his property.

The difficulty arises with the term “unreasonable.” As a result, efforts to state nuisance law in the form of simple rules have been sparse and for the most part unsuccessful. The best known effort to codify nuisance doctrine is the Restatement Second of Torts § 826, 1977, which says:

An intentional invasion of another’s interest in the use and enjoyment of land is unreasonable if:

- (a) the gravity of the harm outweighs the utility of the actor’s conduct, or
- (b) the harm caused by the conduct is serious and the financial burden of compensating for this and similar harm to others would not make the continuation of the conduct not feasible.

However, the Second Restatement’s § 826 is of questionable value because it refers to the actor’s *conduct* rather than his *activity*. One can draw an important distinction between these terms. Conduct often refers to an action or a series of actions within a short time span. Activity refers more broadly to an occupation or a significant pastime. For example, batting a baseball is a type of conduct, while playing professional baseball is an activity.

The reference to conduct could easily lead readers to believe that Restatement § 826 is equivalent to the balancing test observed in negligence law – i.e., the Hand Formula of *Carroll Towing*. The balancing test known as the Hand Formula says that the defendant is negligent if he fails to take care in a setting where his additional care would have been less costly than the additional injury costs that would have been avoided by that care. The language of Section 826 is easy to confuse with the analysis required by the Hand Formula.

If, instead, “conduct” in Restatement § 826 is understood to mean “activity,” then it becomes difficult to understand how the balancing test announced in 826 should be conducted. How is it possible to compare the gravity of the victim’s harm to the utility of the defendant’s activity? Suppose, again, that we are talking about baseball. A ball is hit out of the baseball yard and injures a passerby on the street. How should one go about comparing the gravity of the victim’s injury to the utility of playing baseball? Because this is so difficult to answer, Section 826 provides little guidance to lawyers and judges.

Moreover, part (b) of Restatement § 826 implies that strict liability should be applied to any activity that causes a “serious” interference with the plaintiff’s use and enjoyment of property, provided that the activity would not be bankrupted by such liability. This implies, strangely, that a thinly capitalized activity has an advantage under nuisance law, because it appears to immunize activities that would be bankrupted by a claim for damages. The difficult question in nuisance law is how to balance the external risks and the external benefits of an activity, a question which Section 826 does not even begin to address.

The following test, based on Restatement Second § 520, provides a better summary of nuisance doctrine:

In order to determine if an invasion is unreasonable under nuisance law, the following factors should be examined:

- (a) Existence of a high degree of interference with the quiet use and enjoyment of land of others
- (b) Inability to eliminate the interference by the exercise of reasonable care
- (c) Extent to which the activity is not a matter of common usage
- (d) Inappropriateness of the activity to the place where it is carried on
- (e) Extent to which its value to the community is outweighed by its obnoxious attributes

In the remainder of this entry, I will set out the basic economic theory of nuisance doctrine and explain why it is generally consistent with these factors.

Economics of Nuisances

The literature on the economics of nuisance law can be divided into two branches. One is the *transaction-cost framework*, which began with Coase's discussion of nuisance in his famous article on transaction costs and resource allocation. The transaction cost approach emphasizes the functional differences between nuisance and trespass law and provides a positive theory of the boundary between nuisance and trespass (Merrill 1985; Smith 2004). It has also been applied to explain the law on priority, often described as "coming to the nuisance" (Wittman 1980; Snyder and Pitchford 2003).

The other branch of work on the economics of nuisance law can be labeled the *externality model*, which focuses on the regulatory function of nuisance law (Hylton 1996, 2008, 2010). The notion that liability rules can be used to control externalities has been well understood for a long time in the law and economics literature (Polinsky 1979). The externality approach offers a model of the function of nuisance liability and a positive theory

of the core doctrines of nuisance. The core doctrines examined under the externality model are those of intent, reasonableness, and proximate cause. I will focus on the externality model below and offer a few remarks reconciling the transaction cost and externality models at the end.

Activity Levels, Care Levels, and Externalities

The law and economics literature distinguishes care and activity levels (Shavell 1980). The care level refers to the level of instantaneous precaution that an actor takes when engaged in some activity. For example, an actor can take more care while in the activity of driving by moderating his speed or looking more frequently to both sides of the road. The activity level refers to the actor's decision with respect to the frequency or location of his activity. If, for example, the activity of concern is driving, it can be reduced by driving less frequently.

The invasions associated with nuisance law are external costs connected with activity level choices. Consider, for example, a manufacturer who dumps toxic chemicals into the water, as a byproduct of manufacturing. Suppose the manufacturer is taking the level of care required by negligence law (reasonable care), and, in spite of this, the manufacturing process leads to some discharge of toxic chemicals. In this case, the environmental harm is a negative externality associated with the manufacturer's activity level choice.

The framework below is of activities that impose external costs on society even when they are carried out with reasonable care (Hylton 2008). The question examined is how the law can regulate activity levels in a way that leads to socially optimal decisions.

The Economics of Activity Level Choices

Assume that there are two liability rules that can be applied to actors, strict liability and negligence. Under either rule, actors are assumed to take reasonable care.

For any activity, the actor engaged in it will set his privately optimal activity level at the point which maximizes his utility from that activity. That means the actor will consider the benefits

he derives from the activity as well as the costs and choose a level at which the excess of private benefits over private costs is at its maximum. If we let MPB represent the incremental or marginal private benefits to the actor from his activity, and MPC represent the incremental private costs to the actor from increasing the scale of activity, the actor will increase his activity level as long as the marginal private benefit of an additional unit of activity exceeds the marginal private cost ($MPB > MPC$). The privately optimal level of activity is the level at which the marginal private benefit to the actor is just equal to the marginal private cost ($MPB = MPC$).

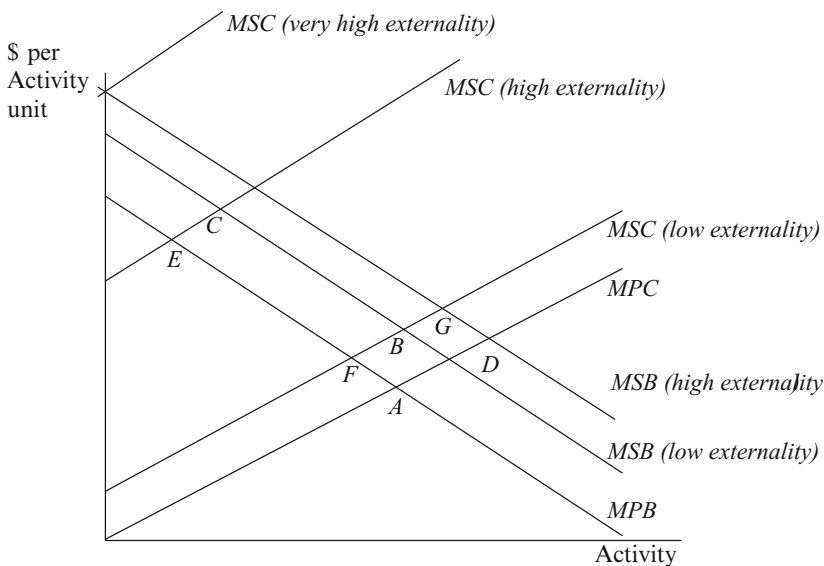
The diagram labeled Fig. 1 can be used to illustrate this argument. Assuming marginal benefits diminish as the actor increases his activity level, the marginal private benefit schedule can be represented by a downward sloping line, as shown in Fig. 1. The marginal private cost schedule is assumed to increase as the actor increases his level of activity (see MPC in Fig. 1). The reason for this is that the incremental cost of the activity goes up as the actor increases his scale. For example, if the activity is driving, the upward-sloping MPC schedule assumes that it is more costly to go from 50 miles per week to 51 than to go from 10 miles per week to 11. (Of course, this assumption may

not be valid in some cases. The incremental cost of going from 50 to 51 miles per week may be the same, in some cases, as the incremental costs of going from 10 to 11, but the results of this analysis are not dependent on this assumption of increasing marginal cost.)

The actor's privately optimal activity level choice is given by the intersection of the marginal private benefit and marginal private cost schedules, shown by point A in Fig. 1. At the intersection point, the net benefits (excess of private benefits over private costs) are at its maximum.

Now I introduce externalities. On the cost side, there are negative externalities (or external costs) associated with many activities. Consider, for example, driving. With each mile driven, the actor imposes some risk of harm from an accident or from pollution on the public in general. Or, if the activity is manufacturing, with each widget produced, a manufacturer who discharges chemicals in the water imposes cleanup costs on others. The marginal social cost of the actor's activity is simply the sum of the marginal private cost and the marginal external cost imposed on society.

On the benefit side, it is possible that there are benefits to society generated by the actor's activity. In the manufacturing case, suppose that instead of producing widgets, the manufacturer is



Nuisance, Fig. 1 Activity levels and externalities

producing a vaccine for some communicable disease. Vaccines cast off substantial external benefits by reducing the risk of disease even to the unvaccinated. The marginal social benefit is the sum of the marginal private benefit and the marginal external benefit of an additional unit of activity.

The final step of this introduction to the economics of activity level choices is to consider the differences between private and social incentives. Consider the case of low externalities on both the cost and benefit sides. Suppose there are external costs and external benefits connected to the activity, but they are relatively modest. They are shown in Fig. 1 by *MSC (low externality)* and *MSB (low externality)*. The socially optimal level of activity, which equates marginal social benefit and marginal social cost, is found at the point *B* in Fig. 1. The socially optimal level of activity (*B*) is roughly the same as the privately optimal level of activity (*A*). The reason is that the modest positive and negative externalities cancel each other out. Given this, there is no reason for the law to intervene to try to reduce the level of activity.

The case just considered is similar to that of an “irrelevant externality” (Buchanan and Stubblebine 1962; Haddock 2007). Although there is an external cost, society should not try to correct it because there is an offsetting external benefit. Buchanan and Stubblebine emphasized the case of offsetting internal benefit, but the concept of an irrelevant externality is equally valid if there is an offsetting external benefit.

Now consider the case of high externality on the cost side and low externality on the benefit side. This is shown by the intersection of the *MSC (high externality)* and *MSB (low externality)*, point *C* in Fig. 1. In this case there is a wide divergence between the privately optimal level of activity (point *A*) and the socially optimal level of activity (point *C*). This case is one in which it appears desirable for the law to intervene to reduce the level of activity. Indeed, in the case of very high externality on the cost side (see *MSC (very high externality)*), it may be desirable to shut down the activity completely.

Consider lastly the case of low externality on the cost side and high externality on the benefit

side. The intersection of the marginal social cost and marginal social benefit schedules occurs at point *D* in Fig. 1. In this case, the privately optimal level of activity (*A*) is substantially below the socially optimal level (*D*). The law should intervene to increase the actor’s level of activity.

Law

I have considered external costs and external benefits associated with activities conducted with reasonable care. Since the actors are assumed to be exercising reasonable care, the negligence rule cannot influence their activity level choices. The negligence rule holds the actor liable only when he fails to take reasonable care.

Strict liability has the property that it imposes liability on actors even when they have taken reasonable care. The legal system can regulate activity levels through imposing strict liability. This part examines the conditions under which strict liability leads to optimal activity levels.

First, consider the case in which externality is high on the cost side and low on the benefit side. The socially optimal scale is point *C* in Fig. 1. In the absence of strict liability, the privately optimal scale is point *A*. Imposing strict liability on the actor is desirable in this case. When strict liability is imposed on the actor, his marginal private cost schedule becomes equivalent to the marginal social cost schedule.

In the case of high externality on the cost side coupled with low externality on the benefit side, the actor’s privately optimal activity level under strict liability will be point *E*. It is not exactly the optimal level, which is at point *C*, but it is close. Social welfare will most likely be improved by using liability to lead the actor to produce at scale *E* rather than at the socially excessive scale *A*.

Consider the case in which externality is low both on the cost and on the benefit side. The socially optimal scale of activity is associated with point *B*. The privately optimal level of activity is associated with point *A*. These are the same activity levels. If strict liability is imposed on the actor, it will reduce his activity level below the socially optimal scale and therefore reduce social welfare. Strict liability will cause the actor to

choose the scale F , which is below the socially optimal scale.

This analysis implies that *strict liability is desirable only when the external cost of the actor's activity substantially exceeds the external benefit associated with the actor's activity*. In this case imposing strict liability reduces activity levels to a point that is closer to the socially optimal scale than would be observed under the negligence rule. When the external benefits are roughly equal to or greater than the social costs associated with the actor's activity, strict liability is not socially desirable.

Another case in which strict liability is not socially desirable is observed when two actors cross-externalize equivalent costs. Put another way, *when the costs externalized by two actors to each other are reciprocal, strict liability is not socially preferable to negligence*. The reason is that under strict liability, you will pay for harms to others, while under negligence (when everyone is complying with the negligence standard), you will pay only for the harms you suffer. Since those harms are the same, activity levels will not differ under the two regimes (Hylton 2008, 2011).

Application to Law: Nuisance and Abnormally Dangerous Activities

To this point, I have presented a model of the economics of externalities and considered its implications for law. In this part, I will examine the extent to which the law conforms to the predictions of the model.

Abnormally Dangerous Activities

The most straightforward application of this model is to the law of abnormally dangerous activities (e.g., blasting). To determine whether an activity is abnormally dangerous, Section 520 of the Restatement (Second) of Torts provides the following factors:

- (a) Existence of a high degree of risk of some harm to the person, land, or chattels of others
- (b) Likelihood that the harm that results from it will be great
- (c) Inability to eliminate the risk by the exercise of reasonable care

- (d) Extent to which the activity is not a matter of common usage
- (e) Inappropriateness of the activity to the place where it is carried on
- (f) Extent to which its value to the community is outweighed by its dangerous attributes

The provisions of Section 520 are in line with the theory set out in the previous part of this entry. First, note that Section 520 can be divided into two parts, the first three provisions and the last three provisions. The first three provisions govern the degree of residual risk and imply that strict liability for operating an abnormally dangerous activity is appropriate only when the residual risk – the risk that remains after the actor takes reasonable care – is high. If the residual risk of the actor's activity is high, strict liability may be appropriate. On the other hand, if the residual risk is relatively low, strict liability would be inappropriate under Section 520. Judge Richard Posner famously applied this component of Section 520 to hold that strict liability would be inappropriate in *Indiana Harbor Belt Railroad Co. v. American Cyanamid Co.*, 916 F.2d 1174 (7th Cir. 1990).

The final three provisions of Section 520 line up with the language in *Rylands v. Fletcher*, L.R. 3 H.L. 330 (1868), which provides the foundation for the law on abnormally dangerous activities. The third factor, common usage, helps us identify activities for which the risks are reciprocal to those of other common activities. If an activity is one of common usage, then actors engaged in those activities will impose reciprocal risks on each other, and there is therefore no basis for adopting strict liability over negligence. The fourth factor, inappropriateness, is another way of determining whether the activity imposes a reciprocated risk. The last provision, comparing benefits and risks, guides courts to compare the external benefits thrown off by the activity with the external costs. If the external costs are great relative to the external benefits, strict liability is appropriate under this provision.

Consider an example. If the actor holds a lion as a pet in his backyard, he will inevitably impose a great risk on his neighbors. Moreover, it is a risk

that remains great even after the actor has taken reasonable care. For this reason, holding a lion as a pet satisfies the first three elements of the Section 520 strict liability test. The last three elements are also satisfied. Holding a lion as a pet is not a common activity – the risk the lion holder externalizes to his neighbors is not equivalent to the risk they externalize to him. The benefits externalized to neighbors from holding a lion as a pet are likely to be far less than the risks externalized to them. For these reasons, it is appropriate under the theory presented here and under Section 520 to apply strict liability to the activity of holding lions as pets.

Nuisance

The law on abnormally dangerous activities is the most obvious application of the theory of this entry. However, the theory applies equally well to nuisance law, the subject of this entry. Most of the standard environmental interferences, such as air or water pollution, have been treated as nuisances under tort law.

The theory of this entry suggests a clear interpretation for the rules governing nuisance law. First, consider the basic legal definition of a nuisance: an intentional, nontrespassory, and unreasonable invasion into the quiet use and enjoyment of property. Intentional, in nuisance law, has had a meaning very similar to its use in the context of trespass law: it is enough if the defendant was aware of the nuisance. There is no need, on the part of the plaintiff, to prove that the defendant aimed to harm the plaintiff. The term nontrespassory has always had the effect of distinguishing between invasions that interfere with exclusive possession of property or a portion of it (e.g., a boulder hurled onto the plaintiff's property) and invasions that merely make it less desirable to remain in possession of property (e.g., smoke).

Perhaps the most important term in the definition of nuisance is unreasonable. The theory of this entry suggests that the factors of Section 520 are equally applicable to nuisance disputes. Paraphrasing Section 520, the appropriate test for unreasonableness under nuisance law can be articulated as follows:

- (a) Existence of a high degree of interference with the quiet use and enjoyment of land of others
- (b) Likelihood that the harm resulting from that interference will be substantial to the typical member of the community
- (c) Inability to eliminate the interference by the exercise of reasonable care
- (d) Extent to which the activity is not a matter of common usage
- (e) Inappropriateness of the activity to the place where it is carried on
- (f) Extent to which its value to the community is outweighed by its dangerous attributes

The first three factors of this test require that the interference be substantial even when the actor is taking reasonable care. As in the case of abnormally dangerous activities, the first three factors should be treated as minimal requirements for nuisance liability.

The last factor asks the court to compare the benefits externalized by the activity and the costs externalized. When the benefits are substantial, the last factor suggests that the court should be reluctant to impose liability on a nuisance theory. Consider, for example, the noise generated by a busy fire station. The noise generated by fire trucks constantly moving in and out of the station with their alarms running could be deemed to substantially interfere with the quiet use and enjoyment of land by neighbors. However, the neighbors also benefit by being located close to the fire station. Since those benefits are substantial and widely dispersed, the neighbors should not be allowed to impose strict liability on a nuisance theory against the fire station, the conclusion reached in *Malhame v. Borough of Demarest*, 392 A. 2d 652 (Law Div. 1978). There is no economic basis for using liability as an incentive to force the fire station to reconsider its location decision.

Nuisance law does not provide for compensation to the extra-sensitive plaintiff (*Rogers v. Elliott*, 15 N.E. 768 (Mass. 1888)). The justification for this well-settled piece of the law is best understood in terms of the model of this entry. A nuisance exists, under the model here, when the externalized costs associated with an activity are

substantially in excess of externalized benefits. The comparison of externalized costs and benefits is made with respect to statistical averages, not to any particular plaintiff. If, on the basis of statistical averages, the externalized costs associated with an activity are not substantially greater than the externalized benefits, then the activity is not a nuisance under the theory here. If a particular plaintiff suffers a severe injury under these conditions, that harm may be actionable under some other legal theory, such as negligence, but it is not actionable under nuisance law.

Local conditions play an important role in nuisance law. In particular, the last three factors (d, e, and f) of the test proposed here depend on local conditions. Most environmental pollutants are regulated because of the risk of harm they impose on people located near the source. In most cases, the risk of harm declines as people move further from the source. Thus, externalized costs are likely to be substantial near the source and declining to zero as one moves further away. Strict liability provides incentives for the pollution generator to locate in regions in which externalized costs are insignificant.

Under the proximate cause rule, courts have limited the scope of nuisance liability to injuries that are connected in a predictable way to the externalized risk. Injuries that are not predictably related to the externalized risk are not within the scope of strict nuisance liability. The externality model suggests a reason for this: to focus liability on the cost-externalizing features of the defendant's activity rather than the activity per se. Suppose the victim drives his car into the defendant's malarial pond. To permit a strict liability action would fail to tax the defendant's activity for the specific risk creation – i.e., the risk of malaria – that nuisance law aims to discourage.

A clearer justification for the proximate cause rule in nuisance law can be based on the model of the previous section. Let the externalized risk component be separated into two subcomponents, where one is the normal risk externalized by activities of the defendant's type and the other is the extraordinary risk that makes the defendant's activity a nuisance. For example, in the case of a

malarial pond, the normal part is the risk externalized by any water storage, and the extraordinary part is the malaria risk. The proximate cause rule excludes liability for the normal-risk component. If, as nuisance law implicitly assumes, normal risks are balanced off by (normal) positive externalities, then excluding liability for normal risk leads to optimal activity levels.

The proximate cause rule leads to the social optimum in activity by excluding the normal-risk component as a source of liability. In terms of Fig. 1, the "low externality" cost increment (*MSC (low externality)*) is representative of the normal risk. If normal positive externalities are present, so that *MSB (low externality)* measures the marginal social benefit of the activity, the socially optimal activity level (assuming the risk consists of both the normal and extraordinary components) is that associated with point C. However, strict liability applied without any offset based on the proximate cause rule would lead the actor to choose the activity level associated with point E. Applying the proximate cause rule of nuisance law, which limits application of strict liability to those injuries attributable to the extraordinary risk, leads the actor to choose the socially optimal activity level (point C). Thus, the proximate cause rule improves on strict liability by leading to an optimal imposition of liability.

Coming to the Nuisance

Sometimes defendants argue that plaintiffs should not be able to recover because they "came to the nuisance." The coming-to-the-nuisance defense is valid in some cases but not in all. The theory of this entry provides a justification for the ambiguous treatment of the coming-to-the-nuisance defense.

Since the goal of nuisance liability is to optimally regulate activity levels, a victim's decision to come to the nuisance is certainly a relevant piece of information. The victim's decision to move is no different from the case of the buyer who contracts with a seller to purchase some item with a latent and dangerous defect. If the buyers are aware of the negative feature of the product, then the resulting market equilibrium would be

socially optimal. Similarly, if a smoke-belching factory sits alone in an area, and the victim moves next door to it, there would be no reason to view the factory's activity as socially excessive. In this case, the coming-to-the-nuisance defense applies.

There are two reasons that the coming-to-the-nuisance defense might not be desirable in this model. First, the victim may not have been aware of the offender's activity when purchasing his property. In *Ensign v. Walls*, 34 N.W.2d 549 (Mich. 1948), the defendant maintained dog-breeding business in the residential area of Detroit. The invasions (odors, noise, occasional escapes, filth) caused by the defendant's activity may not have been obvious to prospective residents; most probably became aware of the nuisance only after moving in.

The second reason the coming-to-the-nuisance defense may not be socially desirable is that the market for real property can be distinguished from most other markets for goods or services. Suppose the community consists of one smoke-belching factory and 99 residents. It is clear in this case that the reciprocal harm condition would not be satisfied; the background risks externalized by the residents would be trivial in comparison to the cost externalized by the factory. If the coming-to-the-nuisance defense were allowed, there would be no mechanism to control the activity level of the factory. The factory could double its level of activity without meeting any liability.

The justifications for the law on priority based on the externality model do not diminish the more traditional transaction-cost-based understanding. A rule favoring priority would encourage socially wasteful races and expropriation (Wittman 1980; Snyder and Pitchford 2003). Snyder and Pitchford (2003) distinguish their analysis from the seminal analysis of Wittman (1980) by noting that their model allows for the court to have limited information and for low transaction costs between the parties after the first move invests. Smith (2004) addresses nuisance law generally from the perspective of information costs, arguing that exclusion rules are favored by the law because they facilitate the production and disclosure of information.

Transaction-Cost Model Versus Externality Model

A complete economic model of nuisance law would consist of the transaction-cost model and the externality model, with the transaction-cost model used to explain the boundaries of nuisance law and the externality model used to explain its regulatory function. The foregoing analysis focuses less on the boundary question that has been the focus of transaction-cost analysis and more on the regulatory function of nuisance.

I have already noted some of the boundary questions examined under the transaction-cost model, specifically the choice between trespass and nuisance and the rule on priority. The transaction-cost model appears to be superior to the externality model as a theory of the boundary between nuisance and trespass law. However, both the transaction-cost and externality models provide justifications for the law's treatment of priority.

One other boundary question, unexamined so far, is the exclusion of protection under nuisance law for aesthetic interests, such as the right to sunlight or to a view of the mountains, *Fontainebleau Hotel Corp. v. Forty-Five Twenty-Five, Inc.*, 114 So. 2d 357 (Fla. App. 1959). The exclusion of aesthetic interests appears to be better explained by the transaction-cost model than by the externality model. It is obviously an externality, in the technical sense, when a landowner erects a fence that blocks the sunlight to another adjacent landowner. There is no reason suggested by the externality model for not treating the harm to the adjacent landowner as potentially a nuisance.

Under the transaction-cost model, there is a clear economic case for excluding liability for aesthetic harms. If aesthetic interests were protected by nuisance law, there would immediately be questions of information and proof. If one adjacent landowner can sue the owner of a hotel for blocking sunlight, why not allow other adjacent landowners? The transaction costs of resolving these disputes in the bargaining process would be enormous. On the other hand, if the law refuses to protect aesthetic interests, then the transaction

costs of resolving disputes would be much more manageable.

Conclusion

Nuisance law is complicated and covers a wide array of land use disputes. However, at its core, nuisance is simple. The law generates optimal activity levels by imposing strict liability when externalized risks are far in excess of externalized benefits or far in excess of background risks. Nuisance doctrine is consistent with this theory.

References

- Buchanan JM, William C (1962) Stubblebine, externality. *Economica* 21:371
- Calabresi G, Douglas Melamed A (1972) Property rules, liability rules, and inalienability: one view of the cathedral. *Harv L Rev* 85:1089
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1
- Haddock DD (2007) Haddock, irrelevant externality angst. *J Interdiscip Econ* 19:3–18
- Hylton KN (1996) A missing markets theory of tort law. *Nw U L Rev* 90:977–1008
- Hylton KN (2008) A positive theory of strict liability. *Rev Law Econ* 4:153–181
- Hylton KN (2010) The economics of public nuisance and the new enforcement actions. *Supreme Court Econ Rev* 43(1):43–76
- Hylton KN (2011) The economics of nuisance law. In: Ayotte K, Smith HE (ed) *Research handbook on the economics of property law*, Edward Elgar Publishing, pp 323–343
- Merrill TW (1985) Trespass, nuisance, and the costs of determining property rights. *J Leg Stud* 14(1):13
- Polinsky AM (1979) Controlling externalities and protecting entitlements: property right, liability rule, and tax-subsidy approaches. *J Leg Stud* 8(1):1–48
- Prosser WL (1971) *Handbook of the law of torts*, 4th ed. p 571
- Restatement (Second) of Torts: unreasonableness of intentional invasion § 826 (1977)
- Shavell S (1980) Strict liability versus negligence. *J Leg Stud* 9(1):1–25
- Smith HE (2004) Exclusion and Property Rules in the Law of Nuisance. *Va L Rev* 90:965
- Snyder CM, Pitchford R (2003) Coming to the Nuisance: An Economic Analysis from an Incomplete Contracts Perspective. *J Law Econ Org* 19(2):491–516
- Wittman DA (1980) First Come, First Served: An Economic Analysis of “Coming to the Nuisance”. *J Leg Stud* 9:557–568

Nuisance Lawsuits

► Frivolous Suits

Off-Premises Contract

- Distance Selling and Doorstep Contracts

Operational Risk

- Risk Management, Optimal

Optimality

- Rationality

Optimization Problems

Roy Cerqueti and Raffaella Coppier
Department of Economics and Law, University of
Macerata, Macerata, Italy

Abstract

This issue deals with the conceptualization of an optimization problem. In particular, we first provide a formal definition of such a mathematical concept. Then, we give some classifications of the optimization problems on the basis of their main characteristics (presence of time dependence and of constraints). In so

doing, we also outline the standard techniques adopted for seeking solutions of an optimization problem. Lastly, some examples taken by the classical theory of economics and finance are proposed.

Definition

An optimization problem is a mathematical obstacle. Its overcoming leads to the identification of some quantities (optimal solutions), contained in a predefined set (admissible region), such that a given (objective) function is optimized (maximized or minimized). Optimization problems have remarkable relevance in the context of economics and law, since taking consistent decisions is associated to the implementation of optimal – in some sense – strategies.

The General Theory of the Optimization Problems

In the context of economics and law, decision makers should act under the guide of optimality criteria. Indeed, it is easy to understand that a decision might drastically undermine the outcome of economic policies or law statement. This said, it is important to develop optimization models and solve them.

In this issue *optimization problems* (OP, hereafter) are treated. Specifically, some classifications of

OP are proposed on the basis of their relevant features. Several clusterings of the huge set of the OP can be proposed. We here report the most important ones under the perspective of the applications. Moreover, the solution strategies for each class of OP are also discussed, having in mind that an optimization problem might not have solutions. In this respect, a warning is in order: OP can be very difficult to solve, and sometimes standard techniques are not sufficient for achieving a solution. The core of Operational Research is exactly the study of new methods for facing OP which are out of the main frameworks. Hence, nonstandard techniques are not treated in this report, even if some of them will be briefly mentioned. Some remarkable economic examples of OP are also proposed.

Classification of the OP

Perhaps, the most intuitive classification of the OP is the one grounded on the dependence of the problems on time and leads to the distinction between static and dynamic OP. More precisely, OP are *dynamic* when the objective function and/or the elements of the admissible region are time dependent, otherwise they are *static*.

The static case is the simplest one, and its treatment requires basic mathematical tools. To discuss static OP, a subclustering of them concerning the presence or the absence of constraints on the involved variables is needed.

Static OP are *unconstrained* if the objective to be optimized is a real-valued function f defined over the n -dimensional Cartesian space $\mathbf{R}^n$, and the optimal variables are searched on the entire set $\mathbf{R}^n$.

Now, assume that all the functions here presented are enough regular, and the derivatives used in the description of the static OP exist. Then the standard solution algorithm runs in two steps: first, the “candidate” optimal solutions are detected by posing the vector of the first partial derivatives of f (the *gradient of f*) identically null, i.e.,: by searching for the vectors $\mathbf{x}_0 \in \mathbf{R}^n$ such that *grad* ($f(\mathbf{x}_0)$) = 0 (*first order conditions*, and the $\mathbf{x}_0$ ’s are the *stationary points*); second, the candidates obtained in the first step are classified through the analysis of the sign of the *Hessian* $Hess(f(\mathbf{x}_0))$,

which is an $n \times n$ matrix containing the second partial derivatives of f and provides information on the convexity of the function f . Specifically, if $Hess(f(\mathbf{x}_0))$ has positive (negative) sign, then f is convex (concave) in a neighborhood of $\mathbf{x}_0$, which is a minimum (maximum) for f over $\mathbf{R}^n$ (*second order conditions*). When $Hess(f(\mathbf{x}_0))$ is indefinite, then second order conditions do not lead to the classification of the stationary point $\mathbf{x}_0$. In this case, $\mathbf{x}_0$ can be classified through a local study of the behavior of f in a neighborhood of it.

Static OP are *constrained* when the optimal variables belong to a proper subset $\mathbf{A}$ of $\mathbf{R}^n$, which is identified through a collection of analytical equations and disequations of the variable $\mathbf{x} \in \mathbf{R}^n$. This class of problems is treated similarly to the unconstrained case but with a remarkable difference: the introduction of the so-called *Lagrangian H* is needed. The Lagrangian is a function including in a unified term the objective f and the equations and disequations generating $\mathbf{A}$. The optimal solutions are then derived by setting the gradient of H equals to zero and then by classifying the stationary points through the study of the Hessian of H . It is worth noting that H is defined also by the introduction of some ancillary variables (the *Lagrange multipliers*), so that the stationary points of H are automatically included in $\mathbf{A}$. It is important to recall here Weierstrass’ Theorem, which can be in general formulated as follows: if $\mathbf{A}$ and f satisfy some properties, i.e.,: $\mathbf{A}$ is a compact (closed and bounded) set of $\mathbf{R}^n$ and f is continuous in $\mathbf{A}$, then the constrained optimization problem admits solutions.

For details and examples on static OP, refer to Simon and Blume (1994).

The dynamic OP are those more involved in economic application, in that they are able to fit with the evolutive nature of the reality. In the dynamic OP, a set $\mathbf{T} \subseteq [0, +\infty)$ collecting the time is introduced. The objective functional J depends on time $t \in \mathbf{T}$ and is optimized with respect to an n -dimensional function of the time t (the *state variable*), denoted as $\{\mathbf{x}_t\}$. Indeed, functional J is usually defined as the aggregation over t (sum over t if $\mathbf{T}$ is a discrete set; integral over t if $\mathbf{T}$ is a continuous set) of a time-dependent

function f which depends also on $\{x_t\}$. The way in which optimization is performed can be *indirect*, i.e., $\{x_t\}$ and J are optimally selected by controlling a further set of variables $\{\alpha_t\}$ included in their definition (the *control variables*) or *direct*, otherwise. In the former case, we define the *optimal control problems* (OCP, henceforth), which is a very relevant class of the dynamic OP.

To fix ideas, assume that time is continuous, and $\mathbf{T} = [t_1, t_2] \subseteq [0, +\infty)$.

Under an operative point of view, the strategy for solving OP in the “direct case” is similar to that presented for the static OP. The role played in the static situation by the stationary points is assigned in the dynamic setting to the so-called *extremals*. In details, under the necessary regularity condition, the first order conditions can be rewritten through a special differential equation (*Euler equation*) of f . The extremals are then classified by studying the convexity of f through the Hessian matrix of the second derivatives of f . This procedure is theoretically formalized in the *Calculus of Variations*, and an illuminating overview on it can be found in Kamien and Schwartz (1991).

For what concerns OCP, their main ingredients are: the state variable $\{x_t\}$, which evolves accordingly to a differential equation (state equation); the control variable $\{\alpha_t\}$, managed by the decider in order to solve the optimization problem; the admissible region, which is a functional set containing the control variables; the objective functional J ; the *value function* V , which is the optimized objective function; and, lastly, the *optimal strategies*, which are given by the pair optimal controls-optimal paths.

A further clustering of OCP can be properly identified, on the basis of the presence of randomness in their formalization. So, we have *deterministic* OCP and *stochastic* ones. The latter are those where f , $\{x_t\}$ and $\{\alpha_t\}$ are random, while all the ingredients of the former are deterministic. In stochastic OCP, the state variable obeys a stochastic differential equation.

The procedures for solving deterministic and stochastic OCP are basically two: the *Pontryagin maximum principle* and the *dynamic programming theory*. The former method may be viewed as a generalization of the procedure employed for

solving the dynamic OP. Indeed, Pontryagin’s Theorem states that the optimal strategies are couples optimal controls-optimal paths which must optimize a special function (the *Hamiltonian*) similar to the Lagrangian of the static OP case and including the objective functional and the constraints. Moreover, the introduction of additional variables (*costates*) and the fulfillment of further conditions (*costate equations*) are also required; the latter method is based on the definition of the value function, which is proved to be formally the unique solution of a specific differential equation (*Hamilton-Jacobi-Bellman equation*, *HJB*) by means of a maximum principle (the *Dynamic Programming Principle*). If the value function is so regular to be the true solution of the HJB, then one can derive the optimal strategies through a *Verification Theorem*. It is worth noting that the regularity of the value function is an important aspect to be treated in the dynamic programming approach, and the employment of a concept of weak solutions of the HJB (the so-called *viscosity solutions*) is often required.

For a survey on the OCT and, in general, on the dynamic OP, see Fleming and Rishel (1975), Bardi and Capuzzo Dolcetta (1997), Yong and Zhou (1999), Fleming and Soner (2006), Kamien and Schwartz (1991). The concept of viscosity solutions is introduced and discussed in Fleming and Soner (2006) and Crandall et al. (1992).

Examples of OP

Some classical examples of OP can be listed in the field of economics.

In Markowitz (1952), the future Nobel Laureate Harry Markowitz developed an optimization model for the selection of the best *capital allocation* among a set of risky assets. The optimality criterion is based on the evidence that an investor aims at maximizing the expected return of a portfolio and, simultaneously, at minimizing its risk level. The resulting problem is a static constrained one, with convex objective function.

In the field of decision theory, a relevant role is played by the concept of utility. A utility function is a tool for ordering preferences of goods and is a

simple assignment of a number representing the satisfaction to be the owner of some quantities of goods. By conventional agreement, a greater value of the utility means a higher level of satisfaction. Hence, *utility maximization* is the ground of a number of static and dynamic OP. For some explicit models, refer to Mas-Colell et al. (1995).

A macroeconomic example of optimization model consists of the *maximization of the growth rate of a country*. In this case, the formalization of the related OP includes the evolutive nature of the phenomenon and the wide set of constraints. Hence, the OP are dynamic. The objective function is usually consumption or utility, and the constraints model human or physical capital accumulation. For this type of models, see Barro and Sala-i-Martin (2004).

Under a microeconomic point of view, of remarkable relevance is the problem of *cost minimization* and *profit maximization*. The solution of the resulting OP provides insights on the strategies to be implemented by the companies for improving their performances. Usually, such models are described through constrained OP, in order to capture the presence of budget and/or technological constraints. A detailed discussion on this family of OP can be found in Varian (1992).

References

- Bardi M, Capuzzo-Dolcetta I (1997) Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations, Systems & control: foundations & applications. Birkhäuser, Boston
- Barro R, Sala-i-Martin X (2004) Economic growth, 2nd edn. The MIT Press
- Crandall MG, Ishii H, Lions P-L (1992) User's guide to viscosity solutions of second order partial differential equations. Bull Am Math Soc 27(1):1–67
- Fleming WH, Soner HM (2006) Controlled Markov processes and viscosity solutions, 2nd edn. Springer, New York/Heidelberg/Berlin
- Fleming WH, Rishel RW (1975) Deterministic and stochastic optimal control. Springer, New York/Heidelberg/Berlin
- Kamien MI, Schwartz NL (1991) Dynamic optimization: the Calculus of variations and optimal control in economics and management, vol 31, 2nd edn, Advanced textbooks in economics. Elsevier B.V, Amsterdam
- Markowitz H (1952) Portfolio selection. J Financ 7(1):77–91
- Mas-Colell A, Whinston M, Green J (1995) Microeconomic theory. Oxford University Press, Oxford
- Simon CP, Blume L (1994) Mathematics for economists. W.W. Norton & Company
- Varian H (1992) Microeconomic analysis, 3rd edn. W.W. Norton & Company
- Yong J, Zhou XY (1999) Stochastic controls. Springer, New York/Heidelberg/Berlin

Option

Danny Cassimon¹ and Peter-Jan Engelen^{2,3}

¹Institute of Development Policy and Management (IOB), University of Antwerp, Antwerp, Belgium

²Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

³Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

Definition

The right (but not the obligation) to buy or sell a certain asset at specific moments in time at a predetermined price

Introduction

An option is a financial instrument that gives its holder the right to buy or sell an underlying asset at a pre-agreed price at specific moments in time (Hull 2011). The first feature of an option contract is that it gives the holder the right to do something, but not the obligation (Bachelier 1900). If an option entitles to buy an asset, the option is referred to as a call option, while a put option is the right to sell an asset (Brennan and Schwartz 1977). The pre-agreed price at which the holder can buy or sell the asset is the strike price or the exercise price. The time to maturity is the time period which indicates when the holder can exercise the option. When an option is exercised during the entire period, it is commonly referred to as

an American option, while a European option contract can be exercised only at the pre-determined expiration date (Cassimon et al. 2007). While the holder determines whether he or she is going to execute or exercise the option, the counterparty (the writer of the option contract) needs to deliver the asset to the holder at the pre-agreed price (in case of a call option) or buy the asset from the holder at the pre-agreed price (in case of a put option). The writer's action is thus dependent on the decision of the holder. For that flexibility, the holder of the option pays an upfront, non-refundable fee to the writer, which is called the premium, reflecting the market price of that option (Cassimon and Engelen 2016; Black and Scholes 1973; Cox et al. 1979; Merton 1973). One can find option contracts on a wide range of assets, including options on common stock (e.g., on a share of Google), options on a stock index (e.g., on the S&P-500, or the Euronext-100 index), options on currencies (e.g., on the EUR/USD exchange rate), options on commodities (e.g., on corn, cattle, soybeans), options on bonds, options on the weather conditions, etc. Options on common stock can also be part of the salary package of senior firm staff. Besides those financial options on a wide range of assets, the concept of options has also been applied in a business context (Cassimon and Engelen 2015; Cassimon et al. 2004).

Options are available on standardized option exchanges or tailor-made from financial institutions in the over-the-counter market. The most important option exchanges include the Chicago Board Options Exchange (CBOE), the CME Group (a merger between the Chicago Mercantile Exchange and the Chicago Board of Trade), the Philadelphia Stock Exchange (now part of NASDAQ), Eurex, NYSE Derivatives (Intercontinental Exchange), the National Stock Exchange of India (NSE), the Bombay Stock Exchange (BSE), and Bovespa. The most traded index options in 2016 included contracts on the Nifty (India), KOSPI 200 (Korea), BankNifty (India), EuroStoxx50 (Eurex), and the S&P500 (CBOE), while most stock options were traded at Bovespa (Brazil), NASDAQ, and NYSE (top three with 47% of the market in 2016). The most

actively traded currency option contract in 2016 was on the USD/INR exchange rate.

Standardization is an important feature of option contracts on exchanges to secure liquidity. This implies that option contracts are only available for specific amounts of the underlying asset, for specific strike prices, for specific maturities, and for specific exercise rights (American versus European). For instance, the CBOE uses the following standardized features for its option contracts on stocks: the underlying value amounts to 100 shares; the strike price intervals are typically set 2.5 points when the strike price is between 5 and 25 USD, 5 points when the strike price is between 25 and 200 USD, and 10 points when the strike price is over 200 USD; and the expiration date is the third Friday of the expiration month, with contracts available for two near-term months plus two additional months from the January, February, or March quarterly cycles. Finally, the stock options are American option contracts. Table 1 gives an example of option price series for different strike prices and time to maturity of 3 months for an underlying stock with a current price of 22.57 USD.

In the following we show that options can be used both for (speculative) investment transactions and to provide hedging. Options are used as an investment when its holder does not hold the underlying asset and is using the investment in the option to (deliberately) create an open position in which he/she is exposed to risk, i.e., to the price evolutions of the underlying asset, hoping that the price evolves in his/her advantage, leading to a gain; however, prices can also move unfavorably, leading to a realized loss. In this sense, these investments are “speculative.” When, on the contrary, one holds the underlying asset (or will acquire it somewhere in the future, during the

Option, Table 1 Example of option price series (in USD)

Strike prices	Call option premia	Put option premia
17.5	5.30	0.10
20	2.90	0.25
22.5	1.15	1.00
25	0.35	2.60
30	0.05	7.70

lifetime, or at expiration of the option), an option can be used to allow the holder to protect (to hedge) the value of the underlying asset against future adverse price evolutions. In section “Basic Payoff Profiles of Options from an Investment Perspective,” we discuss the basic alternative uses of an option as a speculative investment instrument, showing the different payoff (i.e., the gain/loss) schedules; we also show (in section “Options as a Leveraged Investment”) that such option investments are highly leveraged, i.e., enabling high potential returns (but also high losses) at low investment costs (compared to investment in the underlying asset itself). Section “Using Options for Hedging” then discusses the use of an option from a hedging perspective. In section “Option Combinations and the Put-Call Parity Relationship,” we discuss some more sophisticated option combinations and also point to the theoretical so-called put-call parity relationship. Finally, “Conclusion” section discusses some societal consequences of the introduction and use of options. We will use the example of a stock option and the data of Table 1 throughout the remainder of this contribution.

Basic Payoff Profiles of Options from an Investment Perspective

This section gives an overview of the basic payoff profiles of holding or writing call and put option contracts until expiration date. As such, we assume that the parties do not hold the underlying asset, here the underlying stock, and create a particular position, based on their subjective ideas (or superior information) on the future stock price evolution, hoping to make a profit but at the same time exposing themselves to potential losses in case their subjective ideas or information proves wrong and prices move adversely. We will illustrate the basic payoff profiles using the stock option contract with a strike price of 22.50 USD from Table 1. We provide payoff profiles for the holder of the call option (also referred to as a long call) and the writer of the call option (also referred to as a short call). In a similar vein, we provide

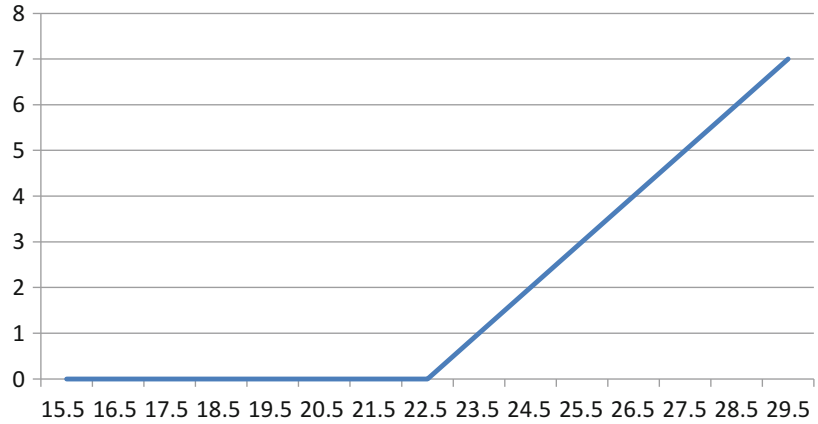
payoff profiles for a long put (holder) and a short put (writer).

Long Call

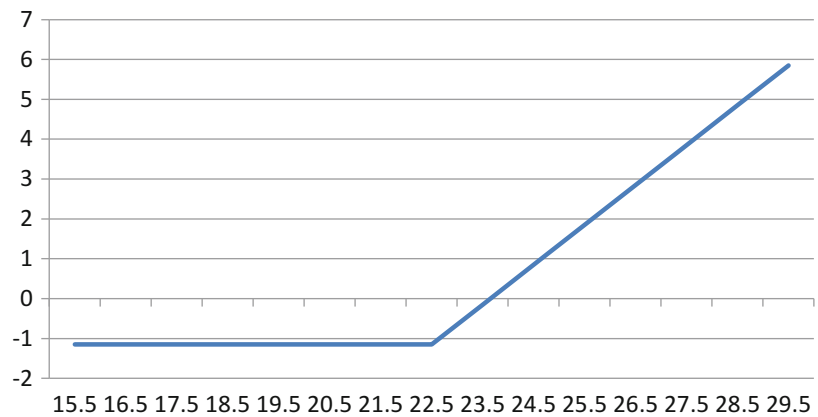
The holder of a call option has the right to *buy* one share of the underlying at the strike price of 22.50 USD. Depending on the stock price at expiration, he will exercise his option or let it expire. For instance, when the stock price at expiration is equal to 20 USD, the holder will let the option expire because one needs to pay 22.50 USD upon exercising the option (right to buy at 22.50 USD), while one can buy the underlying stock directly at the market at a price of 20 USD. In contrast, when the stock price at expiration amounts to 25 USD, the holder will exercise the option and buy the underlying stock from the writer of the contract at a price of 22.50 USD, while its market price is equal to 25 USD. In this way, the holder realizes a profit of 25 USD minus 22.50 USD or 2.50 USD, ignoring any premium paid to acquire the call option.

The example shows that as long as the stock price at expiration is smaller than the strike price or exercise price, the option is worthless and will not be exercised. Put differently, it will take a floor value of zero. When the stock price is lower than the exercise price, the call option is said to be out-of-the-money. As soon as the stock price at expiration is higher than the strike price, the holder will exercise its right and buy the underlying stock at the strike price. The holder will realize a profit, being the difference between the underlying stock price and the strike price. In this case the call option is said to be in-the-money. When the stock price equals the strike price, the call option is said to be at-the-money. This is reflected in Fig. 1 in the asymmetric payoff profile of the call option. Zero profit for stock prices up to the strike price, a positive profit for stock price beyond the strike price. This is the most powerful characteristic of an option. The holder can cut off losses but keep the upside potential open. This powerful mechanism of being able to limit downside losses does not come for free. In order to acquire the option, the holder typically has to pay a premium. In our example, the premium of the call option is equal to 1.15 USD (see Table 1).

Option, Fig. 1 The asymmetric payoff profile of a long call option without the premium



Option, Fig. 2 The asymmetric payoff profile of a long call option including the premium



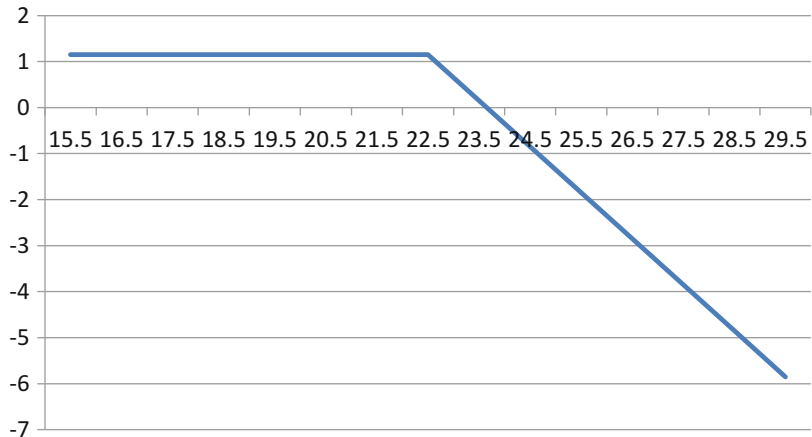
This implies that as long as the call option at expiration ends up out-of-the-money (the stock price is smaller than the strike price), the holder will not exercise and loses the premium of 1.15 USD. As soon as the option ends up in-the-money (stock price higher than the strike price), the holder will exercise and earns the difference between the stock price minus the sum of the exercise price and the premium (see Fig. 2).

Short Call

The payoff profile of a short call contract reflects the position of the writer of the call option contract. The short call is the mirror image of the long call and follows the decisions of the holder of the call option contract. As long as the stock price at expiration is lower than the strike price, the holder of the call option will not exercise the option

contract and let it expire. This implies that the writer of the option does not have to do anything and can just keep the premium for giving the holder the right to buy the share at the pre-agreed strike price. For instance, in the above example, at a stock price of 20 USD, the call option expires, and the writer collects the premium of 1.15 USD. As soon as the stock price at expiration is higher than the strike price, the holder of the option will exercise his or her right to buy the underlying share at the pre-agreed strike price from the writer of the contract. For instance, in the example, at a stock price of 25 USD, the holder will exercise the option and will turn to the writer of the contract to buy the underlying share at the strike price of 22.50 USD, while the market price amounts to 25 USD. The writer is obligated to deliver the share to the holder in return for a cash amount of

Option, Fig. 3 The asymmetric payoff profile of a short call option including the premium



22.50 USD. The writer thus incurs a loss of 22.50 USD – 25 USD or – 2.50 USD. As the writer also received the premium of 1.15 USD, the net loss is 2.50 USD minus 1.15 USD or 1.35 USD. Figure 3 shows the asymmetric payoff profile of the writer of the call option contract. As long as the stock price at expiration is smaller than the strike price, the maximum profit for the writer is the premium; as soon as the stock price at expiration is higher than the strike price, the writer incurs losses. Note, however, that there is no limit to these losses: the higher the realized stock price, the higher the loss; as such, unprotected engagement in the writing of options is a very dangerous strategy. Note also that when one compares the payoff profiles of the holder (Fig. 2) and the writer (Fig. 3) of the call option, it is clear that their payoffs are mirror images of each other.

Long Put

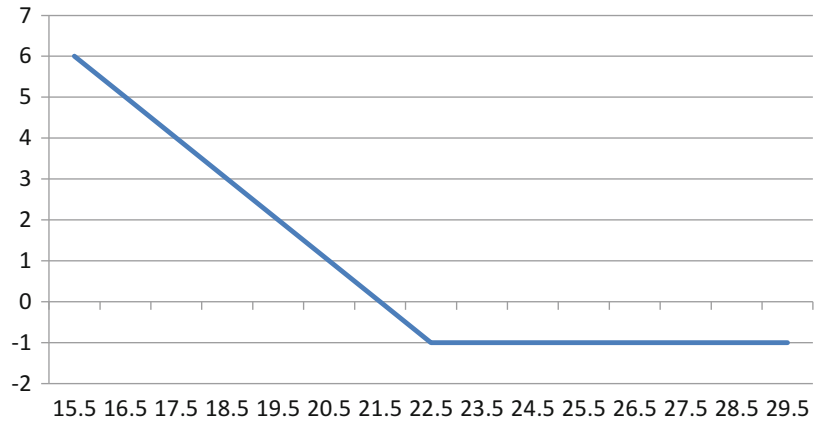
The holder of a put option has the right to *sell* one share of the underlying stock at the strike price of 22.50 USD. Depending on the stock price at expiration, he will exercise his option or let it expire. For instance, when the stock price at expiration is equal to 25 USD, the holder of the put option will let the option expire because one receives 22.50 USD upon exercising the option (right to sell at 22.50 USD), while one can sell the underlying stock directly at the market at a price of 25 USD. The put option is said to be out-of-the-money. In contrast, when the stock price at expiration amounts to 20 USD, the holder will exercise the

put option and sell the underlying stock to the writer of the contract at a price of 22.50 USD, while its market price is equal to 20 USD. The put option is in-the-money. In this way, the holder realizes a profit of 22.50 USD minus 20 USD or 2.50 USD, ignoring any premium paid to acquire the option. If one also takes into account the premium of 1.00 USD (see Table 1), the net profit amounts to 1.50 USD. Figure 4 shows the asymmetric payoff profile of a long put option. As long as the stock price at expiration is higher than the strike price, the put option is out-of-the-money and will not be exercised; hence, the holder just loses the premium. As soon as the stock price at expiration drops below the strike price, the put option ends in-the-money and will be exercised, leading to a positive profit, again largely unlimited (constrained only by the asset becoming completely worthless, in which case the profit here is 21.50 USD).

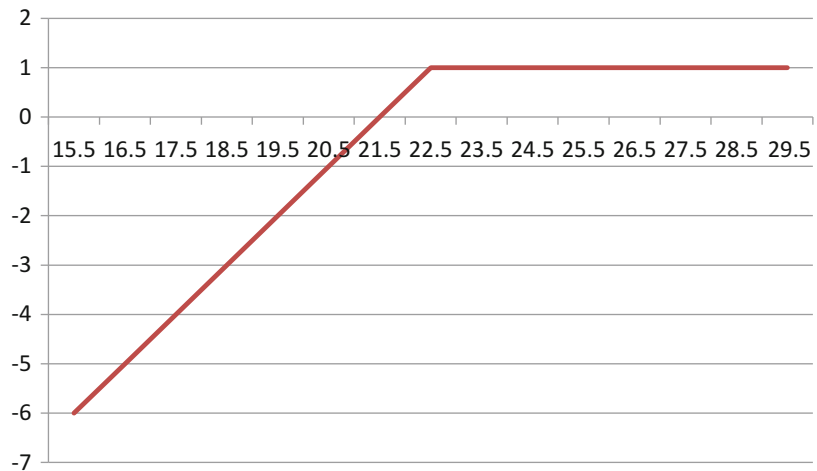
Short Put

Figure 5 shows the payoff profile of a short put or the position of the writer of a put option contract. The short put is the mirror image of the long put and follows the decisions of the holder of the put option contract. As long as the stock price at expiration is higher than the strike price, the holder of the put option will not exercise the option contract and let it expire. This implies that the writer of the option does not have to do anything and can just keep the premium for giving the holder the right to sell the share at the

Option, Fig. 4 The asymmetric payoff profile of a long put option including the premium



Option, Fig. 5 The asymmetric payoff profile of a short put option including the premium



pre-agreed strike price. As soon as the stock price at expiration is lower than the strike price, the holder of the put option will exercise his or her right to sell the underlying share at the pre-agreed strike price from the writer of the contract. The writer will incur losses in this instance, which become higher at lower prices, and again constrained only by the asset becoming completely worthless, in which case the loss here is 21.50 USD. It reiterates the considerable riskiness involved in option writing.

Options as a Leveraged Investment

It is important to understand that options allow for highly leveraged investments. The cost of investing in an option is much lower than that of investing in the underlying asset: investing in say

buying a long call (at 22.50 USD) costs the holder 1.15 USD (the premium), while buying the stock itself requires an investment of 22.57 USD. The evolution of the payoff of both will depend on the stock price evolution, but the payoff of an option is much more sensitive to these price changes than that of the investment in the stock: assume that, at expiration, the realized stock price equals 25 USD. Then the rate of return of investment in buying the stock is about 10.8% (the difference between current and buy stock price, which is 2.43 USD, divided by the investment cost of 22.57 USD), while the rate of return on the option equals 117% (the payoff at 25 USD from Fig. 2, which is 2.50 USD, divided by the investment cost of 1.15 USD). Of course, at adverse price changes, the negative rate of return incurred on a

stock investment is lower than that on the option, which is often equal, but limited, at losing the full investment cost (when the option expires unexecuted).

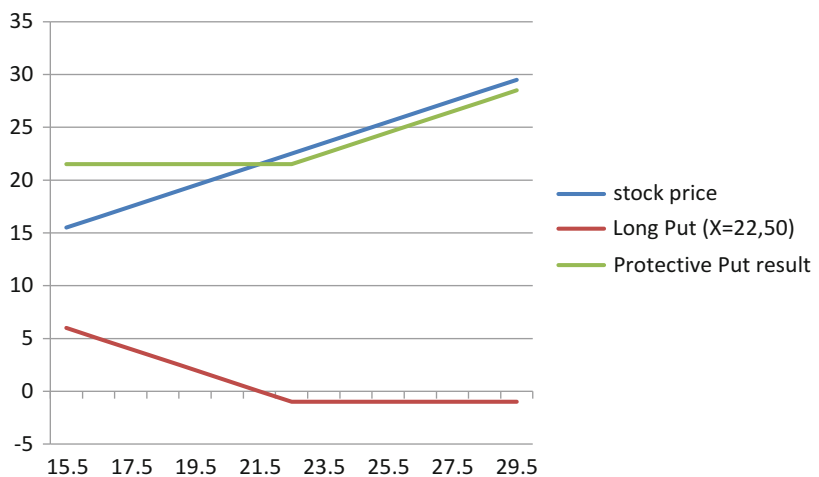
Using Options for Hedging

When one holds the underlying asset, an option can be used to allow the holder to protect (to “hedge”) the value of the underlying asset against future adverse price evolutions. In fact, this was the prime reason why options were designed. More particularly, the holder can buy a long put, which is identical to the example of Fig. 4. The final result is given by Fig. 6: it is a combination of the stock price curve, which shows the result of stock price changes when being unprotected, with the long put payoff curve. The combination of the two curves provides for the final “protective put” result. It shows that when the stock price declines, this is largely compensated by the gains that the holder makes in that case on the long put position, while, as the stock price increases, the option expires unexecuted (with a loss equal to the premium). As a result, the use of this long put allows the holder of a stock to lock in a minimal (“floor”) value of the stock of 21.50 USD (which is the exercise price minus the premium of 1 USD), irrespective of what happens to the stock price, while still being able to enjoy the upward potential

in case the stock price increases. It is clear from comparing the (unprotected) stock price and the protective put curves that hedging does not always lead to the best result (the highest stock value); in case of stock prices higher than 21.50 USD, the unhedged position is better than the hedged one (due to the premium paid to buy the long put option), but hedging does avoid large losses in case of a large stock price decline. The protective put is basically an insurance against a crash of the value of the stock (or a portfolio of stocks). The availability of options with different strike prices will allow for fine-tuning of this position, where locking in a somewhat higher minimal value (by using long put options with a higher strike price) will come at the expense of a lower remaining upward potential.

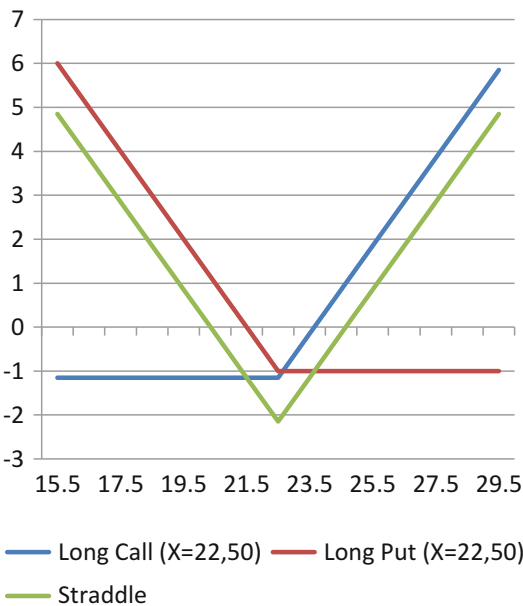
Similarly, when one will acquire the underlying asset somewhere in the future, during the lifetime, or at expiration of the option, an identical strategy of buying a long put can protect against the risk of a drop in the price of that asset in the meantime, leading to the same result as in Fig. 6. If prices turn out lower than the strike price, the holder will exercise the option and sell the share that he/she acquired now at that strike price; as such, a minimal income of 21.50 USD will be secured. At stock prices higher than the strike price, the holder will sell the acquired share in the market at that higher price, and the net income will be that higher price minus the option premium paid.

Option, Fig. 6 The protective put result



Option Combinations and the Put-Call Parity Relationship

Besides the standard payoff profiles of the basic option positions (call/put, long/short), one can create any tailor-made payoff profile by combining several call and put options together, which will accommodate best the particular subjective ideas or information on future price developments, in case of the use of options as an investment vehicle, or the particular desired hedging outcome otherwise. For instance, one could combine a long call and a long put position with the same strike price in one investment portfolio. The combined payoff of holding both option positions simultaneously is referred to as a long straddle position (see Fig. 7). This long straddle will pay off to the holder as long as the stock price at expiration is lower than 20.50 USD or higher than 24.50 USD. If the option ends up at-the-money (at the strike price of 22.50 USD), the holder incurs a maximum loss of 2.15 USD, being the sum of the premium paid for the call option (1.15 USD) and the premium paid for the put option (1.00 USD). This is a valuable trading strategy if one expects the stock price to break out,

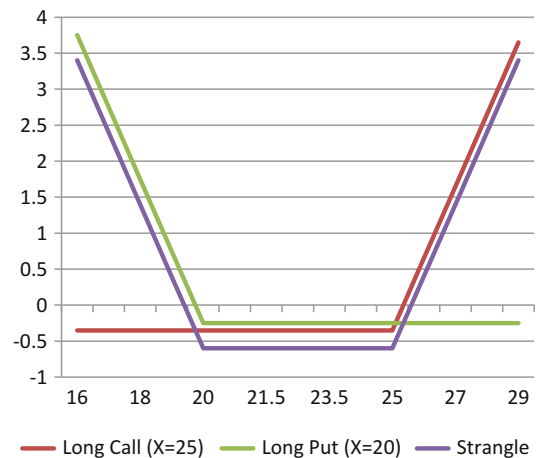


Option, Fig. 7 Payoff profile of a long straddle position

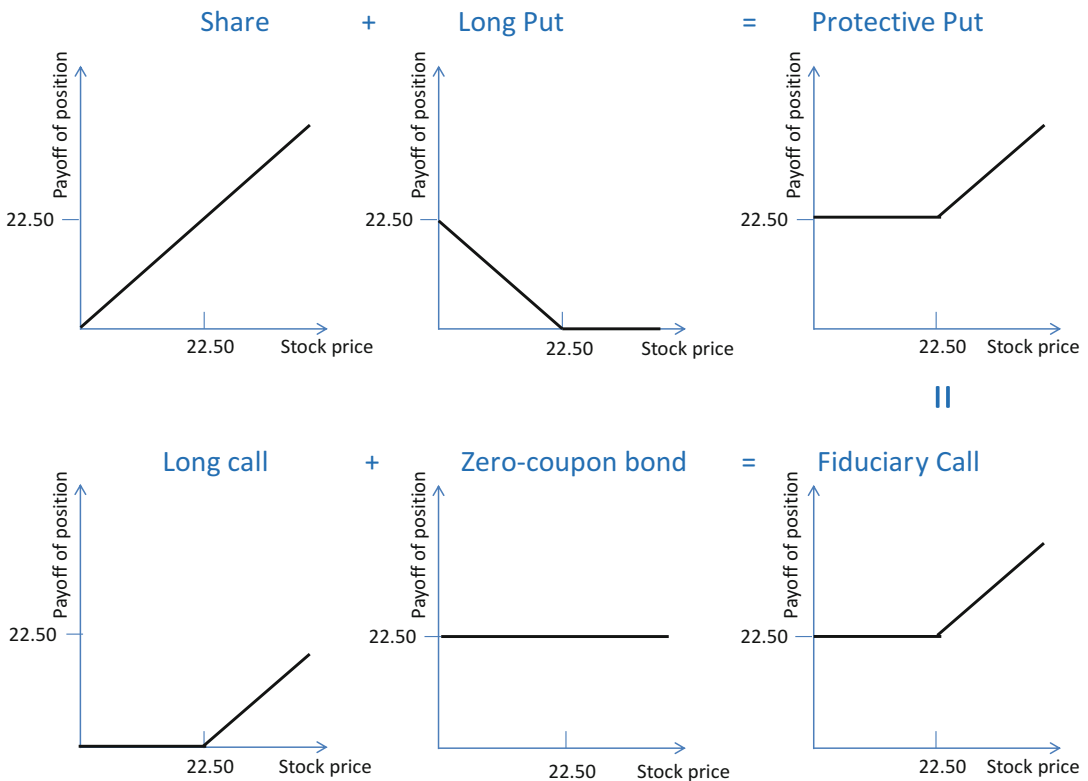
but one is uncertain in which direction. For instance, one expects a pharmaceutical firm to announce news on its R&D program: if the news turns out to be good, one expects the stock price to increase; if the news is unfavorable, one expects the stock price to decline. A straddle position could be a valuable trading strategy in such a situation. In contrast, if one expects the stock price not to move, one could create a short straddle by simultaneously writing a call and put option with the same strike price. If the future materializes as expected, the writer will capture both premia; if the price moves too much in either direction, the writer will incur (unlimited) losses.

An alternative strategy to a long straddle is to combine the long call option and the long put option with different strike prices. For instance, one can combine a long put with a strike price of 20 USD with a long call with a strike price of 25 USD. Such a combination is referred to as a long strangle. In comparison to the long straddle, the maximum loss is smaller (0.60 USD versus 2.15 USD), but the underlying stock price has to move more before the holder will profit (20/26 USD versus 20.50/24.50 USD). By setting the strike prices of the call and put options closer or further away from each other, the holder of the strangle can trade-off reducing the maximum loss with a larger break-even point (Fig. 8).

Finally, Fig. 9 shows a combination of option contracts with other financial assets, which will



Option, Fig. 8 Payoff profile of a long strangle position



Option, Fig. 9 Deriving the put-call parity

also lead to the derivation of the so-called put-call parity. The upper panel again shows the combination of a share with a long put contract, which is the same as the protective put situation of section “Using Options for Hedging” (and Fig. 6), without including the premium. The lower panel shows a fiduciary call position. It involves entering into a long call position and simultaneously investing the present value of the strike price on a risk-free interest account. One can observe that the payoff profile of the protective put and the fiduciary call is exactly the same: the value of a share plus a long put equals the payoff of a long call with the present value of the strike price. Or, put differently, the value of a long put equals the value of a long call plus the present value of the strike price minus the value of the underlying share. This relationship is referred to as the put-call parity and expresses the relationship between the value of a call option and an otherwise identical put option. This implies that once one knows

the value of a call option, one can easily calculate the value of a corresponding put option. The put-call parity is a powerful relationship as an arbitrage constraint on option prices and as a tool to transfer risk between options (Haug and Taleb 2011).

Conclusion

The explanation of their dual use in this contribution has shown that the creation of the option instrument and option markets has brought about real benefits to its users and financial markets at large, first of all through the ability of individual investors, financial and nonfinancial firms, as well as governments to hedge themselves against all kind of underlying financial risks, which provide more financial stability and insurance against catastrophic outcomes, and, secondly, through the creation of additional financial, highly leveraged

investment opportunities, which combine flexibility with low investment input and high potential returns. At the same time, as also shown in this contribution, these benefits come at the expense of potential high costs to the same aforementioned types of users, and even at the world systemic level, due to the potential negative externalities caused when options are used in highly speculative investment strategies that go wrong, in the sense that they result in huge, sometimes unlimited, losses, leading to severe financial distress not only in these entities, but which can then potentially also be transmitted throughout interlinked financial markets, firms, countries, and the financial system at large. Well-known examples include the presence of “rogue traders” such as in the case of Barings Bank or the use of options in speculative currency attacks against countries (see, e.g., Jacque 2015). Preventing such potential catastrophes requires the use of appropriate and rigorous checks and balances at the level of individual firms or governments, self-regulation at the level of financial markets, and other regulatory interventions at the global level (Van Liedekerke and Cassimon 2000).

Cross-References

- [Efficient Market](#)
- [Option Prices Models](#)
- [Real Options](#)

References

- Bachelier L (1900) *Théorie de la spéculation*. Gauthier-Villars, Paris
- Black F, Scholes M (1973) The pricing of options and corporate liabilities. *J Polit Econ* 81:637–659
- Brennan M, Schwartz E (1977) The valuation of American put options. *J Financ* 32:449–462
- Cassimon D, Engelen PJ (2015) Real options. In: Marciano A (ed) *Encyclopedia of law and economics*. Springer, New York, pp 1–10
- Cassimon D, Engelen PJ (2016) Option pricing models. In: Marciano A (ed) *Encyclopedia of law and economics*. Springer, New York, pp 1–6
- Cassimon D, Engelen PJ, Thomassen L, Van Wouwe M (2004) Valuing new drug applications using n-fold compound options. *Res Policy* 33:41–51
- Cassimon D, Engelen PJ, Thomassen L, Van Wouwe M (2007) Closed-form valuation of American call-options with multiple dividends using n-fold compound option models. *Finance Res Lett* 4:33–48
- Cox JC, Ross SA, Rubinstein M (1979) Option pricing: a simplified approach. *J Financ Econ* 7(3):229–263
- Haug EG, Taleb NN (2011) Option traders use (very) sophisticated heuristics, never the black–Scholes–Merton formula. *J Econ Behav Organ* 77(2):97–106
- Hull JC (2011) *Options, futures, and other derivatives*, 8th edn. Pearson, Upper Saddle River
- Jacque L (2015) *Global derivative debacles. From theory to malpractice*, 2nd edn. World Scientific Publishing, Singapore
- Merton RC (1973) Theory of rational option pricing. *Bell J Econ Manag Sci* 4:141–183
- Van Liedekerke L, Cassimon D (2000) Derivatives: power without accountability. In: Van Liedekerke L (ed) *Explorations in financial ethics*. Peeters, Leuven, pp 107–151

Option Prices Models

Peter-Jan Engelen^{1,2} and Danny Cassimon³

¹Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands

²Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

³Institute of Development Policy and Management (IOB), University of Antwerp, Antwerp, Belgium

Abstract

Option prices models refer to the overall collection of quantitative techniques to value an option given the dynamics of the underlying asset. Option prices models can be classified into European-style versus American-style models, into closed-form versus numerical expressions and into discrete versus continuous time approaches.

Definition

Quantitative techniques applied to compute the theoretical value/market price of a financial option given the characteristics and properties of the underlying asset. They can be classified into European-style versus American-style models,

closed-form versus numerical expressions, and discrete versus continuous time approaches.

references to option prices models for other underlying assets at the end.

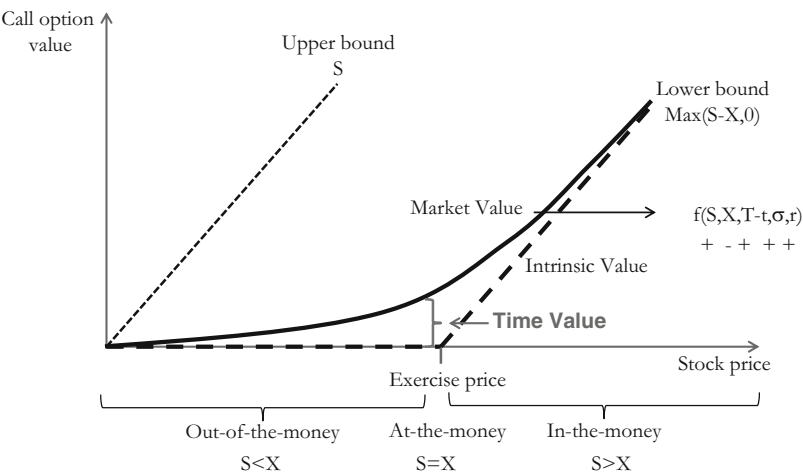
Introduction

Options give the holder the right to buy (in the case of a call option) or sell (in the case of a put option) the underlying asset (e.g., shares, stock indices, currencies) at an agreed price (strike price or exercise price) during a specific period (in the case of American options) or at a pre-determined expiration date (in the case of European options). Holding an option has a particular value, determining the (market) price (the “premium”) at which this option can be bought and sold. While many financial options are actively traded on an exchange and these market prices are easily observable, financial actors also rely on option prices models to calculate the theoretical (market) value of an option. Option prices models refer to the overall collection of quantitative techniques to value an option given the dynamics of the underlying asset. Although a complete taxonomy of option prices models is outside of the confines of this contribution, they can be classified into European-style versus American-style models, into closed-form versus numerical expressions and into discrete versus continuous time approaches. In this contribution we mainly focus on stock options, while we provide

Basic Option Relationships

As an option is a financial instrument which gives the holder the right, but not the obligation, to exercise, the typical payoff structure of an option is asymmetrical. The market value or premium of the option typically consists of the sum of the intrinsic value and the time value. It is indicated by the solid line in Fig. 1. The intrinsic value of an option is its value upon immediate exercise. The dash line in Fig. 1 indicates the intrinsic value and shows the typical hockey stick payoff profile of a call option. When the stock price S is higher than the exercise price X , the call option is said to be in the money as it takes a positive intrinsic value ($S-X$). In case the stock price is lower than the exercise price, the call option will not be exercised as it is cheaper to buy a share directly at the stock market. The call option is said to be out of the money and its value amounts to zero. The call option is said to be at the money when the stock price equals the exercise price. The intrinsic value of the option is its value at expiration date or when it is exercised earlier: the payoff to the holder of the option equals $\max(S-X,0)$. The option’s market price prior to expiration or early exercise is higher than the intrinsic value as one has to add the time value on top of it. The time value reflects the

Option Prices Models,
Fig. 1 Value of a call option as a function of the stock price



flexibility that the holder has in postponing the decision to exercise the option until the expiration date, at the latest. Figure 1 also indicates the upper and lower boundary conditions of the option value. The maximum value of the call option is the underlying stock price S . An option can never be worth more than what it costs to acquire the stock directly. The minimum value of the call option is the intrinsic value, or $\max(S-X, 0)$. Finally, the market value depends on five parameters: the stock price, the exercise price, the time to expiration, the volatility of the stock return, and the risk-free interest rate. With the exception of the exercise price, all parameters have a positive impact on the market value of the option. These are the five value drivers of any option and will be the input parameters of any option prices model. The next section discusses different modeling approaches to calculate the theoretical market price of an option.

European Option Prices Models

European option prices models refer to quantitative techniques used to value options which can only be exercised at their expiration date and where the holder has no decision to make during the lifetime of the option. Such basic plain vanilla options are typically valued using a closed-form expression such as the continuous time Black-Scholes model. When exact formulas are not available, numerical expressions are used to put an approximate value on the option. The discrete time binomial option pricing model is an

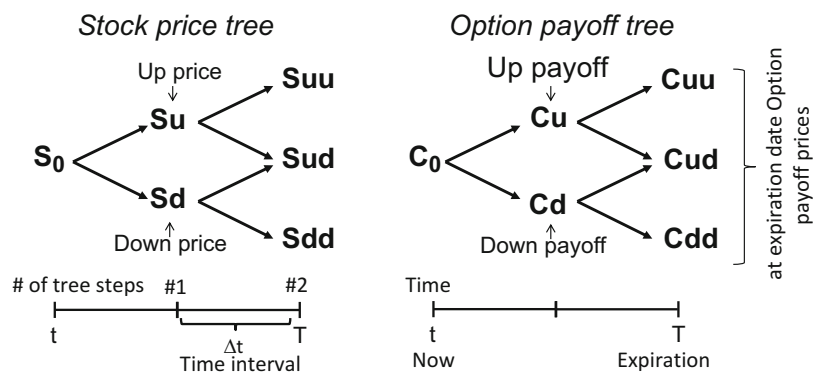
example of the latter. Both models have in common that option valuation is based on the idea of a replicating portfolio. By creating a portfolio of stocks and bonds with the same payoff as the option in every state of the world, the law of one price requires the current value of the replicating portfolio to be equal to the current option value. As one can easily observe the market prices of the assets in the replicating portfolio, one can thus also value the option. A detailed analysis of this concept is outside the confines of this contribution.

The Binomial Model

This model assumes the stock price moves in every time period Δt either upward with factor u or downward with factor d . For instance, Fig. 2 illustrates how the stock price can move in a two-period binomial model. The left panel shows how the possible price path of the underlying stock price, while the right panel shows the corresponding option payoff tree. After two time periods there are three possible stock prices at maturity date and three corresponding option prices. As the option is at expiration, its value can easily be computed as it only takes intrinsic value: the stock price at each node minus the exercise price, or zero, the highest of both. With a process called backward induction one moves one step back in the option payoff tree to calculate the option value in earlier tree steps until one reaches the current moment.

To develop a realistic tree of possible future stock prices, one needs to determine the size of one tree step and the upward and downward

Option Prices Models,
Fig. 2 Two-period
binomial trees



jumps of the stock prices. The time interval Δt depends on the lifetime of the option ($T-t$) and the number of tree steps n . The left panel in Fig. 3 illustrates how one can expand the number of possible share prices at expiration by increasing the number of tree steps. For instance, if time to expiration $T-t$ is 1 year and one uses 250 tree steps n , then the stock moves up or down every time interval $\Delta t = \frac{T-t}{n} = \frac{1}{250}$ or approximately one step equals one trading day.

To determine the value of the upward and downward movement the standard binomial option model of Cox et al. (1979) fixes $u = e^{\sigma\sqrt{\Delta t}}$ and $d = 1/u$. Once both parameters are determined, one can develop the binomial tree and calculate the current option value by discounting at each node of the tree the upward and downward expected option values:

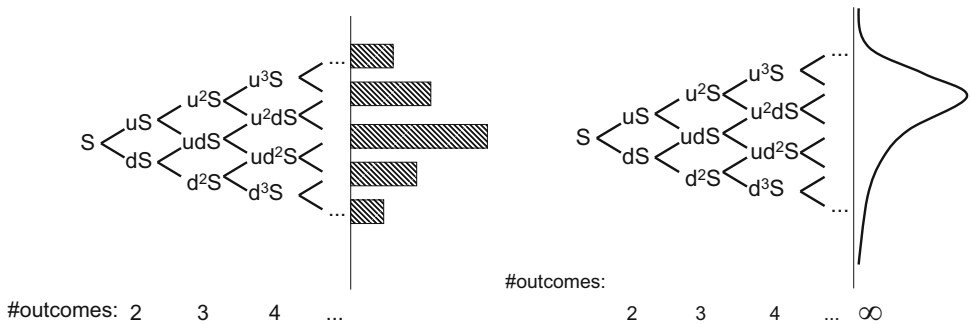
$$C = \{[C_u \times p] + [C_d \times (1 - p)]\}e^{-r} \tag{1}$$

where C is the current option price, C_u is the expected option value in the upward world, C_d the expected option value in the downward world, p the risk-neutral probability that the stock moves upward with $p = \frac{e^r - d}{u - d}$, and r the risk-free interest rate.

The right panel in Fig. 3 shows that increasing the number of tree steps to infinity produces a lognormal distribution of possible share prices at expiration. In that case, the binomial model converges to the continuous time model we discuss in the next section. In practice the binomial model produces a good approximation of the Black-Scholes model from 250 tree steps onwards (see Fig. 4).

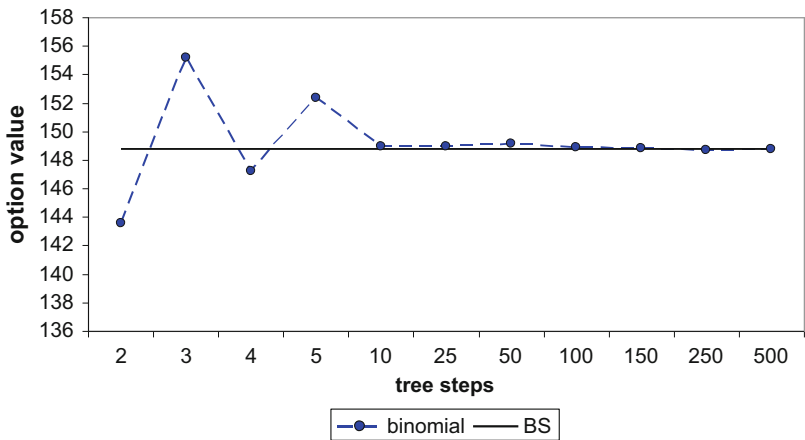
The Black-Scholes Model

Probably the best known model has been developed by Black and Scholes (1973). Its popularity is derived from its closed-form solution and fast



Option Prices Models, Fig. 3 Possible price paths and distribution of stock prices at expiration

Option Prices Models, Fig. 4 Convergence of the binomial model towards the Black-Scholes (BS) model



and relatively simple computation. Its main disadvantage is due to the strict assumptions underlying the model (Hull 2011): (i) frictionless markets, implying no transaction costs or taxes, nor restrictions on short sales; (ii) continuous trading is possible; (iii) the risk-free (short term) interest rate is constant over the life of the option; (iv) the market is arbitrage-free; and (v) the time process of the underlying asset price is stochastic and exhibits a process assuming asset prices to be lognormally distributed and returns to be normally distributed. Obviously, any violation of some of these assumptions may result in a theoretical Black-Scholes option value which deviates from the price observed at the market. The option value C according to the Black-Scholes model can be calculated as:

$$C = S N(d_1) - X e^{-r_c(T-t)} N(d_2) \quad (2)$$

$$d_1 = \frac{\ln\left(\frac{S}{X}\right) + \left(r_c + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \quad (3)$$

$$d_2 = \frac{\ln\left(\frac{S}{X}\right) + \left(r_c - \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \\ = d_1 - \sigma\sqrt{T-t}, \quad (4)$$

where S is the current stock price, X the exercise price, $T-t$ the time to expiration (in years), σ the annualized standard deviation of the stock return, r_c the continuous risk-free interest rate, and $N(d)$ the cumulative normal probability density function.

American Option Prices Models

American option prices models refer to quantitative techniques used to value options which can be exercised earlier than the expiration date or where the holder has to make other decisions during the lifetime of the option. The main difference between European and American option models is thus the incorporation of the early exercise feature. American call options on nondividend

paying stocks are equal in value than an otherwise identical European call option. Early exercise of an American call options on a nondividend paying stock would render the holder the intrinsic option value but would lose the time value of the option. Put differently, the American call option on a nondividend paying stock is worth more alive than dead. Therefore, one can value an American call option on a nondividend paying stock with European option models. The story changes when the stock distributes one or more dividends during the lifetime of the option. In that case, early exercise of American call options on stocks is a valuable alternative and depends on the tradeoff between losing the time value of the call option and capturing the dividend(s). A closed-form solution for the valuation of an American call option with one single dividend payment is offered by Geske (1979a), Roll (1977) and Whaley (1981), while Cassimon et al. (2007) offer a similar solution for the case of multiple dividends. A popular approximation is offered by Black (1975). Other valuation models for American options include trinomial models (Kamrad and Ritchken 1991), finite difference models (Brennan and Schwartz 1977, 1978), Monte Carlo simulations (Broadie and Glasserman 1997; Longstaff and Schwartz 2001; Moreno and Navas 2003), and quadratic approximations (MacMillan 1986; Barone-Adesi and Whaley 1987). A detailed discussion of these models is beyond the confines of this contribution, but we refer the reader to the references for further details.

Other Option Prices Models

Often more complicated option models are needed to handle the specific nature of the option features or the dynamics of the underlying asset. Such models include jump models (Merton 1976), compound option models (Geske 1979b; Cassimon et al. 2004), or barrier models (Li et al. 2013).

In the option literature one can also find option models for other financial assets, such as options on currency (Garman and Kohlhagen 1983; Amin and Jarrow 1991), options on futures (Black 1976;

Ramaswamy and Sundaresan 1985), options on commodities (Hilliard and Reis 1998), options on (stock) indices (Merton 1973; Chance 1986), swaptions (Black 1976), options on interest rates (Ho and Lee 1986), and bond options (Schaefer and Schwartz 1987). Option models are also used in handling credit default risk (Merton 1974; Crosbie and Bohn 2003).

References

- Amin KI, Jarrow RA (1991) Pricing foreign currency options under stochastic interest rates. *J Int Money Financ* 10(3):310–329
- Barone-Adesi G, Whaley R (1987) Efficient analytic approximation of American option values. *J Financ* 42:301–320
- Black F (1975) Facts and fantasy in the use of options. *Financ Anal J* 31(4):36–41, 61–72
- Black F (1976) The pricing of commodity contracts. *J Financ Econ* 3(1):167–179
- Black F, Scholes M (1973) The pricing of options and corporate liabilities. *J Polit Econ* 81:637–659
- Brennan M, Schwartz E (1977) The valuation of American put options. *J Financ* 32:449–462
- Brennan M, Schwartz E (1978) Finite difference methods and jump processes arising in the pricing of contingent claims: a synthesis. *J Financ Quant Anal* 13:461–474
- Broadie M, Glasserman P (1997) Pricing American-style securities using simulation. *J Econ Dyn Control* 21:1323–1352
- Cassimon D, Engelen PJ, Thomassen L, Van Wouwe M (2004) Valuing new drug applications using n-fold compound options. *Res Policy* 33:41–51
- Cassimon D, Engelen PJ, Thomassen L, Van Wouwe M (2007) Closed-form valuation of American call-options with multiple dividends using n-fold compound option models. *Financ Res Lett* 4:33–48
- Chance DM (1986) Empirical tests of the pricing of index call options. *Adv Futur Options Res* 1(Part A):141–166
- Cox JC, Ross SA, Rubinstein M (1979) Option pricing: a simplified approach. *J Financ Econ* 7(3):229–263
- Crosbie P, Bohn J (2003) Modeling default risk, white paper. Moody's KMV, USA
- Garman MB, Kohlhaugen SW (1983) Foreign currency option values. *J Int Money Financ* 2(3):231–237
- Geske R (1979a) A note on an analytic valuation formula for unprotected American call options on stocks with known dividends. *J Financ Econ* 7:375–380
- Geske R (1979b) The valuation of compound options. *J Financ Econ* 7:63–81
- Hilliard JE, Reis J (1998) Valuation of commodity futures and options under stochastic convenience yields, interest rates, and jump diffusions in the spot. *J Financ Quant Anal* 33(01):61–86
- Ho T, Lee S (1986) Term structure movements and pricing interest rate contingent claims. *J Financ* 41(4):1011–1029
- Kamrad B, Ritchken P (1991) Multinomial approximating models for options with k state variables. *Manag Sci* 37:1640–1652
- Li Y, Engelen PJ, Kool C (2013) A barrier options approach to modeling project failure: the case of hydrogen fuel infrastructure, Tjalling C. Koopmans Research Institute Discussion Paper Series 13–01, 34p
- Longstaff F, Schwartz E (2001) Valuing American options by simulation: a simple least-squares approach. *Rev Financ Stud* 14:113–147
- MacMillan L (1986) Analytic approximation for the American put option. *Adv Futur Options Res* 1:119–139
- Merton R (1973) Theory of rational option pricing. *Bell J Econ Manag Sci* 4(1):141–183
- Merton RC (1974) On the pricing of corporate debt: the risk structure of interest rates. *J Financ* 29(2):449–470
- Merton RC (1976) Option pricing when underlying stock returns are discontinuous. *J Financ Econ* 3(1):125–144
- Moreno M, Navas J (2003) On the robustness of least-squares Monte Carlo (LSM) for pricing American derivatives. *Rev Deriv Res* 6:107–128
- Ramaswamy K, Sundaresan SM (1985) The valuation of options on futures contracts. *J Financ* 40(5):1319–1340
- Roll R (1977) An analytic formula for unprotected American call options on stocks with known dividends. *J Financ Econ* 5:251–258
- Schaefer SM, Schwartz ES (1987) Time-dependent variance and the pricing of bond options. *J Financ* 42(5):1113–1128
- Whaley R (1981) On the valuation of American call options on stocks with known dividends. *J Financ Econ* 9:207–211

Recommended Readings

- Derivagem (2014) Option calculator at <http://www-2.rotman.utoronto.ca/~hull/software/index.html>
- Hull JC (2011) Options, futures, and other derivatives, 8th edn. Pearson, UpperSaddle River

Ordoliberalism

Stefan Kolev

Wilhelm Röpke Institute, Erfurt and University of Applied Sciences Zwickau, Zwickau, Germany

Abstract

The purpose of this entry is to delineate the political economy of a group of authors related to ordoliberalism, the German variety of neo-liberalism. To reach this goal, a history of

economics approach is harnessed. First, the historical background around the time of its inception is presented. Second, an overview of the substantive core and of the heterogeneity across the different strands of ordoliberalism is reconstructed. Third, its historical impact on and current potential for political economy are evaluated. While this entry focuses on the general patterns of ordoliberal political economy, separate entries delineate the specificities in the research programs of Walter Eucken, Wilhelm Röpke, and Franz Böhm.

Introduction

A multifaceted group of scholars with a research program at the intersection of political economy, law, and social philosophy has become famous since 1950 under the name of “ordoliberalism” (Moeller 1950). The intellectual history of this group has received much attention, as ordoliberalism has ever since unfolded a seminal impact in Germany, Central Europe, and beyond. This entry has a threefold purpose: first to delineate the history of ordoliberal thought, second to present the overarching research agenda of the ordoliberals, and third to assess its impact and potential for today. Given the heterogeneity and the implicit division of labor within the group, this overview is complementary to the entries on Walter Eucken, Franz Böhm, and Wilhelm Röpke which present the specificities as detailed in the respective primary literature.

“Ordoliberalism” is sometimes used synonymously with the term “Freiburg School,” which is imprecise. While the Freiburg School around Walter Eucken (1891–1950) and Franz Böhm (1895–1977) constitutes the core of ordoliberal group, also authors who strictly speaking do not belong to the Freiburg School have been perceived as representatives of ordoliberal thought. Among them, the most prominent are Wilhelm Röpke (1899–1966), Alexander Rüstow (1885–1963), and Alfred Müller-Armack (1901–1978), but also F.A. Hayek (1899–1992) at certain stages in his evolution can be seen as contributing to the ordoliberal research program

(Oliver 1960, pp. 118–119). Despite all cultural and biographical differences, it is important to note that these thinkers broadly belong to the same generation and also share a common goal in the 1930s and 1940s: to revitalize liberalism whose standing and relevance at that point are at an all-time low. For understanding the inception of ordoliberalism and its particular substantive core, a twofold perspective is necessary: one which sheds light on the specificities of German economics in the 1930s and another which embeds the group in the general movement of international scholarship, self-depicting itself around 1938 as “neoliberalism.”

Formative Years

German economics was in a deplorable state in the early 1930s. The academic landscape was largely characterized by the ruins of the Younger and the Youngest Historical Schools (Rieter 2002, pp. 154–162; Goldschmidt 2013, pp. 127–129). The tragedy of the discipline became most evident in its lacking capability to provide solutions to the extremely pressing problems of economic life, the hyperinflation in the early 1920s as well as the Great Depression in the early 1930s. Instead, many German economists still fought methodological battles or collected empirical evidence, often lacking the theoretical underpinnings for such inquiries. Dissidents to this development constituted in the 1920s the group of the so-called Ricardians, a collection of younger scholars who intended to re-link the discipline back to modes of systematic theorizing – with Rüstow, Eucken, and Röpke among the most active members of the group (Janssen 2009, pp. 34–48; Köster 2011, pp. 224–228). Even though the annus horribilis of 1933 physically separated them – with Rüstow and Röpke leaving as exiles to Turkey and Eucken staying as a “half-exile” (Johnson 1989, p. 57) in Germany – the intellectual ties remained and even solidified over the decades to follow. In the immediate aftermath of 1933, Eucken formed a scholarly community at Freiburg with Franz Böhm and another law professor, Hans Großmann-Doerth (1894–1944),

which became the core of what would later be called the Freiburg School of Ordoliberalism or the Freiburg School of Law and Economics (Vanberg 1998, 2001). In 1936, the three published a manifesto under the title “Our Mission” (Böhm et al. 2008). Eucken and Böhm opposed National Socialism in manifold academic contexts, among others vis-à-vis the 1933–1934 rectorate of Martin Heidegger as well as in diverse intellectual resistance groups preparing for the age after National Socialism, later to be called the Freiburg Circles (Rieter and Schmolz 1993, pp. 95–103; Nicholls 1994, pp. 60–69).

The overarching goal of the Freiburg School and its 1948-founded publication organ, the “ORDO Yearbook of Economic and Social Order,” is the search for an order which entails a minimum of power, considering all forms of power relations – stemming from the state, from the market, and from other social orders (Sally 1998, pp. 110–117; Foucault 2008, pp. 129–158). The seminal analytical distinction of the Freiburg School is between “order” and “process”: while the level of economic order is one of the “rules of the game,” the level of economic process is one of the “moves of the game” (Wohlgemuth 2013, pp. 157–159). Building upon this distinction, the Freiburg School conceived a term which until today is considered by many its trademark – “Ordnungspolitik.” Based on his theory of orders and the above distinction, Eucken formulated the task for the power-fighting state: to set the “rules of the game” by “Ordnungspolitik” (translatable as “order-based policy” or, more broadly, as “rule-based policy”) for enabling Eucken’s normative vision of the “competitive order” but to abstain from intervening in the “moves of the game” which are the protected domain of the private individuals with their interactions. With this, Eucken aimed to draw the border to interventionism which disregards such constraints and arbitrarily intervenes in the “moves of the game” of the economic process for improving the performance of the economy – instead, his theory and his perspective on economic policy are very much in line with the “problem of constitutional choice, i.e., as a question of how desirable economic order can be

generated by creating an appropriate economic constitution” (Vanberg 2001, p. 40).

Complementary to this agenda at Freiburg, the exiles Röpke and Rüstow developed a strand in ordoliberalism which would later be called “sociological liberalism” (Renner 2002, p. 61). Their inquiry was focused on the question how social stability and cohesion are possible amid the challenges of “enmassment” of modern society with its uprooting of the individuals from their traditionally cohesive small groups (Friedrich 1955, pp. 512–525; Kolev 2015, pp. 426–427). Just as Eucken and his associates, Röpke and Rüstow understood the economy and its framework as being embedded into the other social orders, but unlike Eucken they saw the task of economists to also look beyond the context of the economic order and to explicitly theorize its interrelationship to the other social orders, as the potential sources of instability within such an interdependent system of social orders can lie well beyond the economic order (Zweynert 2013, pp. 116–120). The task for a political economist is also to be aware of one basic property of the economic order: that it tends to use up its own stability resource, so that these resources and pillars (called by Röpke “anthropological and sociological” prerequisites) have to be permanently checked and, if necessary, stabilized anew by the state and other players in civil society (Kolev 2013, pp. 133–136).

Impact in Later Decades and Relevance for Today

The quests of the ordoliberal exiles and “half-exiles” are also to be contextualized within the broader intellectual perspective of Europe and the United States at their time. Freiburg was not completely isolated – Röpke, Rüstow, and Hayek stayed in touch with Eucken, in Röpke’s case even well into the year 1943. Even though Eucken did not attend the Colloque Walter Lippmann in 1938, Röpke and Rüstow were among its most active participants, proposed the term “neoliberalism” for the movement of revitalizing liberalism, and also had severe debates on what neoliberalism

should mean along the maxims of ordoliberal political economy, in particular with Ludwig von Mises (Gregg 2010, pp. 82–86; Burgin 2012, pp. 67–78; Kolev 2016, pp. 14–17). Hayek’s evolution is of special interest here, as in the late 1930s and in the 1940s, his incipient political economy, including “The Road to Serfdom,” displayed strong similarities to the ordoliberal research program (Streit and Wohlgemuth 2000; Kolev 2010, 2015, pp. 432–436). Interestingly, a research program with yet again strong similarities to the one in Freiburg came up in Chicago of the time (Van Horn 2009; Köhler and Kolev 2013), later to be called by James M. Buchanan the “Old Chicago” School (Buchanan 2012), and it was Hayek who let Freiburg and “Old Chicago” meet in the session on the potential of the competitive order to become an alternative to *laissez-faire* at the very first meeting of the Mont Pèlerin Society in 1947 (Kolev et al. 2014, pp. 15–20).

The influences to and from ordoliberalism are not confined to the realm of academic economics. All ordoliberals named above were not of the ivory-tower scholar type – their political economies explicitly aimed at becoming tools for practical economic policy. The ways for channeling their theories into practical economic policy were different: Müller-Armack’s 1946-coined term of the Social Market Economy became a successful device for introducing the market economy to capitalism-skeptical German postwar society (Dornbusch 1993, pp. 881–883; Watrin 2000) but has later been used in countries as diverse as Chile, Belarus, or the European Union to frame the economic policy agenda of different administrations. Eucken’s “Ordnungspolitik” is until today a powerful rhetorical tool in Germany to introduce market-oriented reforms or to block interventionist policy proposals as being “Ordnungspolitik”-incompatible. Röpke’s countless appearances in Swiss, German, and international media made him the most active public intellectual of all ordoliberals and until today he is an authority in economic policy matters quoted especially often in Switzerland. Böhm made a longer digression into the realm of parliamentary politics and, after several years of intense debates especially with the Federation of German Industry, was a principal

figure in introducing the “Act against Restraints of Competition” (GWB) in 1958, later to become a leading norm for the formulation of antitrust law in the European Union (Giocoli 2009). Until today, also macroeconomic issues like the optimal monetary policy of the European Central Bank or the fiscal policy debt-brake legislation in European Union countries have been discussed with reference to the rule-based approach of “Ordnungspolitik” (Dullien and Guérot 2012; Kundnani 2012; Wren-Lewis 2012; Feld et al. 2015; Bofinger 2016).

In addition, there have been attempts not only to perceive ordoliberalism as a history of economics artifact or as a guiding tool for practical economic policy, but also to connect it to new rule-based approaches in institutional and constitutional economics (Vanberg 1988; Ebeling 2003; Leipold and Wentzel 2005; Boettke 2012; Zweynert et al. 2016). Considering the multiple economic crises of our time but also the multiple crises of today’s economics as an allegedly useless academic discipline in the eyes of the general public and of many politicians, students, and neighboring social sciences, ordoliberalism’s rich heritage can be one of the sources necessary to revitalize the discipline and to regain the confidence of its currently dissatisfied “customers.”

Cross-References

- ▶ [Austrian School of Economics](#)
- ▶ [Böhm, Franz](#)
- ▶ [Constitutional Political Economy](#)
- ▶ [Eucken, Walter](#)
- ▶ [Hayek, Friedrich August von](#)
- ▶ [Mises, Ludwig von](#)
- ▶ [Political Economy](#)
- ▶ [Röpke, Wilhelm](#)
- ▶ [Schmoller, Gustav von](#)

References

- Boettke PJ (2012) Methodological individualism, spontaneous order, and the research program of the workshop in political theory and policy analysis: Vincent and Elinor Ostrom. In: Boettke PJ (ed) *Living economics*:

- yesterday, today, and tomorrow. Independent Institute, Oakland, pp 139–158
- Bofinger P (2016) German macroeconomics: the long shadow of Walter Eucken. CEPR's Policy Portal of June 7 2016. Available at: <http://voxeu.org/article/german-macroeconomics-long-shadow-walter-eucken>
- Böhm F, Eucken E, Großmann-Doerth H (2008) Unsere Aufgabe. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, pp 27–37; English translation. In: Peacock A, Willgerodt H (eds) (1989) Germany's social market economy: origins and evolution. St. Martin's Press, New York, pp 15–26
- Buchanan JM (2012) Presentations 2010–2012 at the Jepson school of leadership, Summer Institute for the history of economic thought, University of Richmond. Available at: <http://jepson.richmond.edu/conferences/summer-institute/index.html>
- Burgin A (2012) The great persuasion: reinventing free markets since the depression. Harvard University Press, Cambridge, MA
- Dornbusch R (1993) The end of the German miracle. *J Econ Lit* 31(2):881–885
- Dullien S, Guérot U (2012) The long shadow of ordoliberalism: Germany's approach to the euro crisis. Policy paper 49, European Council on Foreign Relationships, London
- Ebeling RM (2003) The limits of economic policy: the Austrian economists and the German ORDO liberals. In: Ebeling RM (ed) Austrian economics and the political economy of freedom. Edward Elgar, Cheltenham, pp 231–246
- Feld LP, Köhler EA, Nientiedt D (2015) Ordoliberalism, pragmatism and the Eurozone crisis: how the German tradition shaped economic policy in Europe. *Eur Rev Int Stud* 2(3):48–61
- Foucault M (2008) The birth of biopolitics. Lectures at the College de France 1978–1979. Palgrave MacMillan, New York
- Friedrich CJ (1955) The political thought of neo-liberalism. *Am Polit Sci Rev* 49(2):509–525
- Giocoli N (2009) Competition versus property rights: American antitrust law, the Freiburg school, and the early years of European Competition policy. *J Competition Law Econ* 5(4):747–786
- Goldschmidt N (2013) Walter Eucken's place in the history of ideas. *Rev Austrian Econ* 26(3):127–147
- Gregg S (2010) Wilhelm Röpke's political economy. Edward Elgar, Cheltenham
- Janssen H (2009) Nationalökonomie und Nationalsozialismus. Die deutsche Volkswirtschaftslehre in den dreißiger Jahren des 20. Jahrhunderts, 3rd edn. Metropolis, Marburg
- Johnson D (1989) Exiles and half-exiles: Wilhelm Röpke, Alexander Rüstow and Walter Eucken. In: Peacock AT, Willgerodt H (eds) German neo-liberals and the social market economy. St. Martin's Press, New York, pp 40–68
- Köhler EA, Kolev S (2013) The conjoint quest for a liberal positive program: “Old Chicago”, Freiburg, and Hayek. In: Peart SJ, Levy DM (eds) F. A. Hayek and the modern economy: economic organization and activity. Palgrave MacMillan, New York, pp 211–228
- Köster R (2011) Die Wissenschaft der Außenseiter. Die Krise der Nationalökonomie in der Weimarer Republik. Vandenhoeck & Ruprecht, Göttingen
- Kolev S (2010) F.A. Hayek as an ordo-liberal. Research paper 5–11, Hamburg Institute of International Economics, Hamburg
- Kolev S (2013) Neoliberale Staatsverständnisse im Vergleich. Lucius & Lucius, Stuttgart
- Kolev S (2015) Ordoliberalism and the Austrian school. In: Boettke PJ, Coyne CJ (eds) The Oxford handbook of Austrian economics. Oxford University Press, Oxford, pp 419–444
- Kolev S (2016) Ludwig von Mises and the “ordo-interventionists” – more than just aggression and contempt? Working paper 2016–35, Center for the History of Political Economy at Duke University, Durham
- Kolev S, Goldschmidt N, Hesse J-O (2014) Walter Eucken's role in the early history of the Mont Pèlerin society. Discussion paper 14/02, Walter Eucken Institute, Freiburg
- Kundnani H (2012) The eurozone will pay a high price for Germany's economic narcissism. *Guardian* of January 6 2012. Available at: <https://www.theguardian.com/commentisfree/2012/jan/06/eurozone-germany-ordoliberalism>
- Leipold H, Wentzel D (eds) (2005) Ordnungsökonomik als aktuelle Herausforderung. Lucius & Lucius, Stuttgart
- Moeller H (1950) Liberalismus. *Jahrb Natl Stat* 162:214–240
- Nicholls AJ (1994) Freedom with responsibility. The social market economy in Germany, 1918–1963. Clarendon, Oxford
- Oliver HM (1960) German neoliberalism. *Q J Econ* 74(1):117–149
- Renner A (2002) Jenseits von Kommunitarismus und Neoliberalismus. Eine Neuinterpretation der Sozialen Marktwirtschaft. Vektor, Grafschaft
- Rieter H (2002) Historische Schulen. In: Issing O (ed) Geschichte der Nationalökonomie, 4th edn. Vahlen, Munich, pp 131–168
- Rieter H, Schmolz M (1993) The ideas of German ordoliberalism 1938–1945: pointing the way to a new economic order. *Eur J Hist Econ Thought* 1(1):87–114
- Sally R (1998) Classical liberalism and international economic order: studies in theory and intellectual history. Routledge, London
- Streit ME, Wohlgemuth W (2000) The market economy and the state: Hayekian and ordoliberal conceptions. In: Koslowski P (ed) The theory of capitalism in the German economic tradition: historicism, ordo-liberalism, critical theory, solidarism. Springer, Berlin, pp 224–271
- Van Horn R (2009) Reinventing monopoly and the role of corporations: the roots of Chicago law and economics. In: Mirowski P, Plehwe D (eds) The road from Mont Pèlerin: the making of the neoliberal thought collective.

Harvard University Press, Cambridge, MA, pp 204–237

Vanberg VJ (1988) “Ordnungstheorie” as constitutional economics. The German conception of a “social market economy”. *ORDO Jahrb Ordn Wirtsch Ges* 39:17–31

Vanberg VJ (1998) Freiburg school of law and economics. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*, vol 2. Palgrave MacMillan, London, pp 172–179

Vanberg VJ (2001) The Freiburg school of law and economics: predecessor of constitutional economics. In: Vanberg VJ (ed) *The constitution of markets. Essays in political economy*. Routledge, London, pp 37–51

Watrin C (2000) Alfred Müller-Armack – economic policy maker and sociologist of religion. In: Koslowski P (ed) *The theory of capitalism in the German economic tradition: historism, ordo-liberalism, critical theory, solidarism*. Springer, Berlin, pp 192–220

Wohlgemuth M (2013) The Freiburg school and the Hayekian challenge. *Rev Austrian Econ* 26(3):149–170

Wren-Lewis S (2012) Anti-Keynesian Germany. Blog “mainly macro”: comment on macroeconomic issues of March 9 2012. Available at: <https://mainlymacro.blogspot.com/2012/03/anti-keynesian-germany.html>

Zweynert J (2013) How German is German neo-liberalism? *Rev Austrian Econ* 26(3):109–125

Zweynert J, Kolev S, Goldschmidt N (eds) (2016) *Neue Ordnungsökonomik*. Mohr Siebeck, Tübingen

Organization

Mehmet Bac

Sabancı University, Istanbul, Turkey

Abstract

Organizations exist because they cost-effectively coordinate and provide incentives relative to alternative modes of transacting. The article exposes some of the main economic explanations for why organization emerges, what it does and how its scale and scope are determined, highlighting the role of a hierarchical command structure in aligning the members’ objectives with those of the organization and successfully overcoming problems of asymmetric information and collusion.

Synonyms

Firm; Government; Institution

Definition

Organization is an entity designed to influence behavior and reach collective objectives for its members.

Numerous classifications of organization can be made to emphasize its different functions or characteristics, according to objectives, output type, ownership structure, hierarchical network, degree of delegation of decisions and divisionalization, financial structure, reliance on technology and innovation, measurability of outputs and inputs, power of incentives, size in labor force, output or asset value, and scope measured vertically or horizontally in terms of diversification of output and tasks. In the modern state, almost all economic and social activities involve some form of integration and organization. The *raison d’être* of organization is cost-effective coordination and provision of incentives in transactions where private negotiations by independent units fail to generate satisfactory outcomes.

The new institutional approach to organization of activity, common in economics and legal studies, puts the *transaction* to the forefront as the unit of analysis. The key to understanding the organization and its boundaries is thus the nature of transaction costs, viewing the organization as a solution by the actors to executing transactions at minimal cost (Coase 1937).

The informational requirements for value-maximizing coordination of a wide range of activities are huge, by far exceeding the signals contained in market prices. The organization coordinates along multiple dimensions. It allocates productive and consumptive resources across tasks and members. It coordinates over time as well: to “fit the pieces” properly, avoid effort duplication, and produce value in rather unpredictable environments, the organization solves assignment and synchronization problems. It induces mutual adjustments and achieves standardization in processes, outputs, knowledge, and even norms and values. By creating a moral order and identity in the minds of its members, the organization copes better with the rise in communication and coordination costs that follow specialization of its labor motivated by productivity.

Transaction costs have two main sources. The first, *bounded rationality*, is the limited capacity of humans to foresee the relevant contingencies and describe with precision those foreseen, including the corresponding set of actions that should be taken in each contingency. Bounded rationality implies that the parties to a transaction will draft an incomplete contract, with gaps and missing provisions. Incompleteness is manifest in the contractual terms of idiosyncratic transactions in which parameters such as “quality” and “know-how” are important, yet difficult to define unambiguously. The resulting cost for the parties can be huge because contract incompleteness invites opportunistic behavior. Organization may cope better with contract incompleteness because it replaces the costly renegotiations in unforeseen contingencies by a command structure where people are told what to do. Integration of a transaction is supplemented by implementing an appropriate incentive scheme, instituting a monitoring and command system, and bringing to the forefront career concerns of the parties, inside and outside the organization.

The second source of transaction costs is information problems. Members at all levels of the organization have private information, but their goals are not necessarily congruent with those of the organization. The simplest nonmarket paradigm to study the information problems and their incentive implications is the principal-agent framework. In its canonical form, the principal makes a take-it-or-leave-it contract offer to the agent who, if the parties agree, performs a task which the principal cannot for lack of time or expertise. The principal faces three types of informational obstacles in motivating the agent to take the appropriate action: *hidden action* (moral hazard), which refers to unobservability of the agent’s actions; *hidden knowledge* (adverse selection), which refers to the agent’s task-relevant private information; and *verifiability* of relevant parameters and contract terms by third parties, such as a court of law.

The challenge in controlling the hidden action problem is instituting objective and verifiable performance measures that are somehow correlated with the agent’s unobservable actions, then, using

these measures to tailor an incentive package which indirectly and cost-effectively aligns the agent’s goals with those of the principal. However, because the relation between the agent’s actions and performance measures is typically non-deterministic, raising the power of incentives inevitably exposes the agent to risk. This generates a utility loss which the principal bears in the form of a “risk premium” if the agent is risk averse. The solution to the hidden action problem thus poses a dilemma for the organization, namely, the need to balance incentive provision costs against the costs of risk bearing (Grossman and Hart 1983). In practice, firms may choose to offer stock options to their CEOs to motivate them so that they in turn motivate the employees, but the considerable indeterminacy in stock prices coupled with the obstacles in measuring performance typically reduces the efficacy of such performance-based payment schemes (Chiappori and Salanié 2003). These negative consequences amplify in complex and bureaucratized, multi-layer organizations (Williamson 1985). Smaller, owner-controlled business organizations have an advantage in coping with incentive problems that are rooted in unobservability of its members’ actions.

The second informational obstacle, hidden knowledge, usually takes the form of private information about individual characteristics, some task specific and others general, such as effort cost or endurance, creativity, and productivity in teams. These are crucial pieces of information for efficient matching of tasks with employees, physical with human resources. Screening methods before and during employment such as interviews, contract design, and probation terms can be used to extract information from the agents, yet in practice the power of these instruments is limited and agents enjoy information rents. To protect their rents, low-cost agents have no incentive to reveal their productivity if the principal is not able to commit to not using the information to the detriment of the agent. The expectation that revelation of productivity will raise future performance standards is known as the ratchet effect (Freixas et al. 1985).

The inability to measure and verify relevant performance parameters such as quality, effort, output, and payments is the third informational obstacle to motivation. The distinction between observability and verifiability is important here. The scope for opportunistic behavior remains strong even if the parties can perfectly observe the relevant parameters, provided courts are not able to verify them. The theory of repeated games suggests that third-party verification is not an issue because cooperative behavior can be sustained as a (subgame-perfect) Nash equilibrium as long as the parties perfectly observe the outcome, but this conclusion collapses when the parties heavily discount future payoffs or if the frequency of transaction is not sufficiently high.

The information problems and their inefficiency consequences amplify in teams, where two or more agents are expected to contribute to a set of common objectives. Each agent has an incentive to free ride, withhold effort, and rely on others. Peer monitoring and pressure are often not strong enough to fully overcome this problem. Restoring optimal incentives in a team context can be very costly too, as the principal may have to pay huge individual rewards for a successful team outcome to fully eliminate the possibility of free riding. Handling these costs leads to differences in the organization's size, incentive structure, and architecture.

Organization outperforms alternative modes of governance for idiosyncratic transactions that are projected to occur frequently, based on specialized needs (Williamson 1985). Task specialization and relationship-specific investments on assets and knowledge generate dependencies. Because these investments are worth more for the transacting parties in their relation than outside, ideally the relationship should be governed by long-term contracts that fully protect relationship-specific investments. But contracts are incomplete and open to opportunism, which suggests the best environment to protect relationship-specific investments is the vertically integrated command structure of the organization (Grossman and Hart 1986).

The organization can be viewed as a nexus of incomplete contracts between principals and

agents in a cascade of relationships (Alchian and Demsetz 1972). The interactions are reflected in a formal architecture, usually a hierarchical structure that shows the degree of centralization, how the constituent units are arranged, authority allocated, and tasks assigned. The channels of this architecture produce decisions, performance supervision at multiple levels, and manage information flows. Aspects of information processing include data generation and establishing a network through which information, processed or raw, is transmitted with maximal processing capability and minimal delay. The lesser the informational disadvantage of a unified governance and the greater the need for coordination, the greater the advantage of organization relative to its alternatives. The organization also has an advantage in performing pressing tasks with minimal delay (Bolton and Farrell 1990).

An organization is informationally efficient if its network minimizes the need for extra information in checking whether a new project or plan of action is worthwhile. Though a reversed tree form of hierarchy can minimize processing delays, usually this is not a fully symmetric, balanced hierarchy. Rather, it is optimal to involve decision nodes at upper levels of the hierarchy in various stages of the data processing exercise (Radner 1992). Placing a group of equally effective people in parallel as a polyarchy ends up selecting a larger number of available projects than a linear hierarchy order where final approval requires sequential approval by all. Therefore, the incidence of type-I statistical error is higher in hierarchy, whereas polyarchy displays higher errors of type II. The relative performance of the two modes, polyarchy or hierarchy, depends on the importance of the two error types (Sah and Stiglitz 1986). Because projects or new ideas vary in this respect, the organization's flexibility to form polyarchic or hierarchic screens in decision-making is an asset. Organizational form also matters, in particular, regarding coordination of changes and the scale of experimentation with new projects. The M-form has an advantage in coordinating attribute matching and flexibility in choosing the scale of experimentation, whereas the

U-form organization achieves better coordination in attribute compatibility (Qian et al. 2006).

Although the formal chart may look hierarchical, many activities and interactions within the organization do not necessarily follow hierarchy. Delegation and decentralization is the norm. The level of decentralization is often defined in terms of the span of control. So, of two organizations with six people, the one in which five specialists support a middle-level manager is more centralized than one in which two pairs of specialists support two middle-level managers. But the real allocation of authority may be quite different, as the organization may selectively and informally authorize the subordinates while formally assigning authority to an overloaded superior. This structure is shown to serve indirectly the motivation of subordinates who know that their ideas and preferences will sometimes be implemented (Aghion and Tirole 1997). The optimal overlap between real and formal authority balances these motivational benefits against the risks resulting from dissipation of control, a trade-off that depends on the degree of congruence between the goals of people working at different formal levels of the hierarchy.

Large organizations in particular are vulnerable to *influence activities*. Members spend time and resources to affect the decisions made at upper levels, typically in the form of producing and manipulating evidence to develop credentials for promotion. The resulting loss in the organization's output is termed influence costs (Milgrom and Roberts 1988). In response, the organization should seek an optimal balance between incentives in promotion and incentives in actual positions, as well as credibly modifying the criteria to reduce the attractiveness of positions awaiting promotion. Though improved transparency by adopting clearer rules and criteria is generally considered an organizational asset, it also reduces the cost of identifying whom to influence. If the latter effect dominates, marginal improvements in transparency may well lead to an increase in influence activities (Bac 2001).

Public organizations, including government and most nonprofits, differ from profit-seeking organizations in both fundamental and qualitative

respects. The differences are reflected in their incentive structures and hierarchy. It is generally observed that members of organizations with social or political missions face weaker direct incentives than the employees of a firm. This is partly explained by the higher degree of congruence in the goals of the social or political organization and its members, which reduces the need for powerful incentives. However, the prime explanation is in the feasibility of direct incentive schemes. Public organizations pursue multiple goals which typically are hard to measure and even define. Some of their outputs and services are laden with amenity potential, contributing directly to the owners' and top managers' utility (Demsetz and Lehn 1985). Multitask models of hidden action have shown that when some goals are measurable and others not, an incentive scheme based on the measurable goals distorts actions to the detriment of the non-measurable goals and, hence, may well worsen the overall performance (Holmstrom and Milgrom 1991). If it does, the organization should reduce the power of direct incentives and possibly rely more on alternative instruments of motivation such as monitoring and emphasizing promotion and career concerns (Holmstrom 1999). A more radical response is to create independent units with clear and distinct goals or, if possible, isolate and confine the measurable goals to new specific divisions where high-powered incentive schemes become feasible and effective. Indeed, units and agencies of government are often characterized by independent missions (Wilson 1989; Tirole 1994). Multi-divisional frameworks may generate a loss of coherence in the overall decision-making process, but the benefits from subjecting the units to different masters under tight systems of checks and balances along financial, operational, and implementation dimensions can be notable. In essence, if incentivized to properly defend the cases for alternative causes, competition among advocates of specific goals and interests in the organization may give rise to good policy setting, similar in function to the adversarial adjudication system in courts of law.

Another threat against which all organizations must carefully safeguard is the possibility of

collusion, a formation of an informal coalition to promote the common interests of its members (Tirole 1986). These organizations inside the organization are products of the information problems, combination of hidden action, hidden knowledge, and non-verifiability. Internal collusion comprises members only, whereas external collusion, named “capture” in the regulation context, involves outsiders as well. A supervisor who conceals an employee’s corruption for a benefit in return is internally colluding. Favoritism in transactions with outsiders and the capture of a procurement officer’s or an environmental agency’s decision by a private firm are incidences of external collusion. Internal and external collusion can coexist, as in the case of a coalition of officials which organizes the collection and sharing of corrupt proceeds from the public. Collusion can be temporary, a spontaneous deal *ex post* at the occasion, or it can take a sustained form whereby the parties involved coordinate *ex ante* their actions toward the common collusive objective (Bac 1996).

Sustaining collusion is subject to the same qualitative problems the organization faces in sustaining itself. But the threat of collusion imposes new constraints on incentive schemes, the rule-discretion balance, and the span of control. Stricter decision rules reduce the need to rely on the private information of members, thereby, the stakes of outsiders whom they are transacting with. The cost of a shift to rules is the foregone benefit from occasional and selective use of discretion. The rule to auction all contracts, for example, works extremely well on the price dimension and minimizes the likelihood of capture at the cost of other contract dimensions such as quality, speed, reliability, and reputation of the contractor, which typically are non-measurable but observable by an expert, manager, or official. Alternative instruments to reduce the payoffs to collusion include raising the rewards to supervision and evidence of good performance, introducing rotation, and intensifying monitoring. The cures carry along their risks: monitors are subject to the same threat of collusion, unless replaced by incorruptible electronic accountability systems or cameras. Large rewards to human monitors can prevent collusion

but are obviously costly and risk inducing extortion with fabricated evidence of wrongdoing. As for rotation policies, they prevent development of trust-based relations that help in sustaining collusive behavior, but trust is also an input for cooperation and productivity in teamwork.

Governance of internal and external transactions and acts is subject to labor, administrative, tort, and criminal laws. Internal administrative procedures are activated upon reports of embezzlement and misdemeanor such as harassment, rules on hiring and firing decisions as well as ensuring job safety standards are imposed by labor laws, and strict (vicarious) liability is instituted for torts and crimes committed by employees. Just as the internal rules of the organization serve to align the objectives of its members with those of the organization, these laws can be viewed as a response to the dissonance between the goals of the organization and the society. So, by holding firms strictly liable for the actions of their employees rather than imposing individual liability, the state *de facto* delegates monitoring and partially also enforcement of offenses to the firms. Because the firm is in close relation to its employees and better informed about its environment, it can, less expensively than the state, monitor, prevent, and sanction. The firm is then expected to align incentives and partially shift the liability to its employees, passing the sanctions along to those in violation of the law using its own instruments such as wage policy, firing, and other measures. The lower the ability of the employees to pay for the social damage of their acts and the more effective the firm is in sharing its liability with its employees, the stronger the case for strict liability laws (Posner 1999).

The dominant new institutional economic model of the large corporate organization views shareholders as principals and the board of directors as their agents who monitor the managers on behalf of the shareholders. As an intermediate layer between shareholders and managers, the board is not immune to collusion with management against the shareholders. The risk of collusion is in part remedied by the market for corporate control, in part by equity holdings of the directors. This model is criticized for

neglecting the fact that organization is a distinct legal entity with rights on its own, protected by the state. It is challenged by new theories proposed to understand corporate governance with directors as the principals and the firm as a productive team where relationship-specific investments are emphasized (Blair and Stout 1999). From a legal point of view, shareholders' primacy is questionable, and ownership of the shares does not entail ownership of the corporation, which is an autonomous legal entity. The board of directors is entrusted to act on behalf of and for the benefit of the shareholders, with a fiduciary duty to review the decisions and ensure that they serve the corporation's interests, and a legal right to control the managers' decisions. Directors are motivated not only by their extrinsic compensation package but reputational concerns as well. They act as a mediating principal balancing the competing interests of the team, which justifies their highly independent legal status to carry out this role effectively.

Theories of organization lack a unified approach in explaining the emergence and function of organization as well as its internal practices comparatively, in contrast with a comprehensive set of alternative modes of transacting. A key feature of these alternatives is that they, too, are "organized" to some extent. Markets and the spectrum of hybrid contractual arrangements such as alliances, partnerships, networks, franchising, and subcontracting practices are not spontaneous, improvised systems. They combine aspects of impersonal exchange and formal organizations, just as organizations combine internal markets, decentralization, and command. But the most important obstacle to the research program on a unified theory of organization is the lack of consensus on modeling-bounded rationality of actors whose decisions and choices shape the form and function of the organization.

Cross-References

- [Governance](#)
- [Incomplete Contracts](#)
- [Institutional Economics](#)

References

- Aghion P, Tirole J (1997) Formal and real authority in organizations. *J Polit Econ* 105:1–29
- Alchian A, Demsetz H (1972) Production, information costs and economic organization. *Am Econ Rev* 62:777–795
- Bac M (1996) Corruption, supervision and the structure of hierarchies. *J Law Econ Organ* 12:277–298
- Bac M (2001) Corruption, connections and transparency: does a better screen imply a better scene? *Public Choice* 107:87–96
- Blair M, Stout L (1999) A team production theory of corporate law. *V Law Rev* 85:247–328
- Bolton P, Farrell J (1990) Decentralization, duplication, and delay. *J Polit Econ* 98:803–826
- Chiappori PA, Salanié B (2003) Testing contract theory: a survey of some recent work. In: Dewatripont M, Hansen L, Turnovsky S (eds) *Advances in economics and econometrics*, vol 1. Cambridge University Press, New York
- Coase R (1937) The nature of the firm. *Economica* 4:386–405
- Demsetz H, Lehn K (1985) The structure of corporate ownership: causes and consequences. *J Polit Econ* 96:1155–1177
- Freixas X, Guesnerie R, Tirole J (1985) Planning under incomplete information and the Ratchet effect. *Rev Econ Stud* 52:173–191
- Grossman S, Hart O (1983) An analysis of the principal-agent problem. *Econometrica* 51:7–45
- Grossman S, Hart O (1986) The costs and benefits of ownership: a theory of lateral and vertical integration. *J Polit Econ* 94:691–719
- Holmstrom B (1999) Managerial incentive problems: a dynamic perspective. *Rev Econ Stud* 66:324–340
- Holmstrom B, Milgrom P (1991) Multi-task principal-agent analysis: incentive contracts, asset ownership and job design. *J Law Econ Organ* 7:24–52
- Milgrom P, Roberts J (1988) An economic approach to influence activities in organizations. *Am J Sociol* 94: S154–S179
- Posner RA (1999) Employment discrimination: age discrimination and sexual harassment. *Int Rev Law Econ* 19:421–446
- Qian Y, Roland G, Xu C (2006) Coordination and experimentation in M-form and U-form organizations. *J Polit Econ* 114:366–402
- Radner R (1992) Hierarchy: the economics of managing. *J Econ Lit* 30:1382–1415
- Sah RK, Stiglitz J (1986) The architecture of economic systems: hierarchies and polyarchies. *Am Econ Rev* 74:716–727
- Tirole J (1986) Hierarchies and bureaucracies: on the role of collusion in organizations. *J Law Econ Organ* 2:181–214
- Tirole J (1994) The internal organization of government. *Oxf Econ Pap* 46:1–29
- Williamson O (1985) *The economic institutions of capitalism*. Free Press, New York
- Wilson JQ (1989) *Bureaucracy: what government agencies do and why they do it*. Basic Books, New York

Organizational Liability

Klaus Heine¹ and Kateryna Grabovets²

¹Rotterdam Institute of Law and Economics,
Erasmus School of Law, Erasmus University
Rotterdam, Rotterdam, The Netherlands

²Rotterdam Business School, Rotterdam
University of Applied Sciences, Rotterdam,
The Netherlands

Abstract

Companies are acknowledged to be distinct legal persons that bear liability both for organizational failures and misconduct of their agents. Civil liability has long been imposed on companies and organizations. Corporate criminal liability has rapidly been expanded in recent years. Following the common law jurisdictions, in which notions of corporate criminal liability were introduced already in the early twentieth century, many civil law countries also recognize the possibility of holding companies criminally liable. However, distinctions in the use and theoretical underpinning of organizational liability remain across countries.

Synonyms

[Company liability](#); [Corporate liability](#)

Corporate Liability Around the World

Laws that render corporations liable have been strengthened after several high-profile corporate scandals, such as accounting crimes and financial fraud scandals (*Enron*, *WorldCom*, *Tyco*, *Parmalat*), the bribery of foreign officials (*Siemens*, *Samsung*), and gross negligence that caused the explosion of the oil rig (*British Petroleum*). Corporations can be held liable for wrongdoings, being subject to civil sanctions and criminal charges. Corporate criminal liability is currently applied in all common law countries

(inter alia, the USA, the UK, Ireland, Canada, and Australia) and in many civil law countries, including the Member States of the European Union; China; some Middle East countries, including Jordan, Syria, and Lebanon; and some post-Soviet countries, such as Baltic states, Moldova, Ukraine, Georgia, and Kazakhstan).

Common Law Countries

The USA

Until the early 1900s, corporations could not be held criminally liable for mens rea crimes. To constitute a mens rea crime, a guilty act (*actus reus*) has to be accompanied by a guilty state of mind or mens rea (Fischel and Sykes 1996, p. 320). Since corporations lack souls that are able to form a culpable intent and bodies to be punished, criminal liability was not imposed on corporations initially (Coffee 1981). But it has long become a common practice to attribute individuals' intents to corporations and make corporations criminally liable (Laufer 2006; Khanna 1996; Stessens 1994). Corporate liability is theoretically based on the legal fiction that corporations are persons distinct from their shareholders, directors, and employees (DiMento and Geis 2005, p. 160; Laufer 2006). Corporations can incur liability based on different grounds.

Vicarious Liability

The common law doctrine of respondeat superior, initially originated in tort law, forms a basis for vicarious liability, which implies that companies are strictly liable for the negligent actions of their agents acting within the scope of employment or authority and with intent to benefit the corporation (Wilkinson 2003; Laufer 1999).

From the perspective of deterrence, vicarious liability is expected to be most effective when a company can successfully monitor and control its agents (Sykes 2012, p. 2164). One of the reasons that justifies vicarious liability is that it mitigates the judgment proof problem. This problem implies that individual defenders have limited wealth and therefore may be unable to fully compensate damages. Compared to individuals,

companies possess larger assets (“deep pockets”) which can be used for paying sanctions and plaintiff compensation. Another justification of vicarious liability is that it prevents companies from inefficient excessive expansion of their activities (Sykes 1984, p. 1246).

Vicarious liability is applied, among others, in the Foreign Corrupt Practices Act (FCPA), enacted in 1977, and in the Alien Tort Statute (ATS), enacted in 1789. These acts are applicable to both domestic and foreign corporations.

Enterprise Liability

According to the theory of enterprise liability, advanced by *Fleming James* and *Friedrich Kessler* in the 1940–1950s, companies should internalize the costs associated with their activities, including the direct and indirect costs of injuries from corporate activities (Witt 2003, p. 39; Keating 2001). Enterprise liability is justified by the presumption that companies are in the best position to control safety and that they can best spread (financially diversify) the costs of injuries caused by defective products or services (Witt 2003, pp. 2, 39; Priest 1985, p. 492). Enterprise liability (firm-level liability) is widely applied in product liability and in ultrahazardous industries.

Compliance-Based Approach

Organizational criminal liability can be imposed based on the duty-based rules (fault-based or negligence rules). The duty-based liability rules foresee a mitigation of liability if corporate compliance programs, aimed at the prevention of wrongdoing, have properly been adopted (Arlen and Kraakman 1997). According to the law and economics literature, negligence liability is generally less efficient than strict liability because under the former tortfeasors do not fully internalize the harm inflicted by wrongdoings. This leads to a too high activity level of tortfeasors, which results in too many wrongdoings (Shavell 1980; Fischel and Sykes 1996, p. 328). The elements of duty-based liability are, however, increasingly applied by means of the compliance-based approach (Moot 2008). For example, the Federal Sentencing Guidelines take into account

corporate compliance programs in imposing liability on companies (Krawiec 2005). The promulgation of the compliance-based approach denotes a shift from respondeat superior liability to duty-based liability of corporate entities (Laufer 1999; Krawiec 2005). The enforcement of effective corporate compliance programs aims at the early prevention of misconduct in organizations, higher compliance efforts, and the creation of a positive law-abiding corporate ethos.

Long before the adoption of the guidelines, corporate compliance programs have been essential in the enforcement of antitrust law, securities law, and environmental law. Ethics codes and voluntary reporting of corporate misconduct, as part of compliance efforts, have also been promoted in the organizations of the defense industry (Walsh and Prych 1995).

The UK

Two high-profile statutes are specifically related to corporate liability in the UK – the Corporate Manslaughter and Corporate Homicide Act of 2007 (CMCH Act) and the Bribery Act of 2010. The Bribery Act foresees corporate liability for bribery. It can be imposed both on domestic and foreign companies, which conduct at least part of their business in the UK. The entity can defend itself by proving that it took adequate precautions and implemented adequate procedures to prevent misconduct.

The CMCH Act imposes homicide criminal liability on a corporate body for a gross breach of a duty of care, which caused a work-related death. Thereby the CMCH Act places an emphasis on the organizational and management failures, and it removed the identification doctrine, which was an obstacle for imposing criminal liability for manslaughter on a corporate body. According to the identification doctrine (the alter ego principle), a guilty individual (a “directing mind”) who occupies a senior position within a corporation has to be determined to attribute liability to the corporation. Unlike vicarious liability, the identification doctrine presumes that certain managers and employees act as the company itself, rather than on its behalf (Wells 1999, p. 120).

Civil Law Countries

Corporate criminal sanctions are applied in most civil law jurisdictions. Starting from 1990s many countries, which had only corporate civil liability before, actively adopted corporate criminal liability (Norway in 1991, Iceland in 1993, France in 1994, Finland and Slovenia in 1995, Belgium in 1999, Estonia and Hungary in 2001, Malta in 2002, Switzerland, Croatia, Poland and Lithuania in 2003, Slovakia and Romania in 2004, Austria 2006, Luxembourg and Spain in 2010, Czech Republic in 2011; the legislative initiative on corporate criminal liability has been introduced in Russia in 2015). Italy, Sweden, and Bulgaria have quasi-criminal corporate liability (Engelhart 2014, p. 56).

Germany

German law recognizes corporate liability imposed for wrongdoings of company representatives or employees. The German Penal Code (Strafgesetzbuch or StGB) follows the Roman rule *societas delinquere non potest*, according to which a legal entity cannot commit a crime. Criminal liability is applicable only to natural persons. Financial sanctions can nevertheless be imposed on legal entities for wrongdoings committed by corporate representatives or employees on behalf of the company (Weigend 2008; Böse 2011). These sanctions ensure that companies do not derive any benefits from the committed offenses, and they can be imposed irrespective of any criminal sanctions applied to natural persons with regard to the same offense. Corporate sanctions are justified by the standpoint that the failure to provide proper organization and supervision of corporate representatives and employees constitutes corporate guilt (Böse 2011, p. 231). The important role of organizational negligence in German law contrasts with both the no-fault enterprise liability approach, in which fault is irrelevant, and the doctrine of vicarious liability, according to which individual negligence is automatically attributed to a corporate entity (Kyriakakis 2009, p. 344).

In the German Civil Code (*Bürgerliches Gesetzbuch* or BGB), organizational negligence is interpreted in the principal-agent context. The

BGB (§823) specifies that a principal (“a person who uses another person to perform a task”) “is liable to make compensation for the damage that the other unlawfully inflicts on a third party when carrying out the task.”

France

Similar to Germany, the principle of *societas delinquere non potest* guided for a long time the approach toward corporate liability in France. Only the Penal Code (*Code pénal*) of 1992 (effective since 1994) introduced corporate criminal liability into the French law, which marked a significant shift from the idea that corporations cannot be criminally liable, which traces back to the Napoleonic Code (Coffée 1999, p. 23). Initially limited to a number of specific offenses, corporate criminal liability has been extended to all offenses and all legal persons, except from the state and local public authorities (Tricot 2014). To impose criminal liability on a legal entity, an offense has to be committed by the entity’s organs or representatives on behalf of the entity (Arts. 121–122 Penal Code).

Italy

Similar to Germany and France, the principle of *societas delinquere non potest* was dominant in the Italian law for a long time. Article 27(1) of the Italian Constitution states that criminal liability is personal, which implies that criminal law applies only to natural persons. Some European and OECD conventions, however, require the states which comply with these conventions to enact corporate liability. This fact induced Italy to enact Decree no. 231 (*Decreto Legislativo no. 231*) in 2001, which introduced administrative liability of corporate entities for crimes committed by their employees.

To attribute a criminal offense to a corporation, the offense should be committed in the interest or to the advantage of the corporation (Sant’Orsola and Giampaolo 2011). Article 6 of Decree no. 231 sets out the possibility for organizational defense, allowing a company to avoid liability if it “adopted and effectively implemented appropriate organisational and management models.”

The Netherlands

Already in 1950, the Dutch law acknowledged that economic offenses can be committed not only by individuals but also by corporate bodies (the Economic Offences Act, *Wet op de economische delicten*). In 1976 criminal liability of public and private corporate bodies was introduced in the Dutch Penal Code (*Wetboek van Strafrecht*), Art. 51. Criminal liability can be imposed on a corporate body if a wrongful act or omission was committed by a corporate employee or agent within the normal course of business, given that a company benefited from the wrongful conduct and accepted such conduct or failed to take adequate precautions to prevent it (Keulen and Gritter 2011). Thus corporate liability is based on the negligence standard. To attribute liability to the corporation, it is sufficient to prove that management or multiple individuals within the corporation possessed knowledge about the criminal activity or the risks thereof, but adequate measures toward crime prevention were not taken. Collective knowledge (the *mens rea* of different individuals) can be aggregated to establish organizational fault (Coffee 1999, p. 10; Stessens 1994, p. 512). This is a distinctive feature compared to the British identification theory (Coffee 1999, p. 22).

Criminal Liability Imposed by International Conventions

Apart from national laws, corporate liability is also envisaged by international conventions. For example, the OECD Convention on Combating Bribery of Foreign Public Officials in International Business Transactions (1997), which came into effect in 1999, prescribes the imposition of criminal liability on legal persons for bribery and corrupt business practices across borders. The signatories of the Convention are obliged to enact domestic legislation conforming to the Convention's standards.

Some other international conventions, such as the International Convention for the Suppression of the Financing of Terrorism (1999), the UN Convention Against Transnational Organized Crime (2000), and UN Convention Against Corruption (2003), also foresee corporate liability.

Moreover, International Standards on Combating Money Laundering and the Financing of Terrorism and Proliferation are an important soft law instrument that contains provisions on corporate criminal liability.

References

- Arlen J, Kraakman R (1997) Controlling corporate misconduct: an analysis of corporate liability regimes. *N Y Univ Law Rev* 72:687–779
- Böse M (2011) Corporate criminal liability in Germany. In: Pieth M, Ivory R (eds) *Corporate criminal liability: emergence, convergence and risk*. Springer, New York, pp 227–254
- Coffee JC Jr (1981) “No soul to damn: No body to kick”: an unscandalized inquiry into the problem of corporate punishment. *Mich Law Rev* 79:386–459
- Coffee JC Jr (1999) Corporate criminal liability: an introduction and comparative survey. In: Eser A, Heine G, Huber B (eds) *Criminal responsibility of legal and collective entities*. Edition iuscrim, Freiburg im Breisgau, pp 9–37
- DiMento JFC, Geis G (2005) Corporate criminal liability in the United States. In: Tully S (ed) *Research handbook on corporate legal responsibility*. Edward Elgar Publishing, Northampton, pp 159–176
- Engelhart M (2014) Corporate criminal liability from a comparative perspective. In: Brodowski D, Espinoza de los Monteros de la Parra M, Tiedemann K, Vogel J (eds) *Regulating corporate criminal liability*. Springer, London, pp 53–76
- Fischel DR, Sykes AO (1996) Corporate crime. *J Leg Stud* 25:319–349
- Keating GC (2001) The theory of enterprise liability and common law strict liability. *Vanderbilt Law Rev* 54:1285–1335
- Keulen BF, Gritter E (2011) Corporate criminal liability in the Netherlands. In: Pieth M, Ivory R (eds) *Corporate criminal liability: emergence, convergence and risk*. Springer, New York, pp 177–191
- Khanna VS (1996) Corporate criminal liability: what purpose does it serve? *Harv Law Rev* 109: 1477–1534
- Krawiec KD (2005) Organizational misconduct: beyond the principal-agent model. *Fla State Univ Law Rev* 32:571–615
- Kyriakakis J (2009) Corporate criminal liability and the ICC Statute: the comparative law challenge. *Neth Int Law Rev* 56:333–366
- Laufer WS (1999) Corporate liability, risk shifting, and the paradox of compliance. *Vanderbilt Law Rev* 52: 1343–1420
- Laufer WS (2006) *Corporate bodies and guilty minds: the failure of corporate criminal liability*. The University of Chicago Press, Chicago

- Moot JS (2008) Compliance programs, penalty mitigation and the FERC. *Energy Law J* 29:547–575
- Priest GL (1985) The invention of enterprise liability: a critical history of the intellectual foundations of modern tort law. *J Leg Stud* 14:461–527
- Sant’Orsola FC, Giampaolo S (2011) Liability of entities in Italy: was it not *societas delinquere non potest*? *New J Eur Crim Law* 2:59–74
- Shavell S (1980) Strict liability versus negligence. *J Leg Stud* 9:1–25
- Stessens G (1994) Corporate criminal liability: a comparative perspective. *Int Comp Law Q* 43:493–520
- Sykes AO (1984) The economics of vicarious liability. *The Yale Law J* 93:1231–1280
- Sykes AO (2012) Corporate liability for extraterritorial torts under the Alien Tort Statute and beyond: an economic analysis. *The Georgetown Law J* 100: 2161–2209
- Tricot J (2014) Corporate liability and compliance programs in France. In: Manacorda S, Centonze F, Forti G (eds) *Preventing corporate corruption: the anti-bribery compliance model*. Springer, London, pp 477–490
- Walsh CJ, Pyrich A (1995) Corporate compliance programs as a defense to criminal liability: can a corporation save its soul? *Rutgers Law Rev* 47:605–691
- Weigend T (2008) *Societas delinquere non potest?* A German perspective. *J Int Crim Justice* 6:927–945
- Wells C (1999) A new offence of corporate killing – the English Law Commission’s proposals. In: Eser A, Heine G, Huber B (eds) *Criminal responsibility of legal and collective entities*. Edition iuscrim, Freiburg im Breisgau, pp 119–128
- Wilkinson M (2003) Corporate criminal liability – the move towards recognizing genuine corporate fault. *Canterbury Law Rev* 9:142–178
- Witt JF (2003) Speedy Fred Taylor and the ironies of enterprise liability. *Columbia Law Rev* 103:1–49

Cases

- Escola v. Coca-Cola Bottling Co.* 24 Cal. 2d 453, 150 P.2d 436 (1944)
- Greenman v. Yuba Power Prods., Inc.* 59 Cal. 2d 57, 27 Cal. Rptr. 697, 377 P.2d 897 (1963)
- Henningsen v. Bloomfield Motors, Inc.* 32 N.J. 358, 161 A.2d 69 (1960)

Panel Data Analysis

Aitor Ciarreta, María Paz Espinosa and
Ainhoa Zarraga
University of the Basque Country, UPV/EHU,
Bilbao, Spain

Abstract

A panel data set is one that follows a given sample of units over time. Thus, panel data analysis refers to econometric tools that deal with the estimation of relationships that combine time series and cross-sectional data. Appropriate estimation methods are discussed depending on the characteristics of the data.

Denote by y_{it} an observation of the dependent variable for unit i at time t and x_{it} the set of K -independent variables observed for unit i at time t . Consider a linear regression model:

$$y_{it} = \alpha_{it} + \beta'_{it}x_{it} + u_{it}, \quad i = 1, \dots, N, \quad (1) \\ t = 1, \dots, T,$$

where the error term u_{it} is independently and identically distributed over i and t , and α_{it} and β'_{it} are (1×1) and $(K \times 1)$ vectors of parameters to be estimated.

This model cannot be estimated because the degree of freedom, NT , is less than the number of parameters to be estimated, $NT(K + 1)$.

Therefore, a structure needs to be imposed. Assuming constant parameters over time, the basic initial test is the covariance test for homogeneity across units, where three hypotheses can be made (Greene 2008; Hsiao 2003):

H₁: Both intercept and slope coefficients are the same: $\alpha_{it} = \alpha$ and $\beta'_{it} = \beta'$.

When **H₁** is rejected, two different types of heterogeneity across units can be considered:

H₂: Slope coefficients are the same and intercept coefficients are not: $\alpha_{it} = \alpha_i$ and $\beta'_{it} = \beta'$.

H₃: Intercept coefficients are the same and slope coefficients are not: $\alpha_{it} = \alpha$ and $\beta'_{it} = \beta'_i$.

Behind the implied models underlies the assumption that the effects of the omitted variables are of two types: individual time-invariant and period individual-invariant. The simplest representation is to introduce dummy variables for specific cross-sectional units that stay constant over time and time-specific effects constant across units. There are several methods to estimate Eq. 1 depending on the structure of the model. First, assume lagged dependent variable does not enter in the equation as a regressor. Then, the fixed-effects (FE) and the random-effects (RE) models are commonly used.

The FE model removes the effect of time-invariant characteristics from the explanatory variables which are unique to the unit and not correlated with other individual characteristics. Thus,

we can assess the net effect on prediction. The equation to estimate is

$$y_{it} = \alpha_i + \beta' x_{it} + u_{it} \quad i = 1, \dots, N, \quad (2)$$

$$t = 1, \dots, T.$$

The least-squares dummy variable estimator, also called the within-group estimator, is obtained applying OLS to Eq. 2 in which variables have been previously transformed to subtract the corresponding time series means. This estimator is unbiased and consistent when either N or T or both tend to infinity.

The RE model, on the contrary, assumes that the variation across units is random and uncorrelated with the independent variables included in the model:

$$y_{it} = \alpha + \beta' x_{it} + v_i + u_{it} \quad i = 1, \dots, N, \quad (3)$$

$$t = 1, \dots, T,$$

where v_i is the between-individual error term. It is assumed that v_i is not correlated with the explanatory variables. If you have any reason to believe that differences across units have some influence on the dependent variable, then random effects are more appropriate. An advantage of RE is that you can include time-invariant variables, whereas in the FE model, these variables are absorbed by the intercept.

In Eq. 3, OLS estimator is inefficient because the error term is correlated. In this case, GLS gives an efficient estimator.

The choice between FE and RE can be performed using the Hausman test where the null hypothesis is that the best model is RE.

There are other tests the analyst has to perform before choosing one model or the other. If the FE model is selected, then test whether time-fixed-specific effects, λ_t , are needed. If they are needed, then the model to estimate is

$$y_{it} = \alpha_i + \lambda_t + \beta' x_{it} + u_{it} \quad i = 1, \dots, N, \quad (4)$$

$$t = 1, \dots, T.$$

The basic model can be extended in several directions. Assume that T is large enough, then

the model is dynamic in nature and lagged dependent variables may be included:

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \lambda_t + \beta' x_{it} + u_{it} \quad \text{where}$$

$$i = 1, \dots, N \text{ and } t = 1, \dots, T. \quad (5)$$

Finally, another extension is to assume variable coefficient models:

$$y_{it} = \sum_{k=1}^K \beta'_{kit} x_{kit} + u_{it} \quad \text{where}$$

$$i = 1, \dots, N \text{ and } t = 1, \dots, T. \quad (6)$$

For advanced reading see Baltagi (2008).

References

- Baltagi BH (2008) *Econometric analysis of panel data*. Wiley, New York
- Greene WH (2008) *Econometric analysis*, 6th edn. Prentice Hall, Upper Saddle River
- Hsiao C (2003) *Analysis of panel data*, *Econometric society monographs* no 34., 2nd edn Cambridge University Press, Cambridge/New York

Parallel Economy

► [Informal Sector](#)

Parenthood

► [Adoption](#)

Party Competition: Definitions, Sources, and Economic Effects

► [Political Competition](#)

Passed Money

► Counterfeit Money

Passive Minority Interests

Nikolaos E. Zevgolis and Panagiotis N. Fotis
Hellenic Competition Commission, Athens,
Greece

Abstract

Passive minority interests consist of various types of non-controlling interests in competitors, such as minority shareholdings and interlocking directorships. Loans and other financial products involving competitors may also play a crucial role in creating mutual interactions among the parties. A minority interest is the portion of a consolidated entity that is not owned by the consolidating entity. A consolidated entity may be formed through management/shareholding control and also depends on the prevailing accounting/regulatory environment.

passive minority shareholder may be in a position to receive direct information regarding the victim firm's main operations. In such a case, despite the fact that the practical ability of the owner of the passive minority interest to have access to this kind of information is not enough to enforce him influencing the strategic behavior of the victim firm, it is adequate to provide him with a sort of control over the victim firm (OFT 2010).

Acquisitions of non-controlling minority interests in a rival may also reduce the quality of the product. However, economically speaking, the said links are less likely to trigger unilateral and coordinated effects than full acquisitions or acquisitions of minority shareholdings that influence the strategic behavior of the target firm.

Passive minority interests consist of various types of non-controlling interests in competitors, such as minority shareholdings and interlocking directorships. Loans and other financial products involving competitors may also play a crucial role in creating mutual interactions among the parties. Generally speaking, a minority interest is the portion of a consolidated entity that is not owned by the consolidating entity. A consolidated entity may be formed through management/shareholding control and also depends on the prevailing accounting/regulatory environment.

Synonyms

Non-controlling minority interests

Introduction

Passive or non-controlling minority interests (minority shareholdings and interlocking directorships) among competitors of oligopolistic markets are links that tend to raise the price and reduce the quantities sold in the market, even in cases where a cartel, tacit or explicit, is not detected. Such links may force competitors to compete less vigorously and adopt behavior more conducive to joint profit maximization by facilitating price and other strategic information sharing among them. This may be happening because there are cases where a

Minority Shareholdings

Minority shareholdings among firms refer to situations where shareholders hold less than 50% of the voting rights of other firms' equity capital. That is, they are portions of a consolidated entity which do not confer control in the legal sense to their owners and are therefore not notifiable under the worldwide merger notification systems. Minority shareholdings are also called structural links (SWD 2014a, Annex 1 and 6).

Active or Passive Minority Shareholdings

Minority shareholdings may be either *active* (controlling) or *passive* (non-controlling) in nature. Active minority shareholdings refer to minority interests where their owner, a person or a firm, may exercise some form of control over the

victim firm(s) (i.e., the issuing firm of the asset or claim for which a minority interest is established). Passive minority shareholdings refer to situations where a minority shareholder cannot be represented in the decision-making bodies of the issuing firm. They constitute purely passive financial interests. However, there are circumstances where a passive minority shareholder may receive direct information about an issuing firm's main operations. Even though the practical ability of the passive minority shareholder to have access to this kind of information does not provide him with the potential to influence the strategic behavior of the issuing firm, it does provide a sort of control over this firm.

Pre-existing Minority Shareholdings

Pre-existing minority shareholdings exists in the context of a notified transaction and therefore may be analyzed regarding its anti- or/and pro-competitive effects within the notified transaction. For instance, a national competition authority or the Commission could order the divestiture of the pre-existing minority shareholding in the case that it is found to be anticompetitive. Where a minority shareholding is acquired after the examination of a notified transaction, the competition authority lacks the competence to review this acquisition.

Horizontal or Non-horizontal Minority Shareholdings

Horizontal minority shareholdings are minority shareholdings in horizontal competitors, while *non-horizontal* minority shareholdings are interests either between firms operating at different levels of the supply chain (*vertical* minority shareholdings) or between firms that are neither purely horizontal nor purely vertical (*conglomerate* minority shareholdings).

Direct or Indirect Minority Shareholdings

When a firm holds a *direct* minority shareholding in another firm it means that it holds it directly without the intervention of a third firm. However, an *indirect* minority shareholding exists when a firm holds a minority shareholding in another firm via the intervention of a third firm.

One-Way or Reciprocal Minority Shareholdings

If both firms hold shares in one another equity capital, then the minority shareholding is called *reciprocal* minority shareholding. The opposite of reciprocal minority shareholding is the *one-way* minority shareholding.

Interlocking Directorships

Interlocking directorships refer to situations where horizontal and nonhorizontal (e.g., vertical) competitors share on boards of directors one or more directors, top executives, nonexecutives, managers, and close relatives, e.g., spouses and parents. Competitive concerns may be stronger in cases where interlocking directorships involve managers rather than directors since the former may more easily understand the day-to-day organization of the issuing firm than the latter and therefore enjoy more decisive influence over its strategic behavior in the interest of their employers.

Types of Interlocking Directorships

As in the case of minority shareholdings, there exist various forms of interlocking directorships such as active or passive, horizontal or non-horizontal, direct or indirect, and one-way or reciprocal interlocking directorships. Except from these basic types, there exist other types of interlocking directorships between firms. For example, in some circumstances firms' officers may hold a seat on the board of directors of competitors. This type of directorship is called *management interlock* (Arreda and Turner 1978). However, there are competitors that share common officers without involving directorships. This type of interlock seems not to raise competition concerns and it is called a *manager to manager interlock* (O'Brien and Salop 2000).

Passive Minority Interests and Competition Law

In a Nutshell

As a general rule, active or controlling minority shareholdings fall under the scope of merger

notification systems, while passive or non-controlling minority shareholdings are not subject to the majority of control systems found worldwide. Against the backdrop of the Council Regulation (EC) No. 139/2004 (*The EC Merger Regulation*), the Commission cannot examine minority shareholdings in stand-alone investigations. Nevertheless, in some national merger laws, there are provisions regarding the acquisitions of minority shareholdings in competitors (see, for instance, Germany, United Kingdom, Austria, and Lithuania), but the majority of them do not incorporate such provisions in their merger notification systems. In that case, the regulation of passive minority shareholdings falls under the mechanism of antitrust law. Preexisting minority shareholdings constitute exceptions to this rule.

Passive Minority Interests and Mergers and Acquisitions

Active along with pre-existing minority shareholdings are the only forms of minority interests that fall under the scope of the Merger Regulation 139/2004. More specifically, active minority shareholdings fall directly under the scope of this Regulation; preexisting minority shareholdings fall indirectly under the scope of the same Regulation. A few merger cases in EU (see, *inter alia*, IV/M.042 *Alcatel/Telettra*, IV/M.113 *Courtaulds/SNIA*, IV/M.833 *Coca-Cola/Carlsberg*, IV/M.890 *Blokker/Toys “R” Us On 23*, IV/M.1082 *Allianz/AGF*, IV/M.1378 *Hoechst/Rhône-Poulenc*, IV/M.1383 *Exxon/Mobil*, IV/M.1453 *AXA/GRE*, IV/M.1673 *VEBA/VIAG*, IV/M.1940 *Siemens/Framatome/Cogéma*, COMP/M.1980 *Volvo/Renault*, COMP/M.2050 *Vivendi/Seagram*, COMP/M.2567 *Nordbanken/Postgirot*, COMP/M.3547 *Banco Santander/Abbey National Transactions*, COMP/M.3653 *Siemens/VA Tech*, IV/M.3696 *E.ON/MOL*, COMP/M.4150 *Abbott/Guidant*, COMP/M.4153 *Toshiba/Westinghouse*, COMP/M.4439 *Ryanair/Aer Lingus*, COMP/M.5096 *RCA/MAV Cargo*, COMP/M.5406 *IPIC/MAN Ferrostaal*, COMP/M.6541 *Glencore/Xstrata*, COMP/M.6662 *Andritz/Schuler*) constitute examples where a party to a transaction has pre-existing structural links to competitors or other firms and where the

parties quite often offered to divest these minority shareholdings with the intention of remedying the competition concerns raised by the transactions.

An active minority shareholding exists if the major prerequisites of the EU Merger Regulation are satisfied. In particular, the merger notification system in the EU is based on three pillars: (i) the notion of concentration, (ii) the existence of acquisition of control, and (iii) the dimension of the notified transaction. According to the Commission Consolidated Jurisdictional Notice under Council Regulation . . . (2008), para 7, “a concentration only covers operations where a change of control in the undertakings concerned occurs on a lasting basis.” So, “the concept of concentration is intended to relate to operations which bring about a lasting change in the structure of the market.” The notion of concentrations also applies to joint ventures, which perform “on a lasting basis all the functions of an autonomous economic entity.”

In terms of European law the concept of concentration cannot be extended to cases in which control has not been obtained and the shareholding at issue does not, as such, confer the power of exercising decisive influence on the other undertaking. However, the European Commission seems to support the need to extend Regulation 139/2004 to the acquisition of non-controlling minority shareholdings (see SWD 2014a, 221, para 62).

Passive Minority Interests and Anticompetitive Practices

In General

Articles 101 and/or 102 of the Treaty on the Functioning of the European Union (TFEU) may apply to passive minority interests in situations where there is evidence of an anticompetitive agreement or concerted practice among the investigated firms or the firms that are engaged in the acquisition of non-controlling stakes and/or one or more firms have a dominant position. Nevertheless, according to the Commission Staff Working Document (SWD 2014a, 221), “the Commission’s ability to use Article 101 and Article 102 TFEU to intervene against anti-competitive minority shareholdings may be limited.”

More specifically, in its *White Paper* and the accompanying Commission Staff Working Paper of July 2014, the Commission states that an acquisition of a minority shareholding may not constitute an agreement which by object or effect restricts the effective competition. As a consequence, it may not be possible to intervene in cases that involve potentially anticompetitive minority shareholdings. Furthermore, regarding Article 102 TFEU, the Commission states the same argument unless the acquirer or the issuing firm or both of them have a dominant position in the markets under scrutiny (see White Paper (2014b) para 40; SWD (2013) 239, Annex II, p 6, and SWD (2014a) 221, para 62).

Application of Article 101 TFEU

In some cases, an agreement which by object or effect restricts the effective competition and violates Article 101 TFEU does not exist. For instance, acquisitions of minority interests via a stock exchange may not contain an agreement among the firms of the sellers and the buyers. In this case, a memorandum among shareholders and/or particular firms' articles covering their relationship may be viewed as an agreement under Article 101 TFEU. Generally speaking, stock market acquisitions of minority shareholdings involve an agreement between the firm, who acts as a purchaser, and the shareholder(s) of a firm, who act as the seller(s) or the issuing firm in the transaction. In this transaction, the difficulty which arises for assessing the potential anticompetitive effects of such an acquisition is that the firm is not (at least typically) a participant in the transaction. Therefore, an agreement between the two firms does not exist.

On the other hand, an agreement between the two firms does exist if the corporate bodies of the two firms are involved directly in the transaction. Furthermore, Gilo et al. (2006) have proved that a controlling shareholder (whether a person or a parent corporation) can facilitate tacit collusion further by making a direct passive investment in rival firms. Such investment particularly facilitates collusion if the controller has a relatively small stake in his own firm.

Potential Application of Article 101 TFEU in

Combination with Application of Article 102 TFEU

The establishment of coordinated effects (see section "[Coordinated Effects of Passive Minority Interests](#)") as a consequence of an acquisition of passive minority interests implies that the firms engaged in the transaction coordinate their strategic behavior, i.e., they raise prices and harm effective competition. Coordination may occur by firms that prior to the acquisition did or did not coordinate their behavior. In the former case, a minority shareholding may make coordination more effective and easier.

An acquisition of a minority shareholding may serve as an instrument for influencing the commercial or strategic policy of a direct competitor (especially in the case where reciprocal minority shareholdings exist), or it may enable the acquiring firm to acquire control over the acquired firm at a later stage. In that case, even though the acquisition of minority shareholding does not lead by itself to a restriction of competition for the purposes of Article 101 TFEU, i.e., it has neither the object nor the effect of distorting competition by creating the premises for cooperation and/or exchange of commercially sensitive information among the engaged firms, its potential impact on competition in the long run can be seen as an agreement which violates Article 101 TFEU.

As a consequence, such structural links may eliminate the incentives of the competitors to compete with each other in the market(s) in which they carry out business (Levy 2014). As well as potential commercial cooperation among firms engaged in a minority shareholding transaction, there are other factors that may serve in this direction. For instance, providing that the market under scrutiny is an oligopolistic market, such factors are barriers to entry and the ability of the acquiring firm to achieve effective control (legal or de facto) over the commercial and strategic policy of the acquired firm.

In particular, the level of concentration in the market under scrutiny plays a critical role in the existence or not of coordinated behavior. The rule is that the higher the level of concentration, the lower the number of the effective firms in the

market and the higher the likelihood of cooperation among them. This depends crucially on the magnitude of the influence of the strategic behavior of the acquired firm, since the more the acquiring firm knows about the target firm, the easier it is to coordinate with it in order to enhance the profits of both firms. A concentrated market enhances coordination among its players without the need for an agreement that may violate Article 101 TFEU; the possibility of collective dominance – and perhaps abuse of it under Article 102 TFEU – in such a case is quite strong. However, two factors that independently contribute to such an enhancement must be considered: firstly, the existence of reciprocal and indirect minority shareholdings among the firms in the market; and secondly, the existence or not of a maverick firm. (The existence of a maverick undertaking would probably diminish the possibility of collective dominance.) More specifically, the existence of reciprocal and indirect minority shareholdings among the firms in the market enhances the ability of the acquiring firm to monitor the strategic behavior of the target firm and at the same time the strategic behavior of all the other firms that are connected with it via the existence of minority shareholdings.

Application of Article 102 TFEU

Article 102 TFEU applies only in the case where the engaged firms in the transaction under scrutiny hold a dominant position, either independently or collectively, in the concerned relevant market(s). Therefore, an acquisition of minority shareholding requires the existence of a dominant position between the parties that engaged in it. However, in theory it has been stated that an abuse of a dominant position requires that an acquisition of a minority shareholding may serve as an instrument for influencing the commercial or strategic policy of a competitor, especially in the case where there exist reciprocal minority shareholdings in the relevant market under scrutiny (Fotis and Zevgolīs 2016).

As a matter of fact, this was the judgment of the Court of Justice in the *Phillip Morris* case. Furthermore, in the *Gillette* case the firm abused its dominant position in the relevant market of

disposable razors by acquiring a minority shareholding over its direct competitor, Wilkinson Sword. Gillette had become the main shareholder and creditor of Eemland (the owner of Wilkinson Sword) and could have used its rights to prevent future concentration plans that Eemland would have had, and Gillette would not have approved of. Therefore, in the *Gillette* case the European Commission established an infringement of Article 102 TFEU since the transaction had altered the structure of the market under investigation.

The Economics of Passive Minority Interests

The majority of relevant research in the literature indicates that non-controlling minority interests decrease the level of competition in the markets by enhancing cooperation among rivals or by increasing the probability that a dominant firm will abuse its dominant position. In any case, passive minority interests decrease consumer welfare by increasing the product price and reducing its quantity.

The Anticompetitive Effects of Passive Minority Interests

Economic literature has shown that passive minority interests among rivals (*horizontal passive minority interests*) harm competition either unilaterally or via coordinated effects. The magnitude of these effects depends on the share of the issuing firm's profits to which the acquiring firm is actually entitled as a result of the non-controlling acquisition, and the ability of the acquiring firm materially to influence the issuing firm's strategic behavior. If passive minority interests are accompanied by interlocking directorships, then such interests may alter the competitive behavior of both acquiring and target firm.

The anticompetitive effects of *nonhorizontal passive minority interests* have also been examined in the economic literature. *Vertical passive minority interests* impose input (upward) or customer (downstream) exclusion. *Conglomerate passive minority interests* enhance anticompetitive

issues if the acquiring firm acquires a non-controlling minority stake in the issuing firm which is active in a market related to that of the acquiring firm. A conglomerate non-controlling minority interest may promote acquiring firms' sales of one product contingent on the sales of other products via *bundling*, *tying*, *exclusive dealing*, or *full-line forcing*.

Non-controlling minority interests may block entry both for potential entrants that are not active in the market or for active firms in the market that are interested in expanding their product range. Also, acquiring firm hinders another firm from getting an access to the capital stock of the issuing firm. For instance, in the Aer Lingus/Ryanair case Aer Lingus argued that Ryanair's non-controlling minority interest on its capital stock had a material impact on Aer Lingus's shares, by making them less attractive to potential buyers (*Case T-411/07, "The Ryanair Decision"* and *Case T-342*07 Ryanair v Commission*). However, it should not be ignored that the attractiveness of Aer Lingus on the stock market is not based solely on Ryanair's non-controlling minority interest (SWD 2014a, 221).

Furthermore, the acquiring firm has an enhanced incentive to deter entry by credibly committing to the potential entrant. If the non-controlling minority interest is accompanied by corporate rights, then the commitment is stronger that further prevents a potential entrant from entering to the market. However, if non-controlling reciprocal minority interests are prohibited (allowed), then entry will be deterred (promoted) if the relevant products are complements (substitutes) (Clayton and Jorgensen 2005).

Unilateral Effects of Passive Minority Interests

The anticompetitive effects of non-controlling minority interests are stronger on the horizontal than on vertical level. On the former, they change firms' strategic behavior in favor of coordination among them. If the acquiring firm has an incentive to compete softly with its competitor, may raise its product price or restrict its output (see, *inter alia*, Reynolds and Snapp 1986; Bresnahan and Salop 1986; Farrell and Shapiro 1990; Shelegia and Spiegel 2012).

In a static environment, the acquiring firm abuse part of the issuing firm's diverted sales via its non-controlling minority stake. In this scenario, the degree of competition between the two firms and the amount of sales that diverted to the competitors are of high importance. The more the rivalry between firms, that is the case of close substitute products, the more the diverted sales that are captured by the issuing firm after an increase of acquiring firm's product. This is called *Diversion Ratio*. In duopolistic markets, the *Diversion Ratio*, in absolute terms, is the ratio of the cross-elasticity of demand for firm's A product when firm B raises its product price divided by the own-elasticity of demand for firm's B product multiplied by the ratio of unit sales by firm B divided by the unit sales by firm A (Fotis et al. 2017). Therefore, the acquiring firm earns profit, the magnitude of which depends on the non-controlling minority interest in the capital of the issuing firm.

Vertical passive minority interests harm competition by imposing input (upward) or customer (downstream) exclusion. If they are accompanied by corporate rights, then the acquiring firm gains the ability to exclude the issuing firm's competitors from its input and customer base. The more concentrated the markets are, the more the concerns for either type of exclusion. The acquiring firm's incentive to exclude competitors is higher through corporate rights than through solely financial interests. If the acquiring firm gains access to strategic data of the issuing firm, then the gains from exclusion are more than the gains accrued from a vertical merger. The acquiring firm internalizes the whole stream of profits that come from the other levels of supply chain and at the same time incurs only a small portion of the costs caused by the anticompetitive strategy (see, *inter alia*, Flath 1989; Greenlee and Raskovich 2006).

An example of vertical passive minority interest at the European Union (EU) level is the case *COMP/M.5406 IPIC/MAN Ferrostaal*. In this case, MAN Ferrostaal (the acquiring firm) held a non-controlling minority stake in Eurotecnica (the issuing firm), a provider of technology and engineering services. The EU Commission was

concerned that the acquiring firm would determine the technology licenses distribution by the issuing firm and thus excluded (potential or active) competitors from the market. The suggested remedy by the EU Commission in order to clear the case included the divestiture of the non-controlling minority interest held by MAN Ferrostaal in Eurotecnica (SWD 2014a, 221).

Coordinated Effects of Passive Minority Interests

Non-controlling minority interests also enable involved firms to coordinate their conduct. Through coordination the acquiring firm internalizes part of the issuing firm's profit. Moreover, the level of transparency is enhanced in the presence of corporate rights. On the one hand, in the case of one-way acquisition of information from the issuing to the acquiring firm (*unilateral passive minority interest*), the acquiring firm has the ability to monitor the commercial policy of the issuing firm. If, on the other hand, a reciprocal minority shareholding exists, then the level of transparency is enhanced more than the previous case (see, *inter alia*, Malueg 1992; Reitman 1994; Gilo et al. 2006; Brito et al. 2013).

A market becomes more transparent when *indirect non-controlling minority interests* are accompanied by corporate rights. Such interests enlarge the scope of coordination and enable the coordinating firms to detect possible deviations from the common target which aims at maximizing the monopoly profits (*IV/M.1673 V EBA/VIAG*).

Pro-competitive Passive Minority Interests and Efficiencies Gains

Non-controlling minority interests raise arguments in favor of efficiencies. Clayton and Jorgensen (2005) analyze reciprocal minority interests in a duopoly Cournot market, and they state that consumer surplus and firms' profits with complements products are enhanced more when the said interests are allowed rather than when they are prohibited. If the products are substitutes, then firms' profits are enhanced when passively acquired interests are prohibited rather than when they are allowed. On the contrary, consumer

surplus is enhanced when non-controlling minority interests are allowed.

Brito et al. (2014) argue that consumer surplus is enhanced by turning voting shares (shares with control rights) into nonvoting shares (shares with no control rights). Moreover, the sale of voting shares to a new large shareholder is better than the sale of voting shares to a series of small shareholders. The authors agree with the main results of O'Brien and Salop (2000), but also consider that a financial interest affects not only the incentives of the acquiring but also the incentives of the acquired firm (see also *COMP/M.2416 Tetra Laval/Sidel*).

Passive minority interests enable the transfer of technology and managerial skills and the creation of synergies among firms which positively influence firms' profits (Ono et al. 2004). Amundsen and Bergman (2002) argue that cost reduction through sales cooperation and learning, such as information about production process, are also, *inter alia*, motives which enable a firm to acquire non-controlling minority interests in their rivals' capital stock.

Cross-References

► [Cartels and Collusion](#)

References

- Amundsen ES, Bergman L (2002) Will cross-ownership re-establish market power in the Nordic power market? *Energy J* 23(2):73–95
- Arreda P, Turner DF (1978) *Antitrust law*. Little Brown and Company, Boston
- Bresnahan TF, Salop SC (1986) Quantifying the competitive effects of production joint ventures. *Int J Ind Organ* 4:155–175
- Brito D, Ribeiro R, Vasconcelos H (2013) Quantifying the coordinated effects of partial horizontal acquisitions. CERP discussion papers no. 9536
- Brito D, Ribeiro R, Vasconcelos H (2014) Measuring unilateral effects in partial horizontal acquisitions. *Int J Ind Organ* 33:22–36
- Clayton MJ, Jorgensen BN (2005) Optimal cross holding with externalities and strategic interactions. *J Bus* 78(4):1505–1522
- Commission Consolidated Jurisdictional Notice under Council Regulation (EC) No. 139/2004 on the control

- of concentrations between undertakings (2008) OJ C95/01. Accessed 10 June 2010
- Council Regulation (EC) No. 139/2004 of 20 January 2004 on the control of concentrations between undertakings (the EC Merger Regulation) OJ L 24, 29.01.2004, pp 1–22. Accessed 10 June 2010
- Farrell J, Shapiro C (1990) Asset ownership and market structure in oligopoly. *RAND J Econ* 21(2):275–292
- Flath D (1989) Vertical integration by means of shareholder interlocks. *Int J Ind Organ* 7:369–380
- Fotis P, Zevgolits N (2016) *The competitive effects of minority shareholdings legal and economic issues*. Hart Publishing, Bloomsbury
- Fotis P, Polemis M, Eleftheriou K (2017) Unilateral effects of partial acquisitions: consistent calculation of GUPPI under horizontal merger guidelines within the EU. *Econ E Polit Ind*. <https://doi.org/10.1007/s40812-016-0053-6>
- Gilo D, Moshe Y, Spiegel Y (2006) Partial cross ownership and tacit collusion. *RAND J Econ* 37(1):81–99
- Greenlee P, Raskovich A (2006) Partial vertical ownership. *Eur Econ Rev* 50:1017–1041
- Levy N (2014) Expanding EU merger control to non-controlling minority shareholdings: a sledgehammer to crack a nut? *CPI Antitrust Chron* 12:1–16
- Malueg AD (1992) Collusive behavior and partial ownership of rivals. *Int J Ind Organ* 10:27–34
- O'Brien DP, Salop S (2000) Competitive effects of partial ownership: financial interest and corporate control. *Antitrust Law J* 67(3):559–614
- OFT (2010) Minority interests in competitors. A research report prepared by DotEcon Ltd, OFT1218
- Ono H, Nakazato T, Davis C, Alley W (2004) Partial ownership arrangements in the Japanese automobile industry; 1990–2000. *J Appl Econ* 7(2):355–367
- Reitman D (1994) Partial ownership arrangements and the potential for collusion. *J Ind Econ* 42:313–322
- Reynolds RJ, Snapp BR (1986) The competitive effects of partial equity interests and joint ventures. *Int J Ind Organ* 4(2):141–153
- Shelegia S, Spiegel Y (2012) Bertrand competition when firms hold passive ownership stakes in one another. *Econ Lett* 114:136–138
- SWD (2013) 239 Final, commission staff working document (Towards more effective EU merger control), Annex II. Accessed 5 June 2010
- SWD (2014a) 221 Final, commission staff working document (Towards more effective EU merger control). Accessed 5 June 2010
- SWD (2014b) White Paper (Towards more effective EU merger control) 449 final. Accessed 3 June 2010
- COMP/M.2416 Tetra Laval/Sidel (2004) OJ L38, 1
- COMP/M.2567 Nordbanken/Postgirot (2001) OJ C 347
- COMP/M.3547 Banco Santander/Abbey National (2004) OJ C 255
- COMP/M.3653 Siemens/VA Tech 2006/899/EC, OJ L 353
- COMP/M.4150 Abbott/Guidant (2006) OJ C 256/10
- COMP/M.4153 Toshiba/Westinghouse (2006) OJ C 184/02
- COMP/M.4439 Ryanair/Aer Lingus (2007) OJ C 47/08 (decision of 20 December 2006, OJ C 10, 2007 (the Ryanair decision))
- COMP/M.5096 RCA/MAV Cargo (2008) OJ C 29
- COMP/M.5406 IPIC/MAN Ferrostaal (2009) OJ C 114
- COMP/M.6541 Glencore/Xstrata (2014) OJ C 109/1
- COMP/M.6662 Andritz/Schuler (2013) OJ C 14
- IV/M.1673 V EBA/VIAG (2000) OJ L 188
- IV/M.042 Alcatel/Telettra (1991) OJ L 122/48
- IV/M.113 Courtaulds/SNIA (1991) OJ C 333
- IV/M.833 Coca-Cola/Carlsberg (1997) OJ L 145
- IV/M.890 Blokker/Toys 'R' Us (1998) OJ L 316/1
- IV/M.1082 Allianz/AGF (1998) OJ C 246/4
- IV/M.1378 Hoechst/Rhône-Poulenc (1999) OJ C254/5 90
- IV/M.1383 Exxon/Mobil (1999) OJ 2004 L103/1
- IV/M.1453 AXA/GRE (1999) OJ C 030/6
- IV/M.1673 VEBA/VIAG (2000) OJ L 188
- IV/M.1940 Siemens/Framatome/Cogéma (2000) OJ L 289/8
- IV/M.3696 E.ON/MOL 2006/622/EC, OJ L 253
- Warner-Lambert/Gillette (1993) IV/33.440
- T-342/07 Ryanair v Commission (2010) ECR II-3457
- T-411/07 Aer Lingus v Commission (2010) ECR II-3691
- The Philip Morris case, 142 and 156/84 (1987) ECR 4487 (1988) 4 CMLR 24

Internet Addresses

- Commission Consolidated Jurisdictional Notice. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2008:095:0001:0048:EN:PDF>
- TFEU. <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012E%2FTXT>
- The EC Merger Regulation. <http://ec.europa.eu/competition/mergers/legislation/regulations.html>
- SWD (2013) 239 Final. http://ec.europa.eu/competition/consultations/2013_merger_control/merger_control_en.pdf
- SWD (2014) 221 Final. http://ec.europa.eu/competition/consultations/2014_merger_control/staff_working_document_en.pdf
- White Paper. http://ec.europa.eu/competition/consultations/2014_merger_control/mergers_white_paper_en.pdf

Further Reading

- Fotis P, Zevgolits N (2016) *The competitive effects of minority shareholdings legal and economic issues*. Hart Publishing, Bloomsbury

List of Cases

- COMP/M.1980 Volvo/Renault (2000) OJ C 301
- COMP/M.2050 Vivendi/Seagram (2000) C 261/04

Patent Annulment

► Patent Litigation

Patent Infringement Lawsuit

► [Patent Litigation](#)

Patent Invalidation Challenge

► [Patent Opposition](#)

Patent Litigation

Elisabetta Ottoz

Department of Economics and Statistics Cognetti
De Martiis, University of Turin, Turin, Italy

Abstract

Patent litigation refers to patent infringement lawsuits or revocation proceedings. Infringement is the act of making, using, selling, or offering to sell a patented invention without the permission of the patent owner. Revocation proceedings refer to the claim on patent validity before civil courts that may be carried out by firms interested not to be sued for infringing “wrongly” granted patents. Presently, national courts of the member states of the European Patent Convention are competent to pass judgment on the infringement and validity of European patents, with inevitable consequences in terms of duplication and inconsistencies. In December 2012, the European Parliament approved the EU unitary patent package, whose ratification by the individual member states will give rise to a European patent with unitary effects in all jurisdictions involved and to the creation of a Unified Patent Court (UPC) with exclusive jurisdiction to hear infringement and invalidity actions. The result will be a patent protection for all participating member states based on a single application and validation.

Synonyms

[Patent annulment](#); [Patent infringement lawsuit](#)

Definition

Patent litigation refers to patent infringement lawsuits or revocation proceedings. Infringement is the act of making, using, and selling a patented invention without the permission of the patent owner. Revocation proceedings refer to the claim on patent validity before civil courts.

Introduction

A patent is the exclusive right granted by a government to an inventor to manufacture, use, or sell an invention for a certain number of years in exchange for detailed public disclosure of the invention itself so as to encourage research and development activities fostering knowledge dissemination.

In order to be patentable, an invention must satisfy some requirements concerning novelty, usefulness, and nonobviousness (WIPO 2008).

A patent is not a perfect protection against imitation, but it grants the patent holder the right to sue intruders once they have been identified. Conversely third parties have the right to challenge the validity of patents granted by the patent authority.

Patent litigation can thus take two distinct forms: infringement or revocation proceedings.

Patent infringement is the act of making, using, selling, or offering to sell a patented invention without the permission of the patent owner. The economic significance of a patent depends on its scope: the broader the scope, the larger the number of competing products and processes that will infringe the patent. The claims contained in the application are the basis of the extent of patent protection as they determine what third parties are legally allowed to do.

Revocation proceedings refer to the claim on patent validity before civil courts that may be carried out by firms interested not to be sued for infringing “wrongly” granted patents, either autonomously or as a counterclaim in a cause for patent infringement.

Proceedings before a national tribunal for revocation as part of patent litigation are not to be

confused with patent opposition, which is an administrative procedure contained in the article 99 of the European Patent Convention allowing third parties to question the validity of a patent granted by the European Patent Office (EPO) within 9 months from the its publication. Patent opposition applies to the European patent at the European-wide level, whereas revocation proceedings within litigation apply to national jurisdictions.

Infringement and validity of European patents are currently under the jurisdiction of national courts and authorities of the member states of the European Patent Organisation.

Despite the Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights, relevant differences still exist as patent litigation and court judgments on validity and infringement vary significantly from one country to another (EPO 2013a). In practice, this gives rise to a number of shortcomings: high costs, risk of diverging decisions, and lack of legal certainty (Luginbuehl 2011).

The European patent system is currently undergoing major reforms aimed at overcoming existing fragmentation, which becomes increasingly problematic as innovation and industrial R&D has assumed a global scope. In December 2012, the European Parliament approved the EU unitary patent package, whose ratification by the individual member states will give rise to a European patent with unitary effects in all jurisdictions involved and to the creation of a Unified Patent Court (UPC) with exclusive jurisdiction to hear infringement and invalidity actions.

Institutional Features

European inventors can obtain patent protection by filing several national applications or, alternatively, one patent application to the EPO in which several States adhering to the 1973 European Patent Convention (EPC) are designed. The granting of a European patent allows the applicant to achieve a bunch of patents treated as independent rights, each having a limited scope: whether or not

the national parts of the European patent are infringed or invalid is then determined based on the national laws of the respective member states of the European Union.

Relevant institutional differences among the jurisdictions still exist concerning several aspects such as the existence of bifurcation, remedies for patent infringement, forum shopping, and the allocation of legal costs as illustrated by Graham and Van Zeebroeck (2014).

Bifurcation

In terms of institutional settings, a major difference relates to whether litigants are permitted or required to address infringement and invalidity claims within the same court and suit or, as in a bifurcated system, infringement is heard and determined separately from validity.

The German system is a bifurcated one: invalidity challenges, either standalone invalidation challenges or appeals of decisions rendered by the German Patent Office, can only be brought to the Federal Patent Court, whereas infringement actions can be lodged in any of the twelve competent district courts. In the French, British, Dutch, Italian, and Belgian systems, patent infringement and invalidity actions, either for national patents or for national validations of patents granted by the EPO, are brought to the same court.

The bifurcated patent system is positively considered because of quick decisions and low costs, but it might be potentially biased toward the patentee in two ways. The fast infringement proceedings in the regional courts mean, in fact, that it is possible to get to an injunction before the patent can be invalidated by the slower invalidity courts. In addition, a bifurcated system is subject to the so-called Angora cat problem: the patentee will argue for a narrow interpretation of his claim when defending the patent but an expansive interpretation when asserting infringement.

Injunctions

If a patent holder discovers that his/her patent is being infringed by products or services belonging to other parties, he/she faces the decision to file a

lawsuit and, following that, either to engage the alleged infringer in a pretrial settlement negotiation or to file suit in court claiming damages for the infringement.

The TRIPS Agreement generally provides for injunctions and damages as a remedy to patent infringement, but the specific procedures and standards for awarding these remedies are left to the member states, which apply different conditions and thresholds.

Courts in several member states allow in principle preliminary injunction, which is granted very early in a court action and restrains the defendant from infringing the patent during the pendency of litigation (Cotter 2011).

If a preliminary injunction is issued, the plaintiff will nearly always have to post a bond for securing any costs or damages caused to the defendant in case infringement is not proved. The amount of the bond is left to the court's discretion.

If the plaintiff wins at the trial, the preliminary injunction usually becomes permanent, but if the defendant wins, the preliminary injunction is removed.

Cross-border preliminary injunctions forbidding accused infringers from practicing the litigated patents both in the domestically and abroad have been applied in the Dutch litigation system with the result of a "forum shopping strategy," as patent holders may stop infringement of their patent throughout Europe.

Belgium too has become a preferred venue for patent owners who wish to quickly enforce their patent rights: Belgian courts assume, in fact, in their preliminary injunctions the *prima facie* validity of the patents, even when an opposition was pending or appealed at the EPO.

Damages

Most jurisdictions provide for three "standard methods" for damages assessment based on the calculation of lost profits, reasonable royalty, and infringers' profits (unjust enrichment) (Reitzig et al. 2007).

In the case of lost profits, the patentee shall be reinstated in a position where he/she would have been but for the infringement. The calculation

method is accepted by all major jurisdictions (USA, Japan, Germany, UK, and France).

In the US jurisdiction, the patentee, in order to be awarded lost profits, has to show causation, establishing that "but for" the infringement, she would have made additional profits. In 1978, the Court of Appeals for the Sixth Circuit put forth a test designed to determine whether a patent holder is entitled to recover for lost profits by the infringement (*Panduit Corp. v. Stalin Bros. Fibre Works, Inc.*, 575 F.2d 1152, 1156 (6th Cir. 1978)).

It is a four-step test requiring that the patentee establishes (1) the demand for the patented product, (2) the absence of noninfringing substitutes, (3) the manufacturing and marketing capability to exploit the demand, and (4) the amount of profit that would have been made absent the infringing product.

Assuming that the four Panduit conditions are met, lost profit evaluation requires to infer how the market would have evolved absent infringement and to compare the hypothetical behavior of both the patentee and the infringer with their actual behavior. The difference between the "but for" and the actual profit represents the patent holder's lost profit damage.

Proof of damages is simpler when the patentee and the infringer compete in a two-supplier market, notwithstanding a market share analysis can be used even in case of several firms in the market in order to determine a measure of lost profit award.

Once lost sales are determined, total lost profits can be calculated by measuring the incremental profits on lost sales plus profits lost on price erosion.

To the extent the patentee does not satisfy "but for" causations required by the Panduit steps, he is entitled to a reasonable royalty. The courts attempt to reconstruct the hypothetical bargain that the parties would have negotiated had they willingly tried to do so at the time infringement began. A hypothetical negotiation between a willing licensor and a willing licensee is imagined generally relying on fifteen factors set forth in *Georgia-Pacific Corp. v. United States Plywood Corp.* (1970).

The case law has established a few guiding principles that should be present in any reasonable royalty determination: first of all the best measure of reasonable royalty is an established royalty rate in the industry; second the hypothetical negotiation takes place at the time the infringement began, meaning that infringers' sunk cost are not part of the infringer's anticipated profits; and third, the patentee need not prove any actual harm to be entitled to a reasonable royalty (Frank and DeFranco 2000).

A damage award can be composed of lost profits and a reasonable royalty as, for instance, in the case where lost profits include lost sales for which the patentee had manufacturing capacity, while a reasonable royalty accounts for additional sales that exceeded the patentee's manufacturing capacity. In case the patent holder is a university or a research institute, the damage is entirely based on a reasonable royalty.

The third way of calculating damages relies on infringer's profits. In Germany the damage "is based on the legal fiction that in using another's patent, the infringer undertook a business on behalf of the rights-owner, who would thus be entitled to obtain all profits made from such business" (Reitzig et al. 2007). Granting "infringers' profits" is formally not allowed in France and the USA, even if the US term "unjust enrichment" may be interpreted as a rather close notion.

In case of willful infringement, in the USA, whether the damage award is in the form of lost profits or reasonable royalties, courts have discretionary authority to enhance the damage award by three times.

Courts have sometimes awarded inflated reasonable royalties that do not reflect the market ones, but rather imply a deterrent function against future infringements (Love 2009).

Choice of Fora

When a patent infringement suit involves a defendant domiciled in an EU member state, national courts of member states must exercise jurisdiction in accordance with Articles 2 and 5(3) of the Brussels Regulation. Under this regime, a plaintiff may bring an action in the courts of the defendant's domicile or in the state(s) in which the

alleged infringing product was manufactured or commercialized in breach of a local patent. Dutch courts have even exercised jurisdiction over foreign defendants for violations of foreign patents.

This means that patent litigants in Europe are permitted a choice of fora, which gives rise to an opportunity to engage in "forum shopping," a strategic choice of court venues in order to obtain a favorable outcome leading to economic inefficiencies. Parties try to take advantage of differences in national courts' interpretation of European patent law and in procedural laws, as well as of differences in speed and in the amount of damages awarded.

"Forum shopping is a common practice in the USA too, as any civil action for patent infringement may be brought in the judicial district where the defendant resides or where the defendant has committed acts of infringement and has a regular and established place of business" 28 USC. § 1400 (Lemley 2010).

The Allocation of Legal Costs

Two main systems for allocating litigation costs are applied, namely, the "American system," where each party bears its own costs, and the "British system," where the loser incurs all costs. In an intermediate position lie the systems which allow a partial fee shifting.

The legal-cost allocation rule plays a role in favoring patent litigation or settlement and bears implications on the royalty-bargaining process.

The Unified Patent Court

In order to overcome duplication and inconsistencies, in December 2012 the European Parliament approved the EU unitary patent package, whose ratification by the individual member states will give rise to a European patent with unitary effects in all jurisdictions involved, that is to say subject to the same legal conditions in all member states. The result will be a patent protection for all participating member states, except Italy and Spain, based on a single application and validation putting an end to validations and litigation of the patent in each state.

A Unified Patent Court (UPC), with exclusive jurisdiction to hear infringement actions, invalidity actions and counterclaims, and actions for provisional and protective measures and injunctions for litigation relating to European patents and European patents with unitary effect (unitary patents), will be created considering a transitional period of 7 years during which actions of litigation may be brought before national courts (Agreement on a Unified Patent Court and Statute 2013b)

The UPC will comprise a court of first instance, a Court of Appeal, and a registry. The court of first instance will be composed of a central division (with seat in Paris and two sections in London and Munich) and by several local and regional divisions in the contracting member states to the Agreement. The Court of Appeal will be located in Luxembourg.

Generally, claimants will bring action for revocation before the central division and will bring actions for infringement before a local/regional division in a member state in which the infringement has occurred or where the defendant is domiciled.

The system allows for a choice between bifurcation – the separation of infringement and validity claims into separate court actions as in the German system – and an integrated process for hearing infringement and invalidity cases: the regional courts have, in fact, the discretion to refer counterclaims for revocation raised by the defendant to the central division.

Where a decision is taken finding an infringement of a patent, the Court may grant an injunction against the infringer aimed at prohibiting the continuation of the infringement.

As for the award of damages Art. 68 of the UPC states:

1. The Court shall, at the request of the injured party, order the infringer who knowingly, or with reasonable grounds to know, engaged in a patent infringing activity, to pay the injured party damages appropriate to the harm actually suffered by that party as a result of the infringement.

2. The injured party shall, to the extent possible, be placed in the position it would have been in if no infringement had taken place. The infringer shall not benefit from the infringement. However, damages shall not be punitive.
3. When the Court sets the damages:
 - (a) It shall take into account all appropriate aspects, such as the negative economic consequences, including lost profits, which the injured party has suffered, any unfair profits made by the infringer and, in appropriate cases, elements other than economic factors, such as the moral prejudice caused to the injured party by the infringement; or
 - (b) As an alternative to point (a), it may, in appropriate cases, set the damages as a lump sum on the basis of elements such as at least the amount of the royalties or fees which would have been due if the infringer had requested authorisation to use the patent in question.

Legal costs and other expenses incurred by the successful party shall, as a general rule, be borne by the unsuccessful party.

The Agreement was signed by 25 EU member states on 19 February 2013. It will need to be ratified by at least 13 states, including France, Germany, and the UK to enter into force: at the moment, August 2014, the Agreement has been ratified by five member states.

Empirical Evidence

Characteristics and changes in the US patent litigation system have been studied by a large number of authors (Lanjouw and Schankerman 2001; Atkinson et al. 2009; Henry and Turner 2006).

Empirical evidence on patent litigation in the largest and most judicially active countries of the European Union is provided by recent studies, which give some real insights based on recent patent suits.

Graham and Van Zeebroeck (2014) analyzing a dataset of European patent litigation during 2000–2010, comprising approximately 9000

judicial patent decisions from seven European countries, show that the incidence of litigation and the bases of judicial outcomes diverge radically across the different countries and technology sectors. Relevant differences are also detected in the likelihood of patent litigants raising patent validity and infringement claims.

Litigation rates are highest in Belgium and France, whereas Germany and the UK show a low rate. Patent litigation varies widely across technology sectors, with the majority of cases in Europe focusing on patents granted for industrial processes, civil engineering, consumer goods, machinery, and transport technology.

Litigation costs significantly differ in European jurisdictions ranging from 50.000 to 200.000 € in France, Germany, and the Netherlands, but being considerably higher in the UK, 150.000–1500.000 €, which may explain the lower number of cases brought to court in the UK. It is worth noticing that the average cost in the USA is much higher ranging from 1.000.0000 to 10.000.000 €.

On the assumption that the Unified Patent Court will offer litigation at roughly the same cost level as the three largest low-cost national systems, Harhoff (2009), by utilizing different data sources, estimates that the total savings from the creation of the unified Patent Court are considerably larger than the actual operating costs, even for the most conservative scenarios. For 2013, the benefit-cost ratios would range between 5.4 and 10.5: in other words, duplication of litigation combined with high costs of litigation, in some countries, costs firms about 5.4–10.5 times more than the establishment and annual operation of the Unified Patent Court.

A comparison of 8,323 patent litigation cases across Germany, France, the Netherlands, and the UK, covering cases filed during the period 2000–2008 (Cremers et al. 2013), highlights relevant differences in the four jurisdictions concerning the number of case loads, settlement rate, average time for judgment, outcomes, characteristics of the litigants, fragmentation, sector distribution of litigants, and value of patents.

Out of a total of 6,739 cases in Germany, 5,121 are infringement cases heard by the three regional

courts covered by our study, whereas 1,618 are revocation cases. By far the number of infringement cases heard by German courts exceeds the combined number of cases in all three other jurisdictions. Depending on how cases are defined, Germany has between 12 and 29 times as many litigation cases as the UK; the difference is similar with regard to the Netherlands; compared to France, Germany has around six times as many cases.

The settlement of disputes reveals relevant differences across countries: settlement is more likely in Germany (60% of cases) as compared with UK (40%). As for cases decided by a judge, revocation is the most likely outcome regardless of whether the initial claim was for infringement or revocation in UK, whereas infringement is in Germany and the Netherlands. France is characterized by a large share of patents that is held not to be infringed, but valid.

The time lag between the filing for a claim for infringement and a first decision is less than 1 year in Germany, the Netherlands, and the UK, nearly double in France. Claims for invalidity are decided fastest in the UK (11.2 months), but take a lot longer to be decided in Germany (15 months) and the Netherlands (11.4 months). Again, invalidity cases in France take significantly longer (19.8 months) than in any other jurisdiction.

There are large differences across jurisdictions with regard to case outcomes. Infringement cases with court decision amount to about 22% in Germany, 36% in the Netherlands, 14.7% in the UK, but only 5.6% in France where most patents are held valid, but not infringed. In the UK the large share of revoked patents of cases that allege infringement, 26%, is due to the fact that, in about 60% of cases alleging infringement, the defendant counterclaims for revocation.

Also outcomes of invalidity actions differ considerably across jurisdictions. Whereas in the UK 42% of patents are revoked if the case is decided by the judge, less than half as many invalidity cases end with revocation in Germany and France. The risk of infringing a patent that forms the subject of a revocation action is very low in all jurisdictions (4% in the UK and 7% in Germany).

Fragmentation leading to parallel litigation of the same patent in multiple jurisdictions is low in Germany (2%) and France (6%), but more relevant in the Netherlands (15%) and in the UK (26%). However, as the number of cases in the UK and the Netherlands is considerably lower than in Germany, the upper bound for the share of duplicated cases lies in Germany.

As for the characteristics of the litigating parties involved in the patent cases, half of all cases involve only domestic claimants in Germany and France. The share of cases with only domestic claimants drops below 40% in the UK and the Netherlands. The data look similar for defendants, with the exception of Germany where the share of cases with only domestic defendants exceeds 60%.

By sorting litigants by type – companies, individuals, universities, public research institutes, government, as well as international institutions – the largest differences in the shares of companies and individuals involved in patent cases are found across jurisdictions rather than between claimants and defendants. France, where there are almost twice as many individuals as defendants than there are claimants, is an exception. Overall the share of companies as claimants or defendants is smallest in Germany; on the contrary the greatest share of litigants in the UK falls into the “large” category. In all other jurisdictions, micro- and small companies represent the largest share of litigants.

As for the sector distribution of litigating companies, the share of pharmaceutical companies in the UK is the highest of the four jurisdictions amounting to 30%. In Germany, in contrast, companies are concentrated in manufacturing, notably the machinery and engine industry. In the Netherlands, the share of companies in the services industry (especially finance, insurance, and real estate) stands out. France does not show a strong characterization.

The number of forward citations received worldwide, a proxy for patent value, is significantly higher for the litigated patents compared to the group of non-litigated patents.

Strategic Use of Rules

First economic analyses of patent litigation stressed the role of divergent expectations (Priest and Klein 1984) and the presence of asymmetric information (Spier and Spulber 1993) in fostering litigation.

Later models argue that patent litigation reveals important information for potential entrants (Choi 1998), analyze patent enforcement through litigation when firms have private information (Llobet 2003), and compare the two doctrines of damages, lost profit and unjust enrichment (Schankerman and Scotchmer 2001).

More recently several studies have pointed out that the litigation system is likely to play a crucial role when patents are used as assets or as legal threats. Patent litigation can then be abused to extort licensing payments by patent-assertion entities (also known as “trolls”). Trolls engage in deliberate strategies whose aim is to acquire patents of failed companies or independent innovators using them to threaten suit against alleged infringers, without having the intention of actively using the patents they assert. In particular, in component-driven industries, notably information technology, trolls engage in deliberate tactics allowing them to take product developers by surprise once they have made irreversible investments (Lemley and Shapiro 2007).

As trolling activity, though not illegally, seeks to exploit structural and procedural weaknesses of the patent and judicial system to earn rents, an optimally designed patent litigation system should minimize the room for such welfare-reducing behavior.

The significant occurrence of trolls in the USA may find its roots in the high costs of legal proceedings, cost allocation rules (each party bears its own costs), contingency fees, high damage awards, and injunctive reliefs, characterizing the US litigation procedure. Moreover, questionable examination quality in patent granting and broadly defined patentable subject matter also play a role.

The weaker presence of “trolls” in Europe is presumably because the patent systems in Europe deviate from the US system in several crucial

points. Generally, court proceedings are much less costly, cost allocation favors the winning party, damage awards are not excessive, most courts have sought a careful balance between the rights of the parties, injunctions are not issued automatically, and the quality of patent examination has been considerably better than in the USA.

However, one should not assume that the European system is troll-proof: recently patent funds have acquired patent portfolios consisting of several thousand patents, largely European ones, and may seek to enforce them (Harhoff 2009).

Conclusions

The birth of the UPC will represent a considerable progress for the management of intellectual property in the EU in order to overcome duplication and inconsistencies and to lower litigation costs. The result will be a patent protection for all participating member states based on a single application and validation with a Unified Patent Court (UPC) with exclusive jurisdiction to hear infringement and invalidity actions.

A relevant argument put forth by the proponents of the UPC is that it will also reduce strategic behavior in Europe. This may be true for forum shopping, that is to say a strategic choice of courts venue by litigants to obtain a favorable outcome: nevertheless a new form of forum shopping might be originated by the UPC if local divisions behave differently with respect of the willingness to grant EU-wide injunctions and with respect to the attitude toward bifurcation.

As for the consequences of the EU-wide injunction, it is worth stressing that it might represent an incentive for “trolling activities,” so far not so common in the EU, suggesting a cautious use of injunctions.

References

Atkinson S, Marco A, Turner J (2009) The economics of a centralized judiciary: uniformity, forum shopping and the federal circuit. *J Law Econ* 52:411–443
 Choi J (1998) Patent litigation as an information-transmission mechanism. *Am Econ Rev* 88:1249–1263

Cotter T (2011) A research agenda for the comparative law and economics of patent remedies. Minnesota legal studies research paper no. 11-10
 Cremers K, Ernicke M, Gaessler F, Harhoff D, Helmerts C, McDonagh L, Schliesser P, Van Zeebroeck N (2013) Patent litigation in Europe. ZEW discussion paper 13-072 <http://www.econstor.eu/bitstream/10419/83473/1/769014895.pdf>
 Directive 2004/48/EC. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:195:0016:0025:en:PDF>
 EPO (2013a) Patent litigation in Europe, 3rd edn. European Patent Academy, Munich. http://www.eplit.eu/files/downloads/patent_litigation_in_europe_2013_en.pdf
 EPO (2013b) Agreement on a unified patent court. [http://documents.epo.org/projects/babylon/eponet.nsf/0/A1080B83447CB9DDC1257B36005AAAB8/\\$File/upc_agreement_en.pdf](http://documents.epo.org/projects/babylon/eponet.nsf/0/A1080B83447CB9DDC1257B36005AAAB8/$File/upc_agreement_en.pdf)
 Frank R, DeFranco D (2000) Patent infringement damages: a brief summary. *Fed Cir B J* 10:281–291
 Graham S, Van Zeebroeck N (2014) Comparing patent litigation across Europe: a first look. *Stan Tec L Rev* 17:655–708
 Harhoff D (2009) Economic cost-benefit analysis of a unified and integrated European patent litigation system. Tender No. MARKT/2008/06/D. http://ec.europa.eu/internal_market/indprop/docs/patent/studies/litigation_system_en.pdf
 Henry M, Turner J (2006) The court of appeals for the federal circuit’s impact on patent litigation. *J Legal Stud* 35:85–117
 Lanjouw J, Schankerman M (2001) Characteristics of patent litigation: a window on competition. *RAND J Econ* 32:129–151
 Lemley M (2010) Where to file your patent case. *AIPLA Q J* 38:1–35
 Lemley M, Shapiro C (2007) Patent holdup and royalty stacking. *Tex Law Rev* 85:1991–2049
 Llobet G (2003) Patent litigation when innovation is cumulative. *Int J Ind Organ* 21:1135–1157
 Love B (2009) The misuse of reasonable royalty damage as a patent infringement deterrent. *Mon Weather Rev* 74:909–948
 Luganbuehl S (2011) European patent law: towards a uniform interpretation. Edward Elgar, Northampton
 Priest G, Klein B (1984) The selection of disputes for litigation. *J Legal Stud* 8:1–56
 Reitzig M, Henkel J, Heath C (2007) On sharks, trolls, and their patent prey—unrealistic damage awards and firms’ strategies of “being infringed”. *Res Policy* 36:134–154
 Schankerman M, Scotchmer S (2001) Damages and injunctions in protecting intellectual property. *RAND J Econ* 32:199–220
 Spier K, Spulber D (1993) Suit settlement and trial: a theoretical analysis under alternative methods for the allocation of legal costs. *J Legal Stud* 11:55–81
 WIPO (2008) Chapter 2: Fields of intellectual property protection. In: Intellectual property handbook: policy, law and use. <http://www.wipo.int/export/sites/www/about-ip/en/iprm/pdf/ch2.pdf>

Patent Opposition

Alessandro Sterlacchini

Department of Economics and Social Sciences,
Università Politecnica delle Marche,
Ancona, Italy

Abstract

A patent opposition allows third parties to question the validity of the patents granted by the European Patent Office (EPO) on the grounds that they do not meet patentability criteria, do not fully disclose the invention, or extend beyond the original application. These issues are debated before an Opposition Division and, eventually, a Board of Appeal of the EPO which decides whether opposed patents are upheld as granted, amended, or revoked. The evidence indicates that these three possible outcomes are equally probable. Since the EPO decision applies to all the states designed in the application, the patent opposition represents a unique opportunity for challenging a patent's validity at European-wide level. Along with their relatively lower costs, this explains why, in Europe, patent oppositions are used by far more frequently than patent litigation.

Synonyms

Patent invalidation challenge; Post-grant patent review

Definition

A patent opposition is an administrative procedure adopted by the European Patent Office which allows third parties to challenge the validity of a granted patent on the basis of specific grounds.

Introduction

A patent granted by a public authority is presumed to be valid. This means that the patented invention

is novel, based upon an inventive step (or “nonobvious”), and susceptible of industrial application (useful). Moreover, the patent application must disclose detailed information on the invention so that a person “skilled in the art” should be able to replicate it. To control for these requirements, patent examiners should have a relatively easy and cheap access to the relevant information about the state of prior art in specific technological fields. This condition is difficult to meet for many reasons: among them, the emergence of new fields, such as those of bio- and nanotechnologies, rooted on variegated but complementary disciplines, a staggering increase of patent applications, and a growing number of claims per application (Archontopoulos et al. 2007). To be stressed is that each claim identifies a specific property right that the patent should protect and, as such, must be validated by patent examiners.

The number of patent examiners has not expanded in line with that of applications and claims, and this has determined a growing workload and backlog in patent offices. For having a granted patent at the EPO, an average of 3 years was necessary during the early 1980s, while in the early 2000s, the examination delay increased to 5 years (van Zeebroeck 2011). To avoid an excessive length of the examination process, the evaluation of each application has become less accurate, subject to possible errors, and, thus, likely to increase the number of low-quality patents susceptible of being invalidated.

A growing uncertainty about the validity of patents generates remarkable economic losses to society (Hall and Harhoff 2004). Patent holders could underinvest in some particular technologies, and their rivals could reduce investment in competing technological advances; if both subjects have already undertaken substantial investments on these activities, they will be prone to embark in costly litigation. If the cost and length of time for invalidating patents are too high, large companies with wide financial means have an incentive to inflate their patent portfolio with low-quality patents. In this way they create a strategic barrier to entry for small innovative

companies, and this will reduce the overall pace of innovation.

To avoid these negative consequences, including that faced by the companies that risk to be sued for infringing “wrongly” granted patents, an efficient and not too expensive procedure for reviewing their validity is necessary. In principle *ex post* litigation before civil courts can fix some important errors of patent offices. In practice, however, patent litigation appears to be a too costly mechanism, giving rise to extremely uneven incentives to challenge and defend issued patents (Farrell and Merges 2004): in particular, small firms and independent inventors are likely to be severely discriminated in favor of bigger and financially wealthier patent holders (Lanjouw and Schankerman 2001, 2004; Kingston 2004; Schettino and Sterlacchini 2009).

An effective and cheaper alternative to litigation is an administrative post-grant review in which informed persons or entities have the chance to disclose relevant pieces of information that were not available or not adequately taken into account during the patent’s examination and that could undermine its validity. In the following, the procedure of patent opposition before the EPO is examined along with some evidence about the frequency, the determinants, and the outcomes of oppositions.

Institutional Features

European inventors can obtain patent protection by filing several national applications or, alternatively, one patent application to the EPO in which several states adhering to the European Patent Convention (EPC) are designed. Considering the relative costs, if patent protection is sought in at least four countries, the EPO-centralized route is more convenient. The granting of a European patent allows the applicant to achieve a bunch of patents that are valid in different countries (obviously, upon having paid the national fees and translating the documents).

According to Article 99 of the EPC, within 9 months from the publication of the mention of

the grant of a European patent, any person can challenge its validity by filing an opposition against the granting decision of the EPO. Although most of the opponents are rivals of patent holders seeking to obtain the limitation or revocation of a patent, in principle, it is not necessary for the opponent to have or manifest a particular interest in the patented invention. The notice of opposition may be filed, jointly or separately, by more than one opponent.

Section 2 of the same article makes clear that “The opposition shall apply to the European patent in all the Contracting States in which the patent has effect.” Thus, in Europe, the opposition at the EPO represents a unique opportunity to challenge the validity of a patent at the European-wide level rather than in individual national courts. It should be added that only in 13 out of the 38 states adhering to the EPC, there is a procedure for post-grant patent oppositions (EPO 2013). Among the largest countries, only in Germany the patent law provides for invalidity actions in a unique national court (the Federal Patent Court) separated from other civil courts specialized in litigation for patent infringements. Instead, post-grant oppositions are not allowed in France, the UK, and Italy, while only to a limited extent in Spain. Thus, multiple suits before the civil courts of different countries could be necessary to invalidate a European patent. Compared to the centralized procedure at the EPO, such an alternative implies not only more costs (see below) but also a non-negligible level of uncertainty: in fact, it cannot be taken for granted that different national courts will achieve the same decision.

The grounds for opposition, established in Article 100 of the EPC, are the following:

1. The subject matter of the patent is not patentable (lack of novelty, inventive step, and industrial applicability, according to Articles 55–57 of the EPC).
2. The patent does not disclose the invention in a manner sufficiently clear.
3. The subject matter of the patent extends beyond the content of the filed application.

The opposition process is overseen by an Opposition Division, composed of three technical examiners of the EPO, at least two of whom must not have taken part in the proceedings for granting the opposed patent (Article 19, EPC). If the complexity or specificity of the case so requires, the Opposition Division can decide to include an additional legally qualified examiner who has not taken part in the proceedings for grant. The same requirement holds for being the Chairman of the Opposition Division. In case of parity of votes, that of the Chairman is decisive.

After 2 months from the filing of an opposition, the opponent must present his/her arguments and evidence for asking a revision of the EPO decision. Then, the patent holder has up to 6 months to reply, and the same time is allowed to the opponent for his/her counterarguments. After the exchange of observations, an oral hearing of arguments (normally open to the public) takes place before the Opposition Division. Then, the Division communicates to the parties its decision. On average, all the process is accomplished in about 22 months (Harhoff 2005).

The three possible outcomes from an opposition proceeding are the following (Article 101, EPC):

1. The opposition is rejected and the patent is upheld as granted.
2. The patent is maintained with amendments based on reformulations or cancelations of claims.
3. The patent is revoked.

The parties adversely affected by the above decisions may appeal to the EPO's Boards of Appeal, and this additional procedure can last almost 26 months. If a Board of Appeal confirms the revocation decision, the patent is invalid in all the states designated in the application. If the patent is maintained as granted, the opponents can resort to invalidity actions in the national civil courts of the designated states. In principle, the same possibility is allowed for amended patents, but the modification or cancelation of claims should reduce the likelihood of further legal disputes.

An important element that differentiates the European post-grant review from that of other countries (including the USA) is that the opposition procedure before the EPO is an adversarial process in which the legal representatives of the parties (opponent and patent holder) have the possibility of airing and debating their arguments before an adjudicator (Rotstein and Dent 2009). In this sense, an opposition proceeding resembles a validity suit before a civil court. However, while a patent litigation can be settled "out of court" (and this is what occurs in many circumstances), an opposition procedure may be continued by the EPO of its own motion even when the opposition is withdrawn (Rule 84, EPC).

Empirical Evidence

The frequency of oppositions of the patents issued by the EPO was found particularly high during the first two decades of the office's life: the average opposition rate over the period 1980–1995 was about 8% (Harhoff and Reitzig 2004). Figure 1 shows that, during the subsequent years, the share of opposed patents has constantly decreased, a part from a slight recovery in the mid of 2000s.

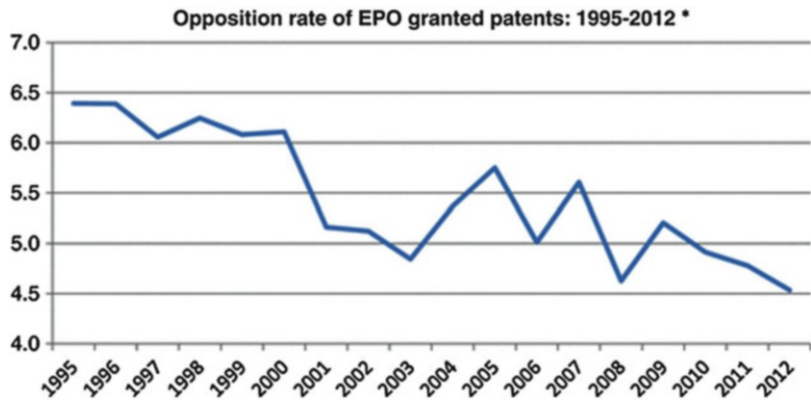
From 1995 to 2000, the mean opposition rate was around 6%, while between 2000 and 2012, it reduced to 5%. The last annual report of the EPO concerned with the year 2013 documents a frequency of opposition equal to 4.7%.

The decrease of the opposition rate in the last decade is due to an almost constant number of opposition cases (about 2,900 per year), while the number of granted patents has increased from 60,000 in 2003 to a record of 66,700 in 2013. However, this does not mean that to obtain a patent from the EPO is becoming easier than it was in the past.

About 60% of all applications filed at the EPO between 1980 and 2002 ended with a grant (van Zeebroeck 2011). Instead, according to the last annual reports of the EPO, over the years 2006–2012, the share of granted patents on the processed applications (through the first search on prior art and, then, the proper examination) dropped to 47%. Almost 23% of the applications

Patent Opposition,

Fig. 1 Opposition rate of EPO-granted patents: 1995–2012 (oppositions in year t on the average patents granted in year t and $t - 1$ (Source: EPO Annual Reports))



were withdrawn by the applicants after the search report, and another 30% were withdrawn or refused during or at the end of the examination process. The actual rate of refusals is not particularly high (around 5%), but it does not take into account that most of the withdrawals during the examination phase are induced by the “toughness” of EPO examiners who, by communicating detailed objections and remarks to the applicants, are able to discourage the less valuable applications (Lazaridis and van Pottelsberghe 2007).

Thus, being coupled with a remarkable drop of the share of granted patents, the observed decrease in the rate of oppositions could be a signal that the average quality of the patents granted by the EPO has improved over time giving less scope for invalidity challenges. This explanation is supported by the results of a survey jointly carried out in 2011 by Thomson Reuters and the *Intellectual Asset Management* magazine (issue of July/August 2011): the EPO was ranked first for patent quality among the world’s five largest patent offices for the second consecutive year. Compared with the Japanese and, especially, the US patent office, the EPO was leading by a wide margin in terms of perceived patent quality and also improved its position with respect to previous years. Consistent findings are provided by de Saint-Georges and van Pottelsberghe (2013) who, by using a composite index of nine variables capturing the transparency of patent systems and the quality of examinations, show that the EPO ranks first among 32 patent offices. It should be added that such a high level of real or perceived quality was not achieved at the expense of the

length in the granting process. On the contrary, in the years 2011–2012, the average delay to grant was lower than 4 years, a duration inferior to that recorded in the early 2000s (see above).

In spite of its declining trend, the rate of patent oppositions in Europe remains much higher than that of patent litigation before civil courts. Due to the presence of multiple and heterogeneous patent litigation systems among European countries, aggregate data on patent litigation at European-wide level are not available. However, some recent works have attempted to collect consistent data for some of the largest countries of the European Union.

Considering the judicial patent decisions from seven European countries published over the period 2000–2010, the litigation rate varies from a minimum of 0.1% in the UK to a maximum of 1.5% in the Netherlands (Graham and van Zeebroeck 2014). The two largest countries of the EU, Germany and France, record a frequency of patent litigation equal to 0.3% and 0.9%, respectively. These percentages underestimate the actual rate of patent disputes because those settled “out of courts” are neglected (for an analysis of patent litigation settlements in Germany, see Cremers and Schliessler 2015). However, it must be stressed that the majority of patent litigation refers to infringement rather than invalidity actions: recent data on patent litigation for France, Germany, the Netherlands, and the UK, collected from 2000 to 2008, show that only 22% of the patent disputes correspond to revocation cases (Cremers et al. 2013). As a consequence, the frequency of nullity actions before civil courts is

much lower than that of oppositions before the EPO.

The prevalence of patent oppositions as compared to invalidity litigation in Europe is mainly due to the different costs involved. Over the last decade, the total costs of an opposition before the EPO (including patent lawyers' fees) vary from €6,000 to €50,000 for each party (Mejer and van Pottelsberghe 2012). Instead, considering the proceedings before first-instance courts and patent cases of small and medium scales, the average cost of patent litigation ranges from €50,000 to €500,000, although in the UK, the maximum cost can be up to €1.5 million (EPO 2006). The country in which patent litigations are less expensive is Germany; much higher costs are documented in the UK, while France and the Netherlands record an intermediate level of litigation expenses (Cremers et al. 2013). It should be added that, in case of multiple litigation, the costs have to be cumulated across the national jurisdictions involved.

In terms of opposition rates, the differences among sectors or technological areas are remarkable. By considering the period 2000–2008, the opposition frequency of EPO-granted patent was found particularly high in the fields of chemicals and pharmaceuticals and lower in information and communication technologies (Caviggioli et al. 2013).

A large body of empirical evidence converges in showing that the opposition probability is significantly associated with the patent quality or value (Graham et al. 2003; Harhoff and Reitzig 2004; Cincera 2011; Schneider 2011). The latter can be approximated by different indicators: the most diffused and effective quality measures are the number of citations received by a patent (forward citations) and the size of patent families (given by the number of countries in which patent protection is sought for the same invention); other employed indicators are the number of backward citations (references to previous patents) and claims. The evidence suggesting that the most valuable patents (according to the above measures) are more likely to be opposed should be interpreted with some caution. In fact, an opposed patent that is revoked by the EPO cannot be

considered of high quality. A similar consideration applies to patents with many claims that could be changed or canceled at the end of the opposition proceeding. In short, what can be safely said is that only the patents that survived an opposition have a higher quality and, as such, are more likely to be successfully enforced in subsequent legal disputes for infringement (Harhoff et al. 2003; van Zeebroeck 2011).

Another interesting issue that can be examined by using patent opposition data is whether the occurrence of an opposition could be due to strategic reasons. The relevance of this question is due to the fact that patent applications are significantly concentrated in a few hands: suffice it to notice that both in 2011 and 2012, the top ten applicants at the EPO accounted for almost 11% of the total applications filed. Are patent oppositions equally concentrated? Are some companies more exposed to the oppositions filed by direct competitors or more prone to challenge the patents held by industry rivals?

Although based on narrowly defined industries, the studies that employ data for individual companies (performing both the roles of attacked patent holders and challengers) suggest that the opposition procedure is essentially a game between the major industry players (Harhoff 2005; Schneider 2011). However, this does not mean that patent oppositions between the largest companies are undertaken for pure strategic motives, such as that of creating uncertainty and inducing the rivals to delay the commercialization of innovations.

In fact, on average, only about 30% of the oppositions before the EPO ended with a patent maintained as granted, that is, with a rejection of the opposition. The prevalent outcome of the opposition proceedings has been the revocation (34%) followed by the amendment (32%) of issued patents. The residual outcome (circa 4% of cases) corresponds to oppositions that were “closed” because the opposition was withdrawn or the patent lapsed because the patent holders stopped to pay the renewal fees. It should be stressed that the percentage of opposition rejected is quite stable over time, while in some years (such as in 2012 and 2103, according to the last

annual reports of the EPO), the share of patents upheld in amended form has been above that of revoked patents. In any case, the fact that around 66% of the opposed patents end with a more or less severe reduction of the property rights of patentees clearly indicates that the opponents are more successful than the defendants of granted patents.

Concluding Remarks and Future Directions

The above findings confirm that the opposition procedure adopted by the EPO is particularly effective in correcting the errors made in the first examination process, improving the quality of granted patents, and, then, reducing the chances of further litigation. In the absence of an effective post-grant review, the only way to fix the errors of patent offices would be that of challenging the patents' validity before national courts. However, as previously stressed, the cost of a patent lawsuit is much higher than that required to pursue an opposition case before the EPO. As a consequence, this kind of administrative patent review, by reducing the scope for further and more expensive litigation, is improving social welfare.

Especially on the basis of these considerations, many scholars have contended that also in the USA, a more effective post-grant patent review, resembling that adopted by the EPO, should be introduced (Graham et al. 2003; Hall and Harhoff 2004; Farrell and Merges 2004; Graham and Harhoff 2006). The Leahy-Smith America Invents Act, enacted into law in 2011, has introduced new procedures that have a limited duration and expand the bases for challenging the patents issued by the United States Patent and Trademark Office (USPTO). These new proceedings differ from the previous USPTO reexaminations which, contrary to the EPO oppositions, have been used by far less frequently than invalidity suits before civil courts. Also in Japan, in the light of the scanty use of the invalidation system (the only means, at present, for obtaining a patent revocation), the Japan Patent Office will probably reintroduce the procedure of post-grant opposition abolished in 2003.

With respect to the important changes that are coming in the European patent system, a final consideration is in order. The unitary patent, valid for the EU countries that signed the agreement, entered into force on 20 January 2013. However, it will apply only after the entry into force of a parallel agreement on a Unified Patent Court. For the purpose of our topic, the establishment of a unified court is important because it will have exclusive jurisdiction for litigation concerned not only with the new unitary patents but also with the current European patents. Both of them will be managed by the EPO, while users will be free to opt for one of the two systems. As a consequence, the availability of unitary patents will not change the procedures that the EPO currently adopts, including the opposition and appeal proceedings. However, the Unified Patent Court will reduce the excessive costs and, especially, the risk of divergent judicial decisions which, at present, make the recourse to patent litigation in Europe particularly burdensome and unlikely. Although not in the near future, this desirable institutional change could diminish the attitude to challenge the patents' validity by mainly resorting to the oppositions before the EPO.

References

- Archontopoulos E, Guellec D, de la Van Pottelsberge PB, Van Zeebroeck N (2007) When small is beautiful: measuring the evolution and consequences of the volume of patent applications at the EPO. *Inf Econ Policy* 19:103–132
- Caviggioli F, Scellato G, Ughetto E (2013) International patent disputes: evidence from oppositions at the European Patent Office. *Res Policy* 42:1634–1646
- Cincera M (2011) Déterminants des oppositions de brevets. Une analyse microéconomique au niveau belge. *Rev Econ* 62:87–99
- Cremers K, Schliessler P (2015) Patent litigation settlement in Germany: why parties settle during trial. *Eur J Law Econ*. 40:185–208
- Cremers K, Ernicke M, Gaessler F, Harhoff D, Helmerts C, McDonagh L, Schliessler P, van Zeebroeck N (2013) Patent litigation in Europe. ZEW discussion paper no 13–072. <http://ftp.zew.de/pub/zew-docs/dp/dp13072.pdf>
- de Saint-Georges M, de la van Pottelsberghe PB (2013) A quality index for patent systems. *Res Policy* 42:704–719

- EPO (2006) Assessment of the impact of the European patent litigation agreement (EPLA) on litigation of European patents. Report of the European Patent Office acting as secretary of the Working Party on Litigation. http://www.eplaw.org/Downloads/EPLA_Impact_Assessment_2006_.pdf
- EPO (2013) Patent litigation in Europe, 3rd edn. European Patent Academy, Munich
- Farrell J, Merges P (2004) Incentives to challenge and defend patents: why litigation won't really fix patent office errors and why administrative patent review might help. *Berkeley Technol Law J* 19:1–28
- Graham S, Harhoff D (2006) Can post-grant reviews improve patent system design? A twin study of US and European patents. CEPR discussion papers no 5680. <http://www.cepr.org/pubs/dps/DP5680.asp>
- Graham S, van Zeebroeck N (2014) Comparing patent litigation across Europe: a first look. *Stanf Technol Law Rev* 17:655–708
- Graham S, Hall B, Harhoff D, Mowery D (2003) Patent quality control: a comparative study of US patent reexaminations and European patent oppositions. In: Cohen W, Merrill S (eds) *Patents in the knowledge-based economy*. The National Academic Press, Washington, DC, pp 74–119
- Hall B, Harhoff D (2004) Post-grant reviews in the U.S. patent system – design, choices and expected impact. *Berkeley Technol Law J* 19:989–1015
- Harhoff D (2005) The battle for patent rights. In: Peeters C, de la van Pottelsberghe PB (eds) *Economic and management perspectives on intellectual property rights*. Palgrave-Macmillan, London, pp 21–39
- Harhoff D, Reitzig M (2004) Determinants of oppositions against EPO patent grants: the case of biotechnology and pharmaceuticals. *Int J Ind Organ* 22: 443–480
- Harhoff D, Scherer F, Vopel K (2003) Citations, family size, opposition and value of patent rights. *Res Policy* 32:1343–1363
- Kingston W (2004) Making patents useful to small firms. *Intellect Prop Q* 4:369–378
- Lanjouw J, Schankerman M (2001) Characteristics of patent litigation: a window on competition. *RAND J Econ* 32:129–151
- Lanjouw J, Schankerman M (2004) Protecting intellectual property rights: are small firms handicapped? *J Law Econ* 48:45–74
- Lazaridis G, de la van Pottelsberghe PB (2007) The rigour of EPO's patentability criteria. *World Patent Inf* 29:317–326
- Mejer M, de la van Pottelsberghe PB (2012) Economic incongruities in the European patent system. *Eur J Law Econ* 41:215–234
- Rotstein F, Dent C (2009) Third-party patent challenges in Europe, the United States and Australia: a comparative analysis. *J World Intellect Prop* 12: 467–499
- Schettino F, Sterlacchini A (2009) Reaping the benefits of patenting activities: does the size of patentees matters? *Ind Innov* 16:613–633

- Schneider C (2011) The battle for patent rights in plant biotechnology: evidence from opposition filings. *J Technol Transfer* 36:565–579
- van Zeebroeck N (2011) The puzzle of patent value indicators. *Econ Innov New Technol* 20:33–62

Path-Dependent Rule Evolution

Jan Schnellenbach

Brandenburgische Technische Universität
Cottbus-Senftenberg Chair for Microeconomics,
Cottbus, Germany

Definition

Path-dependent rule evolution occurs whenever the further change of formal or informal institutions is, at least to some degree, determined by the institutional history of a system.

How rules emerge and change

Different types of rules influence individual behavior. There are formal institutions, such as laws or self-adopted written rules of organizations (Furubotn and Richter 2005); there are informal institutions that are not captured in written form, such as social norms (Young 2008); and there are also habits or routines (Hodgson 2010; Vanberg 2002) that individuals themselves follow. A decision to implement and to follow such rules can be made consciously, but they can also evolve without any individual making a deliberate choice to change them. In any case, the evolution of rules is often path dependent.

Path dependence exists, simply put, when past events and decisions have an influence on and limit the scope of the future evolution of a system (David 2005). A simple example is the decision-making of individuals who have a preference for social approval and who attempt to infer from peer actions what the socially desired behavior is. At the initial stage, let there be a range of behaviors with roughly similar individual payoffs, so social

approval dominates the choice between them. In that case, it can happen that, given enough time, a vast majority of individuals coordinate on one type of behavior (Arthur 1994), which henceforth works like a social norm (Young 2008). It is important, however, that from the ex ante perspective, there were multiple possible equilibria, that is, various different kinds of social norms that individuals could have settled on. Path dependence implies that small differences in the early stages of the process, such as individuals randomly observing one kind of behavior rather than another, can have huge effects on the question which equilibrium is eventually chosen.

From an efficiency-oriented point of view, the biggest problem is that the equilibrium of a path-dependent selection process may not be efficient (David 1985), simply because chance and other individual motives than efficiency have driven the process. It may turn out that a different kind of norm would be, for example, associated with lower transaction costs or a more efficient utilization of technology. However, path dependence often leads to a lock-in (Arthur 1989) where, given the status quo social norm, there are little individual-level incentives to deviate from the norm. A change in the social norm therefore requires a widespread change of individual expectations regarding the desired behavior. This can occur as a result of political interventions, but also through decentralized processes such as social communication. If the latter occurs, the result often appears on the surface as a relatively sudden tipping from one social norm to another (Young 1998).

Some skepticism is however due with regard to the likelihood of efficient, deliberate changes of both informal social norms and formal laws through the political process. Path dependence in opinion-formation can lead to equilibria where publicly voiced political opinions, which are deemed false by a large majority of individuals, nevertheless dominate political discourse (Kuran 1995). Similarly, a majority of individuals can be easily locked-in believing factually false policy-related beliefs to be true and refusing to update them (Schnellenbach 2004). The fact that policy-making rests not on individual but shared beliefs

(Denzau and North 1994; Bischoff and Siemers 2013), with very limited incentives for individuals to invest into holding factually correct beliefs, is therefore one factor that leads to a frequently observed persistence of inefficient rules.

The evolution of the rules that govern society is therefore to be understood as an interdependent process where informal institutions affect the evolution of formal institutions and vice versa (North 2005). In order to overcome inefficient lock-ins, a careful institutional policy-making is necessary, although not always sufficient (Eggertsson 2005).

References

- Arthur WB (1989) Competing technologies, increasing returns and lock-in by historical events. *Econ J* 97:642–665
- Arthur WB (1994) Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbor
- Bischoff I, Siemers LHR (2013) Biased beliefs and retrospective voting. *Pub Choice* 156:163–180
- David PA (1985) Clio and the economics of QWERTY. *Am Econ Rev (P&P)* 75:332–337
- David PA (2005) Path dependence in economic processes: implications for policy analysis in dynamical system contexts. In: Dopfer K (ed) *The evolutionary foundations of economics*. Cambridge University Press, Cambridge, pp 151–194
- Denzau AT, North DC (1994) Shared mental models: ideologies and institutions. *Kyklos* 47:3–31
- Eggertsson T (2005) Imperfect institutions: possibilities and limits for reform. University of Michigan Press, Ann Arbor
- Furubotn EG, Richter R (2005) *Institutions and economic theory*, 2nd edn. University of Michigan Press, Ann Arbor
- Hodgson G (2010) Choice, habit and evolution. *J Evol Econ* 20:1–18
- Kuran T (1995) Private truths, public lies. The social consequences of preference falsification. Harvard University Press, Cambridge, MA
- North DC (2005) *Understanding the process of economic change*. Princeton University Press, Princeton
- Schnellenbach J (2004) The evolution of a fiscal constitution when individuals are theoretically uncertain. *Eur J Law Econ* 17:97–115
- Vanberg VJ (2002) Rational choice vs. program-based behavior. *Ration Soc* 14:7–54
- Young HP (1998) *Individual strategy and social structure*. Princeton University Press, Princeton
- Young HP (2008) Social norms. In: Durlauf SN, Blume LE (eds) *The New Palgrave dictionary of economics online*, 2nd edn. Palgrave Macmillan, Basingstoke

Permit Trading

► Emissions Trading

Piracy, Modern Maritime

Paul Hallwood and Thomas J. Miceli
Department of Economics, University of
Connecticut, Storrs, CT, USA

Abstract

This entry provides an economic analysis of the problem of modern-day maritime piracy by first reviewing the current scope of the problem and then developing an economic model of piracy that emphasizes the strategic interaction between the efforts of pirates to locate potential targets and shippers to avoid contact. The model provides the basis for deriving an optimal enforcement policy, which is then compared to actual enforcement efforts that, for a variety of reasons, have largely been ineffectual. The entry concludes by reviewing the law of maritime piracy and by offering some proposals for improving enforcement.

Synonyms

Hijacking; Looting; Robbery

Definition

An act of robbery or criminal violence committed at sea.

This entry develops an economic approach to the problem of modern-day maritime piracy with the goal of assessing the effectiveness of remedies aimed at reducing the incidence of piracy. To date, these efforts have been largely ineffectual for several reasons, including gaps in domestic laws, reluctance of countries to bear the expense of imprisoning pirates, and the general lack of an

effective international legal framework for coordinating and carrying out enforcement efforts. Indeed, it is the absence of such a framework that bedevils international public law as a whole, not just in the area of maritime piracy.

The theoretical framework is based on a standard Becker-type model of law enforcement (Becker 1968; Polinsky and Shavell 2000), extended to consider the effort level of pirates to locate and attack target vessels and of shippers to invest in precautions to avoid contact. The model provides the basis for prescribing an optimal enforcement policy whose goal is to minimize the cost of piracy to international shipping. It also serves as a benchmark for evaluating actual enforcement efforts within the context of international law (such as it exists). The entry concludes with several proposals aimed at improving enforcement.

Modern-Day Maritime Piracy

Modern-day maritime piracy is a worldwide phenomenon. Over 2,600 attacks, actual or attempted, were reported over the period 2004–2011, but with some recent decline due to the effort of naval task forces as well as a very large increase in the use of onboard armed guards. For example, the most recent data shows that in the first 11 months of 2013, there were 234 boardings worldwide, with 30 in Nigerian waters and 13 in Somali waters. Many incidents have also been reported in Southeast Asia, especially off Indonesia.

Somali pirates principally operate a capture-to-ransom model, with ransoms of up to \$5.5 million per incident being collected. Elsewhere in the world, robbery is the main motive. The overall economic cost of maritime piracy in 2012 was estimated at \$6 billion, down from \$7 billion the year before and as much as \$16 billion a few years earlier. Spending on onboard security equipment and armed guards increased from about \$1 billion to \$2 billion between 2011 and 2012. Other economic costs include additional travel days as a consequence of rerouting of ships; increased insurance costs of as much as \$20,000 per trip; increased charter rates, as longer time at sea

reduces the availability of tankers; cost of faster steaming through pirate-affected seas; and greater inventory financing costs for cargoes that remain longer at sea (Bowden 2010). Also, according to, an additional 10 attacks are associated with an 11% decrease in exports between Asia and Europe at an estimated cost of \$28 billion.

With regard to antipiracy efforts, Anderson (1995) notes that there are economies of scale in this activity. When trade on a given shipping route is sparse, individual merchant ships have to arm themselves, thereby duplicating investment. However, with greater amounts of trade, several shipping companies may reduce costs by hiring armed ships for their protection as they sail in convoy. And with still greater shipping traffic, the least cost protection method has turned out to be patrolling of large areas of ocean space by warships. Today all three methods are used.

Not surprisingly, the efficiency of the pirate organization contributes to its success, both historically and in modern times (Leeson 2007; Psarros et al. 2011). Accordingly, present-day Somali pirates have developed supportive “social” organizations that aid them on land and at sea (Bahadur 2011). Pirate leaders often require new recruits to swear allegiance to the organization and its leaders until death; many Somali pirates are ex-coast guardsmen or ex-militiamen and share a common background and training; there is a common belief that ransoms are like a tax on foreigners who are overfishing Somali waters; and there is even the use of stock exchanges to finance operations.

An Economic Model of Piracy

This section develops a simple model of maritime piracy that focuses on its harmful effects on shipping (Guha and Guha 2010; Hallwood and Miceli 2013). The model accounts for both the efforts of pirates to locate potential target ships and of shippers to avoid contact with pirates. In this sense, the model extends the standard economic model of crime to account for precautionary behavior of potential victims (Shavell 1991; Hylton 1996).

After deriving the equilibrium of the model, we examine optimal enforcement policies.

The model focuses on a representative pirate and a representative shipper who traverse the same geographic area over a fixed period of time. The pirate devotes effort x (measured in dollars) to locate a target vessel, and the shipper invests precaution y (also in dollars) to avoid contact. The pirate’s effort represents the amount of time at sea and/or the number of boats, while shipper’s avoidance can represent the use of alternate (more expensive) routes, less frequent or fewer voyages, or the use of armed escorts. Let $q(x, y)$ denote the probability of a contact over a given time period, where $q_x > 0$, $q_{xx} < 0$, $q_y < 0$, and $q_{yy} > 0$. Thus, pirate effort increases the chances of an encounter, while shipper precaution reduces the chances, both at decreasing rates. The cross partial q_{xy} may be positive or negative, as discussed in more detail below. (A common formulation is $q(x, y) = x/(x + y)$.)

The benefit to the pirate from an encounter is the loot, which can take the form of confiscated cargo, ransom of passengers, or both. Let G be the gross expected gain from an encounter. The net gain, however, must account for the possibility of capture and punishment. Let p be the probability of capture and s the (dollar) sanction upon conviction, both of which the pirate takes as given. Thus, the net gain per encounter is $G - ps$, which we will assume is positive. (This will necessarily be true if $G > \bar{s}$, where $\bar{s}$ is the maximal sanction. We discuss the nature of s in greater detail in the section on enforcement below.) At the time it makes its choice of effort, the pirate’s expected gain is therefore

$$q(x, y)(G - ps) - x. \quad (1)$$

The pirate chooses x to maximize this expression, taking as given y , p , and s . The resulting first-order condition is

$$q_x(G - ps) - 1 = 0, \quad (2)$$

which defines the pirate’s reaction function, $\hat{x}(y, ps)$.

The shipper expects to earn gross profit of π , which will be reduced by any expected costs associated with the threat of piracy. These costs include the losses inflicted directly by the pirate, denoted h (including the loss of cargo as well as damage to the ship and harm to crew members), plus the cost of avoidance actions, y . The net expected return to the shipper is therefore

$$\pi - q(x, y)h - y. \quad (3)$$

The shipper chooses y to maximize Eq. 3, taking x as given. This yields the first-order condition

$$q_y h + 1 = 0, \quad (4)$$

which defines the shipper's reaction function, $\hat{y}(x)$.

The Nash equilibrium occurs at the point where the reaction functions intersect. Differentiating Eq. 2 yields the slope of the pirate's reaction function

$$\frac{\partial \hat{x}}{\partial y} = \frac{-q_{xy}}{q_{xx}}, \quad (5)$$

which has the sign of q_{xy} given $q_{xx} < 0$, while differentiating Eq. 4 yields the slope of the shipper's reaction function

$$\frac{\partial \hat{y}}{\partial x} = \frac{-q_{xy}}{q_{yy}}, \quad (6)$$

which has the opposite sign of q_{xy} given $q_{yy} > 0$. The equilibrium, which we assume exists and is unique, is shown graphically in Fig. 1.

The Impact of Antipiracy Laws

Enforcement laws against piracy involve efforts to capture and punish pirates. Below we discuss the implementation of these laws in practice; here we examine their impact in theory, given the preceding equilibrium.

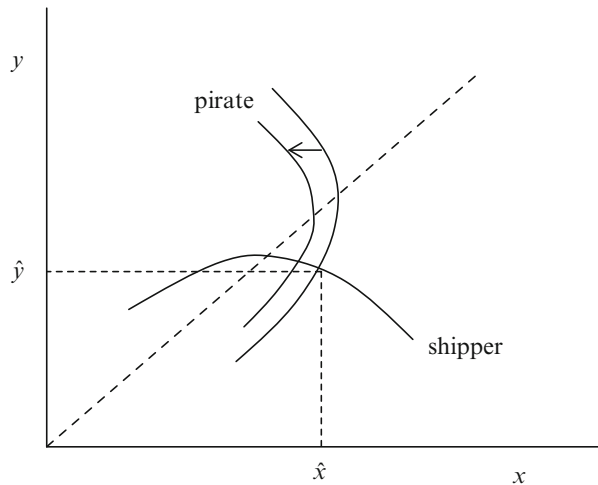
Law enforcement directly affects the behavior of pirates through the expected punishment term, ps , while it indirectly affects shipper behavior through their response to the resulting change in pirate behavior. Consider first the effect of changes in ps on the behavior of pirates. (Note that, given risk neutrality, it does not matter whether this is due to a change in p , s , or both.) Differentiating Eq. 2 yields

$$\frac{\partial \hat{x}}{\partial ps} = \frac{q_x}{q_{xx}(G - ps)} < 0, \quad (7)$$

given $G - ps > 0$. Thus, an increase in the expected sanction for piracy reduces the pirate's investment in effort for any y . In Fig. 1, this results in a leftward shift of the pirate's reaction curve. The new equilibrium involves an unambiguous reduction in the pirate's equilibrium level of

Piracy, Modern Maritime,

Fig. 1 Equilibrium choices of pirate effort (x) and shipper avoidance (y)



effort, but the effect on the shipper's investment in avoidance is ambiguous. As drawn, $\hat{y}$ goes up, but it should be apparent that it could also go down, depending on the location of the initial equilibrium and the amount that the pirate's reaction function shifts. The intuitive reason for these effects is as follows.

The negative effect of greater enforcement on the pirate's effort reflects the standard deterrence argument – a higher expected sanction lowers the marginal benefit of criminal activity. The ambiguous effect of enforcement on the shipper's precaution hinges on the sign of q_{xy} . For the case shown in Fig. 1, $q_{xy} > 0$, so as the pirate's effort declines, the marginal benefit of shipper precaution increases (i.e., q_y becomes more negative), causing an increase in y . In this case, enforcement of laws against piracy and shipper precaution are complementary. However, the reverse would be true if $q_{xy} < 0$, for in that case, greater law enforcement efforts, by lowering x , would substitute for, or “crowd out,” shipper precaution. The actual outcome is therefore an empirical question.

Given the preceding effects of increased enforcement, we now turn to the derivation of the optimal enforcement policy, which involves the enforcement authority (whoever that may be) choosing the probability of apprehension, p , and the sanction, s , to maximize social welfare. An important question here concerns whether or not to count the pirate's gains as part of welfare. The convention in the economics of crime literature has been to count the offender's gains, but there are differing views on this issue (Polinsky and Shavell 2000, p. 48). In the case of pure theft, the value of the loot is simply a transfer payment and thus would drop out of welfare if the thief's gains are counted (Shavell 1991). However, if the gains and losses differ, the possibility arises that the transfer could actually be value enhancing – an “efficient theft” – which most people would find objectionable, especially in the piracy context. Thus, although we will follow the standard convention and count the pirate's gains in welfare, we will assume that the loss suffered by the shipper exceeds the pirate's gains – that is, $h > G$. Consequently, any act of piracy is necessarily inefficient. This could reflect damages or harm to victims on

top of the simple transfer of wealth, as well as any fear or “pain and suffering” incurred by victims of piracy and their sympathizers.

Based on these considerations, we write social welfare as

$$W = \pi - q(\hat{x}(ps), \hat{y}(ps))(h - G + p\beta s) - \hat{x}(ps) - \hat{y}(ps) - c(p), \quad (8)$$

where $\hat{x}(ps)$ and $\hat{y}(ps)$ are the equilibrium levels of pirate effort and shipper precaution, which depend on ps in the manner described above. The total expected enforcement costs are $c(p) + q(\hat{x}(ps), \hat{y}(ps))p\beta s$, where $c(p)$ is the cost of deploying more ships ($c' > 0$, $c'' \geq 0$) and β is the unit cost of increasing s . The enforcement problem is to choose p and s to maximize Eq. 8, subject to $p \in [0, 1]$ and $s \in [0, \bar{s}]$, where $\bar{s}$ is the maximal sanction. The possible interpretations of $\bar{s}$ are (i) the maximum prison term the offender could serve (e.g., life), (ii) a death sentence, or (iii) the harshest punishment that the country charged with carrying out the punishment is willing to impose (as discussed further below).

We begin by deriving a standard result in the law enforcement literature – namely, that $s^* = \bar{s}$, or the optimal sanction is maximal (Polinsky and Shavell 2000). To see why, suppose that $s < \bar{s}$ and $p > 0$. Now raise s and lower p so as to hold ps fixed. As a result, all of the terms in Eq. 8 that depend on ps remain unaffected, but $c(p)$ falls, thus raising welfare. This proves that $s < \bar{s}$ could not have been optimal. The intuition for this result is that the cost of s is only incurred if a pirate is actually captured, so overall costs are lowered by capturing only a few offenders and punishing them harshly.

With $s^* = \bar{s}$, we differentiate Eq. 8 with respect to p and, after canceling terms using Eqs. 2 and 4, obtain the following first-order condition for p^*

$$\begin{aligned} & -q_x \frac{\partial \hat{x}}{\partial p} (h + p\beta \bar{s}) + q_y \frac{\partial \hat{y}}{\partial p} (G - p\beta \bar{s}) \\ & = c' + q\beta \bar{s}. \end{aligned} \quad (9)$$

The left-hand side is the marginal benefit of increased enforcement, while the right-hand side

is the marginal cost. The first term on the left-hand side, which is positive, is the saved costs (victim harm plus punishment costs) as pirates reduce their efforts in response to an increase in the probability of apprehension. This is the direct gain from deterrence. The second term, which is ambiguous in sign, reflects the uncertain effect of an increase in p on shipper effort. Suppose $\partial \hat{y} / \partial p > 0$ (as is the case in Fig. 1), indicating that shippers increase their precaution in response to greater p (i.e., public enforcement and private precaution are complements). Let us also suppose that the term $(G - p\beta\bar{s})$ is negative, as would be true if pirate gains are not counted in welfare. In that case, the overall term is positive (given $q_y < 0$), thus amplifying the marginal benefit of enforcement. Intuitively, when public enforcement elicits increased private precautions, p should be raised, all else equal to encourage such precaution. Conversely, if $\partial \hat{y} / \partial p < 0$, the case of crowding out, the second term on the left-hand side is negative, which works in the direction of reducing p so as not to overly discourage private precaution by potential victims.

Enforcement Problems

The preceding represents the optimal enforcement policy in an ideal setting where there exists a single enforcement authority (or a unified coalition of enforcers), possessing both the will and the resources to carry out the policies implied by Eq. 9. While this may represent a reasonable assumption in many law enforcement contexts, enforcement of international laws against piracy is undertaken by multiple countries with varying degrees of interest in devoting resources to the effort. As a result, enforcement involves a problem of collective action, which may lead to several departures from the prescribed policy.

First, the gains from deterring piracy are enjoyed by all countries who make use of the shipping lanes threatened by pirate attacks. Thus, each country has an interest in reducing piracy in proportion to its expected losses. At the same time, however, deterrence of pirate attacks is a public good in the sense that actions by any one country

to invest in enforcement will benefit all countries. Thus, each country has an incentive to free ride on the enforcement effort of others. Absent some form of credible commitment, therefore, those countries with the largest stake (e.g., the highest value of shipping in the affected area) will undertake the bulk of the enforcement, and all other countries will free ride on that effort. Actual enforcement will therefore be less than the efficient level.

A second factor discouraging enforcement efforts concerns the expected cost of imposing punishment once a pirate is apprehended. If this cost is borne entirely by the country that first apprehends the pirate, then enforcers will likely underinvest in an effort to reduce their probability of incurring that cost. This represents a kind of “reverse rent-seeking” problem in which individual countries underinvest in order to lower the chances that they will be the first to catch the pirate. Note that both of the above problems, which arise from the collective nature of enforcement of piracy laws, will arise in any law enforcement context involving overlapping or undefined jurisdictional boundaries. For example, similar problems plague the enforcement of laws against international drug trafficking (Naranjo 2010) and prosecution of the global war on terror.

A third enforcement problem concerns the credibility of threats to actually impose any punishment at all on a band of pirates once they are captured. Since the pirates’ harmful acts are sunk by the time they are apprehended, enforcers may lack adequate incentives to incur the high costs of detention, trial, and final punishment. Although there may be incapacitative benefits of detention, the probability of any particular pirate committing further harmful acts is small compared to the high cost of punishment. As a result, it may be optimal (in a time-consistent sense) simply to release him. This issue is largely ignored in the economics of crime literature, where it is generally assumed that threats to prosecute and punish criminals are taken as given. The issue is amplified in the piracy context because of the absence of a well-established international tribunal that can develop a reputation over time for carrying out threatened sanctions.

A final problem concerns the choice of the sanction s . As the model showed, the optimal

sanction is maximal, but countries may interpret this prescription differently based on constitutional or other considerations, or they may set *s* based on criteria that differ from that described above (which, if they sympathize with the pirates, could involve setting no punishment at all). As a result, pirates will not be able to predict with any accuracy the actual penalty upon conviction, thereby diluting the deterrent effect of greater enforcement. Countries may also differ in their criminal procedures and evidentiary standards. Although countries can theoretically agree by treaty to uniform standards on these matters, philosophical differences regarding appropriate measures (e.g., based on disagreements over the appropriateness of the death penalty or sympathies for pirates) will make this difficult in practice.

International Law Governing Maritime Piracy

This section evaluates the efficacy of international law in light of the preceding analysis. Piracy is a crime under customary international law and is codified as such in the United Nations *Convention on the Law of the Sea* (UNCLOS) (ratified in 1994). Under this *Convention* states parties agreed to “cooperate” in policing the oceans outside of territorial waters and to arrest, prosecute, and imprison persons suspected and ultimately found guilty of piracy (Articles 100–107). In fact, these articles were taken verbatim from Articles 14–21 of the *Convention on the High Seas*, which was put into force in 1962.

The evidence on actual enforcement of international laws against maritime piracy, as defined by UNCLOS, suggests that these laws have largely been ineffective. For example, over the period between August 2008 and September 2009, Combined Task Force 151 and other navies in the Horn of Africa region disarmed and released 343 pirates, while only 212 others were handed over for prosecution (Ungoed-Thomas and Woolf 2009). The UN Security Council likewise reports that 90% of apprehended Somali pirates were released (UN Security Council 2011).

The discussion in the previous section suggests why this is the case: policing and enforcement is

a public good or at least a mixed good with external benefits for third parties. There are, however, some other considerations as well. The first simply concerns those acts that meet the definition of piracy under the *Convention*. Acts must be for “private ends,” suggesting that they must be motivated by the desire for material gain rather than for political purposes. Thus, terrorist acts would not meet the definition of piracy (Bendall 2010, p. 182) nor would hijacking or acts involving “internal seizure” of a ship by its crew or passengers (mutiny) under the so-called two-vessel requirement for piracy (Hong and Ng 2010, pp. 54–55).

A second difficulty, as discussed above, is the overlapping jurisdiction problem. UNCLOS only applies to acts of piracy on the high seas and in the 200-mile exclusive economic zones, and enforcement relies on the cooperation of all member states. Enforcement in the 12-mile territorial waters is the responsibility of the coastal state, and states vary both in their definitions of piracy and in the availability of resources or the will to enforce anti-piracy laws (Hong and Ng 2010, p. 55; Dutton 2010). Pirates will therefore naturally gravitate toward those areas where enforcement efforts are low or where antipiracy laws are weak. Of course, shippers will also avoid those areas (though at a cost of rerouting), so states with weak laws will suffer economic costs. However, because shipping lanes inevitably cross jurisdictional boundaries, some of those costs will be externalized.

A third problem with enforcement of international law, mentioned earlier, is the problem of successfully prosecuting those pirates who have been apprehended. Article 105 permits the apprehending state to prosecute offenders, but this has often been difficult both politically and logistically. For example, Fawcett (2010) notes that problems of transporting defendants and evidence gathering are significant impediments.

However, in Southeast Asia there has been some success in cooperation against piracy under the *Regional Cooperation Agreement on Combating Piracy and Armed Robbery against Ships in Asia*, which has been functioning since 2006 and now has 17 contracting parties (Noakes 2009). Under this agreement, the parties share

information and perform antipiracy patrols, especially in the Straits of Malacca. What may have helped in this instance is that in this region, there is relatively little area of high seas (none in the Straits of Malacca), and so policing is largely restricted to waters over which sovereign rights exist. As a result, the benefits of enforcement against piracy are more concentrated on the enforcing country, and hence there is less of a public good problem.

Proposals to Improve Enforcement

This section offers three proposals to improve the enforcement of antipiracy laws. The first, suggested by Dutton (2010), involves putting suspected pirates on trial in the International Criminal Court (ICC) rather than in the national court of the apprehending party. The ICC was created by the Rome Statute, which was ratified in 2002 and has 110 signatories, all of whom share in its costs according to an agreed-upon formula (Romano and Ingadottir 2000). However, while the Rome Statute grants the ICC jurisdiction over war crimes, crimes against humanity, genocide, and aggression, at present the ICC has jurisdiction only over the first three of these and that it will not be until 2017 that it can exercise jurisdiction over the crime of “aggression,” which still has to be defined in law but under which piracy could conceivably be classified. Another difficulty with using the ICC against piracy is that some signatories may decline to finance the court for this purpose; that is, a states party choosing to free ride under UNCLOS may also wish to do so under a revised Rome Statute.

The second proposal involves extending the *Convention for the Suppression of Unlawful Acts against the Safety of Maritime Navigation* (SUA Convention) to piracy as well as to terrorism. The SUA came into force in 1992 and by 2011 it had 156 signatories and ratifications. This *Convention* is targeted at policing the oceans against criminal activities, though it specifically targets terrorism rather than piracy. The word “terrorism” appears five times in SUA, but the term is never defined, leading some to believe that, with appropriate reinterpretation,

SUA could be used against maritime pirates (Hong and Ng 2010).

Cognizant of these features Noakes (2009), the chief maritime security officer for the Baltic and International Maritime Council (BIMCO) argued before a US House of Representatives Committee that SUA 1988 can and should be used to combat piracy and that it is incorrect to view this *Convention* as applying only to maritime terrorism and not maritime piracy. It is certainly true that SUA 1988 uses the word “terrorism” sparingly, and this could give the impression that it could be used in the context of maritime piracy. However, SUA 1988 grew out of UN General Assembly Resolution 40/61 in 1985, itself being a response to terrorism on the *Achille Lauro*, and Resolution 40/61 is clearly aimed at terrorist acts at sea and not piracy.

Still, Article 3 of SUA defines seven offenses, the first three being described as follows:

Any person commits an offence if that person unlawfully and intentionally: seizes or exercises control over a ship by force or threat thereof or any other form of intimidation; performs an act of violence against a person on board a ship if that act is likely to endanger the safe navigation of that ship; or destroys a ship or causes damage to a ship or to its cargo which is likely to endanger the safe navigation of that ship.

Although an authoritative legal opinion by each signatory states party regarding exactly what crimes at sea the SUA encompasses has yet to be given, the US legislative attorney, R. C. Mason, working for the Congressional Research Service, has suggested, based on Article 3, that the SUA is directed at piracy as well as terrorism at sea (Ploch et al. 2010). However, this is only “guidance” and at present the US position on SUA and piracy remains unresolved.

Kilpatrick (2011) offers a third proposal, arguing that the UN Hague Convention (1970) could be extended from international civil aviation to maritime piracy. This Convention has been widely adopted, with 185 signatories as of 2013, and it compels states to either extradite or prosecute airplane hijackers. It also requires signatory states to punish terrorist acts by “severe penalties” through domestic laws. However, it is questionable that countries will move to extend the Hague

Convention to maritime piracy. This is true for several reasons. First, while the United States is of central importance in global civil aviation, it is much less so in international shipping. In civil aviation, US legislation has a significant impact on global regulation because foreign airlines and flights from foreign airports to the United States that do not meet US security standards are effectively prohibited from accessing its lucrative market. Second, countries are probably more strongly motivated to move against aircraft hijackings because each single incident is likely to affect more people, say, 250 persons on an airplane versus 20 or so on a ship. Finally, aircraft hijackings seem to be given much more prominence in the media than maritime hijackings.

Cross-References

- [Hijacking](#)
- [Ransom Kidnapping](#)

References

- Anderson J (1995) Piracy and world history: an economic perspective on maritime predation. *J World Hist* 6:175–195
- Bahadur J (2011) *The pirates of Somalia: inside their hidden world*. Pantheon, New York
- Becker G (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Bendall H (2010) Cost of piracy: a comparative voyage approach. *Marit Econ Logist* 12:178–195
- Bowden A (2010) *The economic cost of maritime piracy*. Working paper, One Earth Future Foundation
- Dutton YM (2010) *Bringing pirates to justice: a case for including piracy within the jurisdiction of the international criminal court*. Working paper, One Earth Future Foundation
- Fawcett J (2010) Challenges to apprehension and prosecution of East African maritime pirates. *Marit Policy Manag* 37:753–765
- Guha B, Guha A (2010) Pirates and traders: some economics of pirate-infested seas. *Econ Lett* 111:147–150
- Hallwood P, Miceli T (2013) An economic analysis of maritime piracy and its control. *Scott J Polit Econ* 60:343–359
- Hong N, Ng A (2010) The international legal instruments in addressing piracy and maritime terrorism: a critical review. *Res Transp Econ* 27:51–60
- Hylton K (1996) Optimal law enforcement and victim precaution. *Rand J Econ* 27:197–206
- Kilpatrick RL (2011) *Borrowing from civil aviation security: does international law governing airline hijacking offer solutions to the modern maritime piracy epidemic off the coast of Somalia? Oceans beyond piracy working Paper*
- Leeson P (2007) An-arrgh-chy: the law and economics of pirate organization. *J Polit Econ* 115:1049–1094
- Naranjo A (2010) Spillover effects of domestic law enforcement policies. *Intl Rev Law Econ* 30:265–275
- Noakes G (2009) Statement on international piracy. Testimony before the US House of Representatives, Committee on Transportation and Infrastructure, Subcommittee on Coast Guard and Maritime Transportation. February. (www.marad.dot.gov/documents/HOA_Testimony-Giles%20Noakes-BIMCO.pdf)
- Ploch L, Blanchard CM, Mason RC, King RO (2010) Piracy off the horn of Africa. Congressional Research Service, 7–5700, R40528, April. (http://assets.opencrs.com/rpts/R40528_20100419.pdf)
- Polinsky AM, Shavell S (2000) The economic theory of public law enforcement. *J Econ Lit* 38:45–76
- Psarros G, Christiansen AF, Skjong R, Gravir G (2011) On the success rates of maritime piracy attacks. *J Transp Secur* 4:309–335
- Romano C, Ingadottir T (2000) *The financing of the international criminal court: a discussion paper*. Center for International Cooperation, Project on International Courts and Tribunals, June, NYU and School of Oriental and African Studies, University of London
- Shavell S (1991) Individual precautions to prevent theft: private versus socially optimal behavior. *Intl Rev Law Econ* 11:123–132
- Ungoed-Thomas J, Woolf M (2009) Navy releases Somali pirates caught red-handed: a legal loophole has helped scores of Somali gunmen escape justice. *The Sunday Times*, 29 Nov
- United Nations Security Council (2011) Report of the special advisor to the secretary general on legal issues related to piracy off the Somali coast: a plan in 25 proposals. 30 Jan 2011

Piracy, Old Maritime

Peter T. Leeson
Department of Economics, George Mason
University, Fairfax, VA, USA

Abstract

Eighteenth-century pirates were profit-maximizing criminals. Their infamous practices reflect strategies pirates adopted to bolster their bottom line. The “pirate code” was a system of constitutional democracy that sea

dogs developed to govern themselves privately in the absence of government. The “Jolly Roger” – pirates’ black flag of skull and bones – was a signaling device that sea dogs used to facilitate the seizure of prizes without costly conflict. Pirates used heinous torture to develop a reputation that incentivized captives to acquiesce to their demands. And pirates “pressed” – or pretended to conscript – willing recruits to reduce sailors’ legal cost of joining their crews.

Synonyms

Caribbean piracy; Eighteenth-century piracy; Golden Age piracy

Definition

Old maritime piracy refers to the seafaring banditry of early eighteenth-century (c. 1715–1730) criminals. These pirates preyed on merchant ships traveling through the Atlantic Ocean, Indian Ocean, Caribbean Sea, and Gulf of Mexico. The pirate population – at its height, an estimated 2,400 sailors (Rediker 1987, p. 256) – was drawn predominantly from the merchant marine, consisting mostly of ordinary seamen who voluntarily joined a pirate crew’s ranks after the vessel they were sailing on had been seized by its attackers. Old maritime piracy’s most infamous members included Blackbeard (real name, Edward Teach), Bartholomew Roberts, “Calico Jack” Rackham, and at least a few women, such as Rackham’s crewmates, Anne Bonny and Mary Read.

Pirates lived and worked together at sea and on land. When “on the account” – sea dogs’ term for pirating – they operated in independent crews averaging 80 members (Rediker 1987, p. 256). When not at sea, many pirates made homes in New Providence Island, Bahamas – a neglected and thus rogue-infested British colony. Unlike privateers, who were also active in the early eighteenth century and plundered merchant ships for profit, pirates paid allegiance to no government

and lacked the sanction or support of any state. Pirates were considered, and considered themselves, “enemies of all nations.”

The Decision to Pirate

During peacetime, when privateering employment was unavailable and naval employment was difficult to find, most sailors had two employment choices. One possibility was to work in the merchant marine. The pay in this employment was low. Between 1689 and 1740 the average able seaman earned between £15 and £33 a year (Davis 1962, pp. 136–137) – \$4,000–8,800 in current US dollars.

Sailors’ second occupational option was piracy. Piracy was a capital offense throughout eighteenth-century Europe, and pirating was a risky activity. Nevertheless, the potential financial rewards of pirating were sufficient to attract an estimated 4,000 sailors to illicit maritime plunder between 1716 and 1726 (Rediker 1987, p. 30). No data exist that would permit a computation of the average pirate’s wage. However, the evidence that is available suggests that piracy, unlike work as an able seaman, could be extremely lucrative.

In the early eighteenth century, Captain John Bowen’s pirate crew plundered a prize that resulted in a payout of £500 per man (Johnson 1726–1728, p. 480). Several years later the members of Captain Thomas White’s crew each earned £1,200 from its cruise (Johnson 1726–1728, p. 485). In 1720 Captain Christopher Condent’s crew seized a prize that earned each pirate £3,000 (Marx 1996, p. 161). And in 1721 Captain John Taylor’s and Oliver La Bouche’s pirate consort earned an astonishing £4,000 for each crewmember in a single attack (Marx 1996, p. 163). Although more modest prizes were more common, a successful expedition could yield a pirate a sum sufficient to retire. The prospect of immense financial reward that piracy offered was an important lure for sailors who left merchant-marine employment for risky careers as pirates.

A second, and perhaps equally important, consideration for many sailors who chose piracy over

legitimate maritime employment was the work and living conditions available on pirate ships versus those available on merchant ships. For reasons described below, captain mistreatment of ordinary sailors on merchant vessels – financial and physical – was a relatively common feature of merchant-marine employment. In contrast, on pirate ships officer abuse was far less common, rendering life aboard such vessels superior in many sailors' eyes.

Pirates' Profit-Maximizing Strategies

Organization

As criminals, pirates could not rely on state-made institutions of law and order to produce social cooperation among them. Yet to live and work together at sea for months at a time – i.e., for pirating, a necessarily jointly produced activity, to be possible at all – piratical cooperation was essential. Indeed, as persons committed to theft and murder for personal gain, the need for institutions of social order among pirates was likely more pronounced than it was among the members of legitimate societies, most of whom, in contrast to pirates, were not so willing to resort to violence to benefit themselves.

The success many pirate crews enjoyed suggests that pirates overcame the problem of social cooperation they confronted without government. Central to the way pirates accomplished this was their institutional organization grounded in constitutional democracy (Leeson 2007, 2009a). Two chief officers led early eighteenth-century pirate ships – the captain and the quartermaster. The former wielded command in times of battle, when chasing and engaging prey. The latter wielded command at all other times: the quartermaster was in charge of distributing crewmembers' victuals and payment and enforcing pirates' private laws. Pirate crews elected both officers democratically and deposed their officers popularly too.

Pirates' democratic selection of officers contrasts starkly with the selection of officers on early eighteenth-century merchant ships and naval vessels, who were appointed by owners and

governments externally. Likewise, the division of power on pirate ships via the captain and quartermaster contrasts starkly with the centralization of power on period merchant and navy ships, where captains wielded autocratic authority over their crews. Pirates' institutional organization also differed sharply from that of early eighteenth-century European states, whose autocratic forms closely resembled those of legitimate, period sea vessels.

Pirates created their own social rules to prevent conflict and maximize the potential for piratical cooperation, which in turn maximized their crews' illicit profits. Pirates developed a body of private law, sometimes called the "pirate code," but which pirates called "articles of agreement" (Leeson 2009c). Each crew forged its own articles democratically and required the consent of every would-be member to its provisions before he was permitted to join the company.

The specifics of pirate articles differed by crew, but their basic elements were uniform. These included rules prohibiting theft and violence; terms governing crewmembers' compensation; regulations on activities that could generate negative externalities for other crewmembers, such as drinking and smoking; and terms of piratical social insurance, which was provided for crewmembers injured on the job. Pirate articles also established punishments for article violations and provided explicitly for democracy as the governing mechanism for important crew decisions, such as the selection of officers.

Remarkably, pirates enshrined their articles in writing as constitutions. Below is the constitution that governed pirate captain Bartholomew Roberts' crew aboard their ship the *Royal Fortune* (Johnson 1726–1728, pp. 211–212):

- (I) Every Man has a Vote in the Affairs of Moment; has equal Title to the fresh Provisions, or strong Liquors, at any Time seized, and may use them at Pleasure, unless a Scarcity make it necessary, for the Good of all, to vote a Retrenchment.
- (II) Every Man to be called fairly in Turn, by List, on board of Prizes, because, (over and above their proper Share) they were

on these Occasions allowed a Shift of Cloaths: But if they defrauded the Company to the Value of a Dollar, in Plate, Jewels, or Money, Marooning was their Punishment. If the Robbery was only betwixt one another, they contented themselves with slitting the Ears and Nose of him that was Guilty, and set him on Shore, not in an uninhabited Place, but somewhere, where he was sure to encounter Hardships.

- (III) No person to Game at Cards or Dice for Money.
- (IV) The Lights and Candles to be put out at eight a-Clock at Night: If any of the Crew, after that Hour, still remained enclined for Drinking, they were to do it on the open Deck.
- (V) To keep their Piece, Pistols, and Cutlash clean, and fit for Service.
- (VI) No Boy or Woman to be allowed amongst them. If any Man were found seducing any of the latter Sex, and carry'd her to Sea, disguised, he was to suffer Death.
- (VII) To Desert the Ship, or their Quarters in Battle, was punished with Death or Marooning.
- (VIII) No striking one another on board, but every Man's Quarrels to be ended on Shore, at Sword and Pistol.
- (IX) No Man to talk of breaking up their Way of Living, till each shared a 1,000 l. If in order to this, any Man should lose a Limb, or become a Cripple in their Service, he was to have 800 Dollars, out of the publick Stock, and for lesser Hurts, proportionately.
- (X) The Captain and Quarter-Master to receive two Shares of a Prize; the Master, Boatswain, and Gunner, one Share and a half, and other Officers one and a Quarter.
- (XI) The Musicians to have Rest on the Sabbath Day, but the other 6 Days and Nights, none without special Favour.

Constitutional democracy among pirates pre-dates its adoption in the United States by more

than half a century. Equally surprising, pirates' reason for relying on this system of governance echoes the reason given in the *Federalist Papers* – or rather the other way around, since the Federalists did not put pen to paper until decades after pirates put constitutional democracy into practice. According to pirates, their reason for using constitutional democracy was to prevent their officers from abusing the authority that officers' positions of power necessarily conferred on them. Guarding against such abuse was particularly important to pirates, they suggested, given the mistreatment they had suffered in their previous lives as legitimate sailors at the hands of merchant captains who enjoyed near-dictatorial authority over their crews.

Pirate democracy permitted crewmembers to control the behavior of captains through the carrot of election and the stick of deposition. Moreover, the creation of multiple, popularly elected offices aboard pirate ships fostered competition between officers, which encouraged both captains and quartermasters to more faithfully adhere to the will of their crews. Circumscribing officers' authorities in written constitutions helped pirates ensure that their leaders wielded the powers they enjoyed in their crews' interests rather than for leaders' private gain. By making behaviors that constituted transgressions of those authorities explicit and requiring that all crewmembers consent to their terms *ex ante*, pirate constitutions created common knowledge about legitimate and illegitimate officer behaviors within pirate crews and helped coordinate pirates' response to officer behaviors.

The contrasting institutional organization of eighteenth-century pirate ships and merchant ships has an economic foundation (Leeson 2007). Merchant ships confronted a classic owner-employee, principal-agent problem. These vessels and their cargoes were financed by external financiers – wealthy landlubbers who funded commercial voyages. Since merchant-ship owners did not sail on the vessels they invested in, their valuable vessels and cargoes were beyond their eyes or reach when away at sea.

This situation invited sailor opportunism, such as stealing cargo, shirking in work activities, or

even absconding with vessels. To control such opportunism, merchant-ship owners appointed captains who they gave small shares in their ships and endowed with autocratic authority to monitor and financially and physically punish sailors who did not act in their interests. Autocratic captains solved owners' principal-agent problem. But in doing so they enabled captain self-dealing, which, as merchant sailors frequently complained, many captains exploited at their sailors' expense.

Pirates, in contrast, confronted no such owner-employee, principal-agent problem. The reason for this is simple: pirates stole their ships. On pirate ships, crewmembers were both owners and employees – principals and agents. Because of this, pirates did not require autocratic captains to control crewmember opportunism. They still needed leaders who could command in times of battle and administer rules that prevented crewmember conflicts. But in the absence of a divergence between principal and agent interests, pirates could democratically elect their leaders and divide authorities among them without loss, and indeed with substantial benefit: restraining the potential for captain self-dealing. The economic context of eighteenth-century merchant shipping rendered autocracy the efficient institutional organization on merchant ships. The different economic context of eighteenth-century pirating rendered constitutional democracy the efficient institutional organization on pirate ships.

Other Profit-Maximizing Strategies

Pirates employed several other practices to maximize the profitability of their criminal enterprise. Each of these practices is an infamous part of popular pirate lore. One example of this is pirates' black flag of skull and bones, the "Jolly Roger," which pirates used as a signaling device to distinguish themselves from less ominous maritime belligerents who merchant ships might confront and whose advances merchant ships were more likely to resist (Leeson 2009a). A second example is pirates' use of heinous torture against recalcitrant captives and their popular image among contemporaries as fiendish "hair-triggers," both products of pirates' manipulation of their public

image, which pirates used to develop a reputation that encouraged prizes to surrender to them peacefully (Leeson 2010a). A third example is pirates' practice of "pressing" – or pretending to compel crewmembers into their service – a practice pirates used to manipulate laws against piracy that reduced merchant sailors' expected cost of joining pirates' ranks (Leeson 2009b, 2010b).

References

- Davis R (1962) The rise of the English shipping industry in the seventeenth and eighteenth centuries. Macmillan, London
- Johnson C (1726–1728) [1999] A general history of the pyrates Dover, Mineola
- Leeson PT (2007) An-*arrgh*-chy: the law and economics of pirate organization. *J Polit Econ* 115:1049–1094
- Leeson PT (2009a) The invisible hook: the hidden economics of pirates. Princeton University Press, Princeton
- Leeson PT (2009b) The invisible hook: the law and economics of pirate tolerance. *NY Univ J Law Lib* 4:139–171
- Leeson PT (2009c) The myth of the myth of social contract: the calculus of piratical consent. *Public Choice* 139:443–459
- Leeson PT (2010a) Pirational choice: the economics of infamous pirate practices. *J Econ Behav Organ* 76:497–510
- Leeson PT (2010b) Rationality, pirates, and the law: a retrospective. *Am Univ Law Rev* 59:1219–1230
- Marx JG (1996) The pirate round. In: Cordingley D - (ed) *Pirates: terror on the high seas – from the Caribbean to the South China Sea*. Turner Publishing, Atlanta
- Rediker M (1987) Between the devil and the deep-blue sea: merchant seamen, pirates and the Anglo-American maritime world, 1700–1750. Cambridge University Press, Cambridge

Political Competition

Anna Laura Baraldi

Department of Economics, Second University of Naples, Capua, Italy

Abstract

In this entry, we tried to explain the role of political competition in the economy. Political science provided various definitions of

political competition. After analyzing in detail all of those, we mentioned the actors political competition involves. In economic terms, competition in political market has seen as competition in good market. Therefore, economic science studied if this kind of competition is good or bad in terms of economic variables as, for example, growth.

Synonyms

Party competition: definitions, sources, and economic effects

Political Competition: Definitions, Sources, and Economic Effects

Political competition is a complex process at the heart of representative democracy. *It* can be defined as competition for political power, that is, for the right to shape and control the content and direction of public policy.

Broadly speaking, this process involves the interaction of a set of *citizens*, each one with views about the relative desirability of conceivable states of the world; a set of *constitutional rules*; a group of *politicians* who compete in the selection for constitutionally privileged positions within the decision-making process; *political parties* who choose to coordinate their behavior that they see as mutually beneficial; a group of *political representatives* in charge of representing citizens in the decision-making process; a group of *office-holders*, politicians selected under constitutional rules to hold privileged positions in the decision-making process; and a *government*, defined as a unique coalition of office-holders in the most privileged positions in the decision-making process.

The unequal distribution of benefits and costs of living among different individuals and groups is the primary source of political competition in a society. Indeed, the decisions and actions of governments benefit, reward, and advantage only certain areas of the society, while others are compelled to bear more of the costs and burdens.

Therefore, political competition is the continuing controversy over who controls the government and who officially decides how its formal-legal power is to be used.

Under a free and competitive, constitutional democratic political regime, different groups and organizations within the society engage in political competition. This can be for political authority and for political influence.

Competing for political authority means contending for the legitimate right to govern the entire political community (as carried out by the Labour and Conservative Parties in Britain, the Democrats and Republicans in the United States, and the Liberal and Progressive Conservative Parties in Canada); they compete for direct control of the government and for the legitimate right to officially decide how and for what purposes governmental authority is to be used. Each political party seeks to place its own political leaders in the majority of the government, thereby giving them the right to decide for and act in the name of the entire political society, exercising formal-legal authority to make and enforce authoritative, and binding decisions on public policy. The individual candidate, by seeking his/her party's nomination and then election by the voters to a higher government office (such as member of the legislature), is competing for the right to be a formal-legal participant in the processes of authoritative decision-making and action carried out by the government.

Competition for political influence means that *political interest groups* or *pressure groups*, whose members have common interests and views in a single area of public policy, seek to obtain and exercise political influence therein. Their objective is to influence the policy decisions and actions of the government in one or more areas of public policy. An interest group focuses on influencing and shaping public policy in their chosen policy area or areas, rather than attempting to have an impact on public policy in general.

Political competition overlaps two disciplines: economics and political science. The concept of political competition has been developed in the literature of political science and is directly related to the outcome of elections. It is commonly

asserted that the more competitive the parties, the more responsive the political system will be to the desires of the majority. Different interpretations of political competition have been given (Bardhan and Yang 2004) in the relevant literature. According to these interpretations, political competition generates costs and benefits. The definitions of political competition are the following:

1. Political competition as *accountability for incumbents*. This interpretation focuses on the process of political turnover. The intensity of political competition increases when the incumbent leaders can be removed more easily by the public and replaced by their competitors. Political competition affects the behavior of incumbent leaders today via tomorrow's threat of dismissal. The benefits of *accountability for incumbents* are due to the "accountability" of politicians: if the incumbent politician wants to maintain power, their incentive to respond to the public's wishes is stronger when their position is vulnerable. Therefore, more intense political competition makes the incumbent more accountable for his/her actions. This interpretation of political competition also generates costs linked to the strength of the threat of dismissal for the incumbent politician. Actually, when a degree of political competition is too high, the probability of reelection becomes sufficiently unrealistic; thus, the incumbent politician may abandon any hope of reelection in order to extract the maximum rents for himself/herself during his remaining time in office. This then shifts political incentives towards the short term. In an economic sense, this concept of political competition may be an important determinant of corruption of politicians.

Political competition, in the sense of political turnover, may generate another cost. The cost is evaluated in terms of the incumbent politician's incentives to undertake public investments. Although public investments are growth enhancing, they are politically destabilizing: to keep his position secure, the incumbent may decide not to invest, given that investment must be financed by taxes. In this

situation, when political competition is high, the destabilizing effects of public investment weigh more heavily on the incumbent, in many cases leading to a non-investment outcome (Acemoglu and Robinson 2000, 2002).

2. Political competition as *decentralization of political authority*. According to this interpretation, political competition is more intense when political authority is in the hands of a larger number of agents at any given point in time. This kind of political competitiveness has been largely studied in the literature on fiscal federalism. Competition between authorities representing distinct political jurisdictions induces each of them to become more politically efficient. In fact, authorities representing "efficient" political jurisdictions, in terms of limited corruption level or sound economic policy, have the opportunity of attracting mobile resources away from authorities representing inefficient jurisdictions. In these situations, competition between authorities is similar to competition between firms. As in the case for firms, competition between political authorities can generate economic costs whenever externalities are present. The model of fiscal policy is the theoretical framework used to illustrate this point. In this model, decentralized political authorities decide on public spending; they compare the marginal benefits of public spending to only a fraction of the social marginal costs. The reason is that the benefits of public spending are generally concentrated within a particular jurisdiction, or a particular interest group, while the costs are spread out across the whole of society. This leads, in equilibrium, to a depleted pool of public savings (Persson and Tabellini 2000; Drazen 2001).

If we consider the competition in attracting mobile resources, inter-jurisdictional competition may generate potential costs and benefits (Besley and Case 1995). The model considers a multi-jurisdictional framework with heterogeneous elected officials; agents are *immobile*, but they can deduce information about their "types" of local officials by observing the behavior of officials in neighboring

jurisdictions. This leads local officials to be engaged in “yardstick competition,” whereby each recognizes that his performance will be judged in relation to the performance of others. This form of competition can yield benefits to the people because, in order to maximize their probability of reelection, politicians have an incentive to “hide their true colors” (Besley and Case 2003), for example, by restraining taxation. However, this form of competition can also produce costs to the people in situations where this behavior is expected to prove too expensive; politicians may decide to abandon their hope of reelection in order to “seize all they can” from the public then and there.

It is possible that the “market for policy” view is not complete. This can be a result of inter-jurisdictional competition aimed at gaining mobile resources, which are also inclined to distinctive frictions. For example, agreements with coalitions of less mobile constituents are made instead of facing electoral incentives to disregard the intimidations of the mobile (Rodden and Rose-Ackerman (1997)). The distortionary effects can be quite serious when tighter complicity rather than greater numbers are reflected in the political power of the immobile; this is common for oligarchic landed interests.

For a model of the trade-off between local informational advantages of decentralization and the possibility of capture by the local elite, see Bardhan and Mookherjee (2006).

3. Political competition as *electoral competition*. This interpretation of political competitiveness focuses on the conflict between parties and elites to win public support. In this sense, political competition is the competition between essentially identical agents to *acquire* political power or between those already in power. Therefore, it is clearly related to both the first and second interpretations of political competition, which associate political competition with opportunities for political turnover, and decentralization of political authority, respectively.

Now we can highlight some key points characterizing political competition. Firstly, an

increase in political competition (i.e., the number of political parties engaged in electoral competition) can be seen as a sign of the democratization of a society. Secondly, a major part of political literature states that the degree of political competition among political party systems is largely determined by the choice of the electoral system (Cox 1997; Duverger 1954; Lijphart 1994, 1999; Sartori 1976; Taagepera and Shugart 1989). Thirdly, the party system determines the degree of bargaining complexity that may affect government formation and maintenance (De Winter and Dumont 2003; Lijphart 1999; Müller and Strøm 2003; Van Roozendaal 1997) and feature among the determinants of public policy. In this entry, we will also treat political competition from an economic perspective. Since Marshall’s “Principles of Economics” (1890), economists have been trained to believe that market competition maximizes the welfare of the consumers, whereas monopoly and market power create economic rents that make producers better off and consumers worse off. The literature of public choice, followed by political economy scholars, attempted to import this notion into the political market, using economic performance as a benchmark to evaluate the welfare properties of political equilibria (Persson and Tabellini 2000). Stigler (1972) studied analogies between economic and political competition. Political competition, as economic competition, is supposed to be welfare enhancing for the citizens. This is because more vigorous political competition implies the selection of politicians who are better at resisting interest group pressures to obtain transfers financed by distortionary taxation. On the contrary, electoral competition with few parties reduces the degree to which the political system represents the heterogeneous preferences of the electorate (Lipset and Rokkan 1967) and may also reduce the pool of available political talent (Becker 1983). Since modern times, it can be said that political competition enhances economic efficiency and makes a faster income growth possible.

Nevertheless, theoretical literature is not conclusive in stating whether political competition is or is not growth enhancing. In order to analyze how it affects economic performance, such

literature refers to political competition as electoral competition: competition among parties and candidates in elections to obtain public support through votes of citizens (Bardhan and Yang 2004). This concept of political competition seems close to that of accountability for incumbents (Persson et al. 1997): if political competition is intense, the incumbent politician is more accountable for he/she actions in office and he has an incentive to perform well; otherwise, he can be easily removed by the public and replaced with opponents.

If the votes market is considered as a goods market (with politicians competing with each other to win the elections – Becker 1983), the transmission mechanism linking political competition to growth implies that more intense political competition induces incumbent politicians to act in the public interest, maximizing their probability of being reelected (Mulligan and Tsui 2006). Otherwise, when political competition is intense, the electoral base of each party tends to be smaller. In order to cater to their narrow support base, politicians find it expedient to promise pork-barrel policies rather than policies that benefit the electorate as a whole. The resulting policies benefit the supporters of the winning politician, but do not necessarily maximize aggregate welfare (Lizzeri and Persico 2005).

Public Choice literature concentrates on the effect of political competition on government size, according to two interpretations: the interest group theories (Mueller and Murrell 1986) and the fiscal illusion view (Buchanan and Wagner 1977), both assessing a negative relation between political competition and government size. Lipford and Yandle (1990), taking an industrial organization view of the political market, find that the state share of total state and local tax revenues rises with party dominance.

A well-consolidated body of the relevant literature analyzes how political competition affects corruption which, in turn, affects economic performance. The theoretical literature in the field of the relationship between corruption and economic growth is split, for example, in the literature on political competition and corruption. Polo (1998) shows that intense political competition may alter

the form of corrupt behavior. Policy distortions resulting from lobbying activities are likely to be greater when there is little electoral competition. However, when politicians use discretion over the way in which political contributions are spent, greater electoral competition increases the incentive to divert funds for personal use (Damania and Yalcin 2005). Persson et al. (1997) state that intense political competition implies that the incumbent politician is more accountable for his actions in office: either the incumbent has an incentive for good performance or he can be easily removed and replaced (Mulligan and Tsui 2006). Intense political competition may also lead to a low probability of reelection for the incumbent, as for a firm that may lose a share of the market if the latter becomes more competitive; in this case, an incumbent can act in a myopic manner, maximizing rents during his remaining time in office. To sum up, the overall effect of political competition on corruption is complex and difficult to define.

The empirical literature on the effects of political competition on economic growth is still poor and lacks in giving a unanimous answer. Recently, the analysis focusing on the different sources of growth (Pinto and Timmons 2005) has made the effect of political competition on growth unpredictable.

Besley et al. (2010), analyzing the relationship between political competition, economic policy, and economic performance in the United States, show that political competition may positively affect policy and economic growth via the “quality of politicians”. Padovano and Ricciuti (2009) analyze the effect of an institutional reform (i.e., the change in the regional electoral system) on the competitiveness of Italian regional politics. They find evidence of a positive correlation among them for the 15 Italian regions with Ordinary Statutes. Alfano and Baraldi (2012) find an inverted U relationship between the degree of political competition and the growth rate; this result shows the possibility that an “optimal” level of political competition, in terms of growth, may exist. This optimal level resolves the trade-off between political accountability and government instability.

The complexity in the empirical literature on that topic is how to measure political competitiveness among political parties, and a number of measures of political competition have been proposed. Vanhanen (2000, 2003) measures competition by the share of votes captured by “minor” parties in parliamentary elections, while Holbrook and Van Dunk (1993) and Rogers and Rogers (2000) use the “win margin” of the incumbent governor as a measure of competition, and Skilling and Zeckhauser (2002) focus on the length of time a party has been in office. Alfano and Baraldi (2012) construct a “political competition index” equal to one minus a (normalized) Herfindahl index for the political parties in a country, taking the percentage of votes for each political party at elections. All these measures have merits, but they often start from the presumption that some basic democratic structures are in place and they are not invariant to the choice of election rule and do not account well for party structure (Vanhanen 2000).

In order to overcome the drawback that measures of political competition have, Grofman and Selb (2009) have recently identified six properties that any index of competition should have:

1. The measure should be party specific, i.e., it should allow for the possibility that voters of different parties might have different incentives to turn out to vote.
2. For each party, the measure should run from 0 to 1, with 0 indicating situations where voter incentives to turn out are the least and 1 indicating situations where voters incentives to turn out are the greatest.
3. The measure should be summable over all parties to give a weighted average of overall incentives for turnout in a given district. This aggregate measure should, when appropriately normalized, still run from 0 to 1, with 0 again indicating situations where, in the aggregate, voter incentives to turn out are the least and 1 indicating situations where, in the aggregate, voter incentives to turn out are the greatest. The weights should reflect the vote shares of the parties, and aggregation to the legislature as a whole should not be distorted by variation in

district population size as a function of district magnitude.

4. For each party, the maximum value should be obtained if the votes required to win its last seat (s) are such that a vote loss of one vote would convert a win for that seat (those seats) into a loss. The minimum value should be obtained if one candidate/party receives all the votes.
5. The measure should be sensitive to the nature of the voting rule being used. We propose that it should vary with the threshold of exclusion of that rule.
6. For two-candidate plurality elections, the measure should be reduced to a simple function of the difference in vote share between the winner and the loser.

To sum up, neither theoretical nor empirical literature clarifies whether an intense political competition is beneficial for economic growth or not. Nevertheless, the tendency of many democracies is to reduce the number of political parties engaged in electoral competition. A lot of electoral systems are set up to counteract the tendency of parties to multiply. Many parliamentary democracies have thresholds of exclusion that deny representation to parties with a vote share below the threshold (e.g., Germany has a threshold of 5%). With a threshold of exclusion, the number of parties is reduced because parties that anticipate a small vote share do not field candidates. The general point is that several features of electoral systems act to discourage the proliferation of parties. The main conjecture in support of that is that having a large number of parties creates government instability, which is thought to harm economic growth. In fact, when there are many competing parties, the electoral base of each party tends to be smaller. To cater to their reduced support base, politicians find it expedient to promise pork-barrel policies with narrow appeal, rather than policies that benefit the electorate at large. The resulting policies benefit the supporters of the winning politician, but do not necessarily maximize aggregate welfare. For instance, politicians must choose between creating a large or small amount of bureaucracy. The former entails an aggregate deadweight loss but generates some

benefits (jobs, etc.) that can be targeted to supporters; the latter, instead, is efficient but does not allow the politician to target largesse to supporters. Consequently, we should expect a large amount of bureaucracy. The idea is that projects which diffuse benefits (the minimum bureaucracy) are less appealing to office-motivated politicians because these benefits are less targetable and so may be under-provided by the political system. This distortion becomes worse as the number of competing parties increases. The reason is that the smaller the fraction of the electorate as the support base of each party, the higher the gain from targeting a smaller subset of the electorate, and therefore, the temptation for politicians to engage in special interest politics becomes greater.

References

- Acemoglu D, Robinson J (2000) Political losers as a barrier to economic development. *Am Econ Rev Pap Proc* 90:126–130
- Acemoglu D, Robinson J (2002) Economic backwardness in political perspective. NBER working paper no. 8831
- Alfano MR, Baraldi A L (2012) Is there an optimal level of political competition in terms of economic growth? Evidence from Italy. *Eur J Law Econ*, on line first. <https://doi.org/10.1007/s10657-012-9340-5>
- Bardhan P, Mookherjee D (2006) Corruption and decentralization of infrastructure delivery in developing countries. *Econ J* 116:101–127
- Bardhan P, Yang T (2004) Political competition in economic perspective, Department of economics, UCB. Department of Economics, UCB, UC Berkeley. Retrieved from: <http://escholarship.org/uc/item/1907c39n>
- Becker G (1983) A theory of competition among pressure groups for political influence. *Q J Econ* 98:371–400
- Besley T, Case A (1995) Incumbent behavior: vote-seeking, tax-setting, and yardstick competition. *Am Econ Rev* 85(1):25–45
- Besley T, Case A (2003) Political institutions and policy choices: evidence from the United States. *J Econ Lit* 41(1):7–73
- Besley T, Persson T, Sturm D (2010) Political competition and economic performance: theory and evidence from the United States. *Rev Econ Stud* 77:1329–1352
- Buchanan J, Wagner R (1977) Democracy in deficit: the political legacy of Lord Keynes. Academic, New York
- Cox GW (1997) Making votes count: strategic coordination in the world's electoral systems. Cambridge University Press, Cambridge
- Damania R, Yalcin E (2005) Corruption and political competition. World Bank, Mimeo
- De Winter L, Dumont P (2003) Luxembourg: stable coalitions in a pivotal party system. In: Strøm K, Müller WC, Bergman T (eds) Coalition governance in western Europe. Oxford University Press, Oxford
- Drazen A (2001) The political economy of macroeconomics. Princeton University Press, Princeton
- Duverger M (1954) Political parties: their organization and activity in the modern state. Wiley, London/Methuen/New York
- Grofman B, Selb P (2009) A fully general index of political competition. *Elect Stud* 28(2):291–296
- Holbrook TM, Van Dunk E (1993) Electoral competition in the American States. *Am Polit Sci Rev* 87(4):955–962
- Lijphart A (1994) Electoral systems and party systems: a study of twenty-seven democracies, 1945–1990. Oxford University Press, Oxford
- Lijphart A (1999) Patterns of democracy. Yale University Press, New Haven
- Lipford J, Yandle B (1990) Exploring dominant state governments. *J Inst Theor Econ* 146(4):561–575
- Lipset S, Rokkan S (1967) Party systems and voter alignments: cross-national perspectives. The Free Press, Toronto
- Lizzeri A, Persico N (2005) A drawback of electoral competition. *J Eur Econ Assoc* 3:1318–1348
- Marshall A (1890) Principles of economics, vol 1, 1st edn. London, Macmillan. Retrieved 07 Dec 2012
- Mueller D, Murrel P (1986) Interest groups and the size of the government. *Public Choice* 48:125–145
- Müller WC, Strøm K (2003) Coalition government in western Europe. Oxford University Press, Oxford
- Mulligan C, Tsui K (2006) Political competitiveness. NBER working paper no. 12653, Oct
- Padovano F, Ricciuti R (2009) Political competition and economic performance: evidence from the Italian regions. *Public Choice* 138:263–277
- Persson T, Tabellini G (2000) Political economics: explaining economic policy. MIT Press, Cambridge, MA
- Persson T, Roland G, Tabellini G (1997) Separation of powers and political accountability. *Q J Econ* 112(4):1163–1202
- Pinto PM, Timmons JF (2005) The political determinants of economic performance. Political competition and the sources of growth. *Comp Polit Stud* 38:26–50
- Polo M (1998) Electoral competition and political rents. IGER, Bocconi University, Mimeo
- Rodden J, Rose-Ackerman S (1997) Does federalism preserve markets? *Virginia Law Rev* 83(7):1521–1572
- Rogers D, Rogers J (2000) Political competition and state government size: do tighter elections produce looser budgets? *Public Choice* 105:1–21
- Sartori G (1976) Parties and party systems: a framework for analysis. Cambridge University Press, Cambridge
- Skilling D, Zeckhauser K (2002) Political competition and debt trajectories in Japan and the OECD. *Jpn World Econ* 14:121–135
- Stigler G (1972) Economic competition and political competition. *Public Choice* 13:91–106

- Taagepera R, Shugart MS (1989) *Seats and votes: the effects and determinants of electoral systems*. Yale University Press, New Haven
- Van Roozendaal P (1997) Government survival in western multi-party democracies: the effect of credible threats via dominance. *Eur J Polit Res* 32(1):71–92
- Vanhanen T (2000) A new dataset for measuring democracy, 1810–1998. *J Peace Res* 37(2):251–265
- Vanhanen T (2003) *Democratization: a comparative analysis of 170 countries*. Routledge, London

Political Corruption

Roy Cerqueti and Raffaella Coppier
Department of Economics and Law, University of
Macerata, Macerata, Italy

Abstract

Political corruption represents a specific type of public-to-public corruption which implies that one participant of corrupt transaction belongs on the State and the other to the private sector: in fact, public corruption is a particular (and illegal) State-society relationship. Political corruption occurs when politicians, who are delegated to make laws and enforce them by the citizens, act themselves in a corrupt way. More precisely, it appears when policymakers exploit their political strength to pursue their own economic benefits and/or maintain their powerful position.

Introduction

To begin, it is important to define the word “corruption.” According to Jain (2001), “Corruption is an act in which the power of public office is used for personal gain in a manner that contravenes the rules of the game.” Here we are only referring to public-to-public interaction and not private-to-private. While public-to-public corruption usually involves a private party on one side and public official or member of State on the other, private-to-private corruption is any form of illegal transaction between associations or corporations

within the private sector. When, for example, a director or employee gets involved in practices that are not part of his responsibilities and that are detrimental to the company he is working for but advantageous for himself or others, this can be defined as private-to-private corruption: “the type of corruption that occurs when a manager or employee exercises a certain power or influence over the performance of a function, task, or responsibility within a private organization or corporation” (Argandoña 2003).

For many years different political scientists have approached the subject of political corruption and the difficulty in defining the term. In some cases, certain practices may be considered legal (e.g., certain forms of bribery), consequently making corruption appear less prevalent or occasionally also having the adverse effect (Olken and Pande 2012). After Peters and Welch (1978) explained that “the systematic study of corruption is hampered by the lack of an adequate definition” and “What may be ‘corrupt’ to one citizen, scholars, or public official is ‘just politics’ to another, or ‘indiscretion’ to a third,” many have tried to come up with different ways of solving the problem.

As we can deduct from Jain’s definition, corruption usually stems from three different circumstances – weak institutions, discretionary power, and economic rents.

Firstly, regarding weak institutions, the rewards gained from corruption will always have to be more significant than the foreseen payoff from the process of being discovered and punished. Secondly, for corruption to occur, the State member in question must be able to control the designation of resources in a discretionary manner. Lastly, it must be possible to extract rents or create extractable rents on behalf of the discretionary power.

Therefore with regard to public-to-public corruption, as mentioned above, there are two sides – the State (politicians, bureaucrats, civil servants, officials, and anyone who has discretionary power over society) and the private party. When politicians (or any State members) take advantage of their position to unlawfully obtain resources to their advantage (or in favor of others), this is

considered to be a corrupt practice which is carried out on the boundary between public and private environments.

Business and financial corruption which are widespread and indeed any transactions which happen exclusively in the private environment, away from the public sector, will not be considered in this voice.

It is worth identifying what actually constitutes corruption. Some practices are illegal but are not actually defined as corruption because they are not linked to political power. These include money laundering, fraud, drug smuggling, and black market transactions.

Another important point is the close link between politics and organized crime. Furthermore, there is even a new type of “business politician” or “entrepreneur” who “combines mediation in (licit or illicit) business transactions, first-hand participation in economic activity, and political mediation in the traditional sense” as stated by della Porta and Pizzorno (1996).

There is a specific analysis of corruption, known as the “principal-agent analysis” (Rose-Ackerman 1978), which involves considering three main agents or “actors”: the principal (the State), the private agent (citizen or company), and an agent (a public official) who in theory follows the interests of the principal. If the private party wants to obtain an advantage, he pays a bribe to the public official to help him instead of carrying out the interests of the principal. This is considered to be corruption.

Corruption can happen on two levels – bureaucratic and political. Bureaucratic or petty corruption is corruption in the public administration where bureaucrats (known as the “actors”) use their public position for financial or nonfinancial benefits. Political or grand corruption, on the other hand, is when politicians are corrupt and use their power for illegitimate financial gain and to maintain their political position. They are given the right to make decisions and to create laws, but they abuse this right in order to satisfy their own requirements.

Since political corruption has many different aspects, it is a difficult task to come up with a conclusive definition of the term. The

characteristics and objectives of political corruption depend on the polity under discussion. Nye (1967) gives this statement about political corruption as “behavior which deviates from the formal duties of a public role (elective or appointive) because of private-regarding (personal, close family, private clique) wealth or status gains: or violates rules against the exercise of certain types of private-regarding influence.” The various forms of political corruption can be recognized and analyzed by using an overall definition as a starting point.

It is worth noting that political and bureaucratic corruption are in fact inextricably linked as in the two different spheres their areas of competence can become interchangeable. However, for our objectives, the two different forms of corruption should be analyzed in practical terms.

In this encyclopedic voice, the focus is on political corruption, as opposed to general corruption; therefore one can concentrate on the political actors and the public environment in which they perform. Political corruption has a complex nature with a wide range of aspects which should be outlined. Although an extensive classification of the subject could reduce the analytical study, it is necessary in order to fully understand the different forms and characteristics.

Different Types of Political Corruption

Political corruption is not only a question of breach of formal legislations, code of conducts, and court rulings, but it also has a deeper impact on the entire political system. It can have a negative influence on decision-making, leading to the mishandling of procedures and finally the breakdown of political institutions.

While political corruption involves politicians carrying out illegal practices in order to maintain their position and improve their personal situation, it does not include police misconduct or suppression of political rivals. It is also important to clarify if the political corruption is taking place in democratic nations or dictatorial countries, due to the fact that there is lack of accountability between those in power and “the people.” There

are problems with political corruption in dictatorial countries as the legal framework is already fragile, making it difficult to assess corruption and often leading to infringement by the leaders. Therefore the legal infrastructure of the State cannot be used as a point of reference to determine the extent of political corruption.

While bureaucratic corruption can often be approached by using administrative means (auditing, legislation, and institutional organization), political corruption with its detrimental consequences needs a much stronger solution. It is necessary to use legislative, moral, ethical, and political reference points first and foremost to differentiate between legality and legitimacy regarding political corruption.

Major political reforms are essential for fighting local political corruption. In the majority of dictatorial countries, political corruption is a routine event. As bureaucratic corruption is often rife in these regimes, this encourages political corruption, which obviously varies according to the different types of authoritarianism. Rulers of these countries use corrupt means in order to empower themselves and maintain their political status.

With regard to democracies, political corruption is more of a sporadic and incidental event and can therefore often be resolved by means of reforms and reinforcing the political framework and institutions of checks and balances.

Political corruption can be divided into different categories, mainly “personal” and “institutional” or “collective” corruption. The difference depends on what extent the financial (or other) rewards from the corruption are “privatized.” While some may share the extraction with his associates, others will keep most if not all of the rewards for himself. “Personal” corruption is when politicians abuse their status to illegally obtain benefits for “personal” gain. A very common example is the acceptance of bribes by politicians and giving out benefits in return. On the other hand, “collective” corruption is a form of corruption which involves extracting resources for the benefit of a bigger group or organization, not just the private individual. Public power is abused for “private” gain, which does not necessarily mean for the personal benefits of one person

but also for “collective” gain, for private groups. It is often the ruling parties and/or potential ruling parties, national governments, and administrative authorities that become involved in this form of corruption, using the resources at hand to their own advantage.

There is a significant distinction to be made between political corruption and “lobbying” when talking about “collective corruption.” According to some (Damania et al. 2004; Harstad and Svensson 2011; Campos ad Giovannoni 2007), the definition of “corruption” only refers to petty or bureaucratic corruption, whereas any form of practice done to influence policy makers should be referred to as “lobbying.” However, this definition cannot be applied to all political systems as it does not distinguish between what is legal and what is illegal when influencing policy makers and does not take into account the regulations in all countries. While in the United States, where it is completely legal to provide pecuniary payments to policy makers (lobbying) and the definitions of petty corruption and lobbying may be applied, in other nations, the same payments are regarded as illegal and therefore politically corrupt.

It is worth looking at the procedures used to obtain private gains within the political environment in order to distinguish between the different forms of political influence (including lobbying) and political corruption. Since political corruption occurs when politicians obtain private benefits by means of nontransparent, unofficial procedures, one could say that lobbying is a particular form of corruption as it is still a way of gaining benefits from public entities or legislative organizations, with favors being given in return. However, lobbying and political corruption do have their differences which need to be noted. Lobbying is not always a case of bribes and campaign contributions. According to Austen-Smith and Wright (1994), lobbyists are capable of influencing politicians as they share competences and know-how which some politicians lack. Grossman and Helpman (1999, 2001) mention how some lobbyists even threaten politicians by saying they will give voters detrimental information about their policies, while others sway politicians with endorsements.

Other two types of corruption are “transactive” and “extortive” corruption. “Transactive” corruption is an illegal exchange between a donor and a recipient, where both actors benefit. “Extortive” corruption usually involves coercion of some sort so that the donor and people he is close to are not hurt in any way. Transactive practices are in theory based on reciprocity. It is a transaction between public officials (the State) and private actors (society) where both parties benefit from the exchange. In this corrupt exchange, there are two further distinctions to be made between “redistributive corruption” and “extractive corruption.” With “redistributive corruption,” it is the State who provides the private citizen or businessman with the resources, while “extractive corruption” is from the private actor to the State. Even though in theory the relationship is a reciprocal one, in reality this is rarely the case. In some nations or States in the world, the State plays the weaker role in transactions with society. Here redistributive corruption might take place as the mafia is rife and prevalent clientelism, which means that the State loses power as these private individuals or groups form a corrupt relationship with the State but end up with more benefits than the State itself.

“Extractive corruption,” on the other hand, is when the State has the upper hand in the relationship. This is often seen in the political framework of various authoritarian States. The corrupting group or individual plays a passive role as the State (the corrupted) has the advantage in the corrupt exchange.

The Principal Causes of Political Corruption

It is a complex task to try and find the reasons behind political corruption. Rose-Ackerman (1996) states “Many officials remain honest in the face of considerable temptation, and others accept payoffs that seem small relative to the benefits under their control. Others, however, amass fortunes. The level of malfeasance depends not only on the volume of potential benefits, but also on the riskiness of corrupt deals and on the

participants’ moral scruples and bargaining power.”

Various theories explain the widespread phenomenon of political corruption through cultural and moral factors typical of a country. When considering the factors that contribute to the onset and spread of corruption, an important role should be assigned to the cultural, ethnic, and social framework that fix in the “moral” of a country the seriousness of an act of corruption. Cultural norms, which vary from State to State, pose a boundary between a gift and a bribe or favoritism: therefore the definition of what is corrupt or not is a cultural thing (Rose-Ackerman 1999) and what can be defined by a corrupt external observer can be considered as an acceptable gift within a country.

However, the economic analysis of the causes and the extent of political corruption cannot and must not be limited to cultural factors. In fact, in addition to these, one needs to consider structural and institutional (economic and political) aspects that may override cultural interpretations (the latter is sometimes used as a way of justifying corrupt practices in some countries).

It is worth noting that the framework or “structure” of the administration and how well the institutions of a country work can influence the spread of corruption. For example, in Britain where the administrative organization is separate from the political system and the political parties were institutionalized at an early stage, there is notably less local political corruption. If there is a clear boundary between the two different systems (social and political procedures from political and economic ones), it is more difficult for politicians to enter into the bureaucratic environment.

In some countries, the presence of so-called regional brokers or different forms of mafia has meant that the social order of these States has been weakened. Due to the particularistic and fragile financial and administrative network in these countries, the State has had to depend on these “brokers” to function locally.

It is worth mentioning that each State develops in different ways, so it is perhaps incorrect to conclude that the infiltration of these groups

between the bureaucratic authorities and political system is the main cause of political corruption. That said, widespread clientelism may indeed be largely to blame for the corruption problem. Della Porta (1997) mentions how the links between clientelism and corruption and weak administration structures and corrupt practices have become “vicious circles” from which it is incredibly difficult to break free.

Another cause of political corruption identified in literature is government intervention in the economy. This intervention creates income managed on a discretionary basis by public officials, which in turn can generate corruption. Due to the growth of the welfare state and public sector, the political elite now have the monopoly over many financial resources.

It is easier for members of State and private individuals to sidestep rules and regulations as decisions are not influenced by the workings of the market but linked to these rules.

While the focus so far has mainly been on the reasons and opportunities for one to enter into corrupt practices, it is also important to look at “institutional” and “political” factors.

One significant factor is in what political context the corruption takes place. There is a widespread belief that corruption is measured by the level of democracy in a country. The more democratic a State, the less corrupt it should be. According to Friedrich (1989) the more lawful the country, the less likely corruption will evolve.

However, in some authoritarian countries, dictators are able to control the extent of corruption. In other words he will be able to decide who benefits from the State resources. For this reason, in some countries under powerful dictatorships, corruption is actually less rife. The citizens will consider these States a legitimate entity as they have the power and ability to carry out social changes and promote economic production as well as to manage law and order.

Regarding the functioning of political parties, we must consider the different methods of financing of political parties and the funding of elections. In countries such as the United States and Japan, this is now becoming a key matter with regard to political corruption.

Due to the high costs of elections and the political process in general, parties often have to look elsewhere for financial resources. They have few funds available and therefore attempt to find other means to keep their political status and to appear the most efficient in the eyes of the public.

Another important factor which may lead to political corruption is “longevity in power.”

This term refers to how governments can stay in power without being challenged by other rivals. In authoritarian States, the absence of a formal autonomous structure means that bureaucrats or public officials can take advantage and use power in corrupt ways. On the contrary, in democratic countries public accountability is guaranteed by the institutions of the “State of law.” If governments believe that their position of power is infallible, many begin to confuse the difference between State and government and as a result begin to think they are exempt from the laws that apply to everyone else.

Among the causes of greater or lesser spread of corruption are certainly the control system and the extent of the punishment related to corrupt acts. In deciding whether to engage in a corrupt transaction, an individual (be they public officials or firms), by comparing the expected benefits with the expected cost, decides to be corrupt only if the total benefit is at least equal to the total cost. The reforms designed to increase the risk of discovery and the extent of the punishment may reduce the demand for and supply of bribes. With regard to the likelihood of being discovered, it is often endogenous compared to the level of corruption, in that “the more widespread the corruption, [...] the lower the risk of being denounced” (della Porta and Vannucci 2016).

Consequences of Political Corruption

While most analysts concur that political corruption is morally condemnable, its economic and political consequences are less discussed. This is perhaps due to the fact that political corruption – complex as a concept – is difficult to both define and measure.

Corruption, in the past, has even been considered as having a positive role in some developing

countries. This was the opinion of analysts who were working within the scheme of the “modernization theory” in the 1960s. More precisely, the debate on the economic effects of corruption started with the two pioneering works of Leff (1964) and Huntington (1968), who said that corruption would stimulate economic growth mainly through the operation of two mechanisms:

- Corruption if “speedy money” can ensure that individuals are able to circumvent red tape.
- If the bureaucrats are paid directly for their work through bribery, this should make the bureaucrats work better and faster.

Therefore corruption is seen as something constructive in promoting economic growth and investment as well as helping social development in countries with weak bureaucracy, making up for the lack of adequate formal practices. However, these statements have been widely opposed by recent literature that says that corruption can have positive effects only in limited areas: they refer to the benefits arising from the assumption that State intervention creates inefficiency and corruption, shrinking them or eliminating them for the best economic growth of the system. Therefore, only in cases where the political system is inefficient, may political corruption be an improvement (see della Porta and Vannucci 1997).

With the spread of corruption, government costs will increase, public services will no longer be guaranteed as resources will begin to dwindle, revenues may be exposed, and it will no longer be possible to make effective decisions. Corruption can also have a negative influence on the purchase or even production of goods for the public. The whole process of purchasing, tendering, giving contracts, the realization of work, and payment at the end may all be affected.

A key moment in the political process of a democracy is when one needs to respond to public demand. The administration network is faced with the demands and requirements of different social groups. The function of political mediation according to Pizzorno (1992) is “to identify and interpret the needs and desires of the population; select and generalize those which can be

expressed in political terms; propose, justify and criticize policies and measures to achieve these ends or, when necessary, to explain why they cannot be satisfied.” It is the means through which interests are collected and expressed.

The efficiency of this service is then jeopardized by corruption as the demand becomes “internal” not “public.” Public demand is no longer protected as the formal restrictions of the institutions are weakened by corrupters, who are able to make decisions and take the profits from corrupt practices creating the disintegration of public demand to satisfy specific needs.

Public spending is often inextricably linked with corruption and clientelism. Public officials aim to draw as many resources as possible into their domains of power so they can receive payment for mediation as a bribe. In this way, these administrators also try to win public support as a consequence of the influence of public investment on employment. For this reason, public spending is moved into the areas where there are the most benefits from corruption and where there are fewer risks due to the discretionary characteristics of the practices. Citizen needs are not taken into account when these operations and services are carried out and many works remain incomplete or discarded altogether. In addition, in the case where politicians are not content with the sum given through bribes, public demand often remains unfulfilled. It is worth noting, however, that during purchasing transactions, there are competitive rules which mean that those participating have equal treatment once the cost of public spending has been decided.

In fact, in a widespread corrupt society, it can therefore be expected that those who formulate public demand will award the provision of public goods or a “carte” of all the companies participating in tendering. The bids are then able to be correlated and the income obtained then redistributed. Politicians or bureaucrats who have been involved in corruption in the past will then make sure that these individual companies or group of companies will definitely (or most likely) win a tender.

As a consequence, public demand will then be administered in such a way that these “protected”

companies will have an advantage in promoting their own products and will be able to benefit in the future from their “special” connections with politicians. Costs become inflated as public tenders are allocated to companies who lack the required competences, equipment, and qualifications.

Another element which politicians use to manipulate public demand is when they have real (or false) “urgent” interests. Usually the amount of public spending and its administration will be relatively high so corrupt politicians need to create a contrived emergency.

Lack of planning expertise of public administration and convoluted bureaucracy slows down the time taken to complete public operations, which in turn paves the way for corruption as discretionary proceedings in the contracting process become legitimate.

Corruption therefore creates a redistributive process from which politicians benefit: “A corrupt system of government services has the distributional disadvantage of benefiting unscrupulous people at the expense of law-abiding citizens who would be willing to purchase the services legally” (Rose-Ackerman 1978). Corruption attributes property rights to agents who then breach the rights of public interest and consequently further the tax burden which has a detrimental impact on the life of citizens.

Cross-References

- [Administrative Corruption](#)
- [Corruption](#)

References

- Argandoña A (2003) Private-to-private corruption. *J Bus Ethics* 47:253–267
- Austen-Smith D, Wright JR (1994) Counteractive lobbying. *Am J Polit Sci* 38:25–44
- Campos N, Giovannoni F (2007) Lobbying, corruption and political influence. *Public Choice* 131:1–21
- Damania R, Fredricksson PG, Mani M (2004) The persistence of corruption and regulatory compliance failures: theory and evidence. *Public Choice* 121:363–390
- della Porta D (1997) The vicious circles of corruption in Italy. In: della Porta D, Mény Y (eds) *Democracy and corruption*. Printer, London, pp 35–49

- della Porta D, Pizzorno A (1996) The business politicians: reflections from a study of political corruption. *J Law Soc* 23:73–94
- della Porta D, Vannucci A (1997) The “perverse effects” of political corruption. *Political Stud* XLV:516–538
- della Porta D, Vannucci A (2016) *The hidden order of corruption: An institutional approach*. Routledge, London.
- Friedrich CJ (1989) Corruption concepts in historical perspective. In: Heidenheimer AJ, Johnston M, LeVine VT (eds) *Political corruption. A handbook*. Transaction Publishers, New Brunswick. (third printing 1993)
- Grossman G, Helpman E (1999) Competing for endorsements. *Am Econ Rev* 89:501–524
- Grossman G, Helpman E (2001) *Special interest politics*. MIT Press, Cambridge, MA
- Harstad B, Svensson J (2011) Bribes, lobbying and development. *Am Polit Sci Rev* 105:46–63
- Huntington SP (1968) *Political order in changing societies*. Yale University Press, New Haven
- Jain A (2001) *The political economy of corruption*. Routledge, London
- Leff NH (1964) Economic development through bureaucratic corruption. *Am Behav Sci* 8:8–14
- Nye JS (1967) Corruption and political development: a cost-benefit analysis. *Am Polit Sci Rev* 61:417–427
- Olken BA, Pande R (2012) Corruption in developing countries. *Annu Rev Econ* 4(1):479–509
- Peters JG, Welch S (1978) Political corruption in America: a search for definitions and a theory, or if political corruption is in the mainstream of American politics why is it not in the mainstream of American politics research? *Am Polit Sci Rev* 72:974–984
- Pizzorno A (1992) *La corruzione nel sistema politico*. Introduction to Della Porta D, *Lo scambio occulto*. Il Mulino, Bologna
- Rose-Ackerman S (1978) *Corruption. A study in political economy*. Academic, New York
- Rose-Ackerman S (1996) Democracy and ‘grand’ corruption. *Int Soc Sci J* 48(149):365–380
- Rose-Ackerman S (1999) *Corruption and government*. Cambridge University Press, New York

Political Economy

Karsten Mause
Department of Political Science (IfPol),
University of Münster, Münster, Germany

Definition

The term “political economy” (PE) is mainly used in two related contexts. First, it is used to denote a multidisciplinary research field in which political

scientists, economists, legal scholars, and other social scientists investigate the relationship between the political sphere (most notably “the state”) and the economic system of different societies on Earth at different points in time. Second, social scientists, journalists, and other observers sometimes use the term PE to refer to the observable interaction of politics and business in real-world societies. Focusing on the first context, this entry gives an overview of the research field of political economy (PE) and discusses its relationship to law and economics as a research program.

Positive Political Economy: Analyzing What Is

Within the toolkit of PE, there are basically two different approaches which are currently used to analyze the relationship between politics and the economy: positive PE (explained in this section) and normative PE (see next section). Using a positive approach, researchers conduct empirical analyses in the form that they describe and explain the relationship between the political and economic sphere (i.e., what is) – but they usually do not make value judgments in the form of normative statements as to what public and private sector actors *should* do or not do: for instance, positive politico-economic analyses usually do not contain policy recommendations regarding the “right” way for the government to intervene in certain sectors or markets in the economy (i.e., what ought to be). In other words, scholars doing positive PE research first of all *describe* as precisely as possible the extent to which the state or other political actors intervene in the economic system of a society (or different societies, if a comparative perspective is taken) in a certain investigation period.

Political interventions in different sectors of the economy or particular markets may take various forms, such as government subsidies, taxation, state-owned enterprises, or regulations (Den Hertog 2000; Boettke and Leeson 2015), and may be done for various reasons, such as eliminating market failures/allocative inefficiencies, redistributing resources from the rich to the

poor, or stimulating economic growth and employment (see below). As it is often the case that firms, business associations, trade unions, and other actors within the economic system try to influence the process of economic policymaking via lobbying and other forms of leverage, in many contexts we can observe a mutual interference of the political and the economic sphere. This may also include a possible correlation between (a) the economic situation in a society and (b) the popularity and election results of government, opposition parties, or individual politicians (Lewis-Beck and Stegmaier 2013).

Moreover, it has to be taken into account that external factors such as global issues (poverty, climate change, war refugees, etc.), international organizations (e.g., World Trade Organization, European Union, International Monetary Fund, World Bank), developments in international markets, or the activities of foreign governments (e.g., tariff policy, international tax competition, sovereign debt, sovereign defaults) may influence a (sub)national PE understood as the interaction of the political and economic system in a real-world society. The fact that nation states are these days embedded, in various respects, in an international system is analyzed in the literature on “international PE” and “global PE” (see, e.g., Ravenhill 2016). In this context it should also be mentioned that there is a strand of PE research which focuses explicitly on the differences between the national economic systems of the countries in the world including different “Varieties of Capitalism” (Hall and Soskice 2001) as well as the remaining more or less socialist “command economies” or “centrally planned economies” such as Cuba and North Korea (Fine and Saad-Filho 2012, chap. 7).

However, many politico-economic analyses within positive PE do not content themselves with describing the relationship between politics and the economy but, moreover, try to *explain* the observations made in the descriptive phase of research. Which factors can explain why the state intervenes in a particular way in a society’s economy? Which explanatory factors may have driven the transformation of the “interventionist state” over time? Why do some countries show a better macroeconomic performance (economic growth,

employment, price stability, etc.) than other countries? In this spirit, for example, numerous politico-economic studies have empirically analyzed whether factors such as government ideology, powerful interest groups, fiscal pressure, socioeconomic problems (e.g., de-industrialization, unemployment, economic slump), path dependence, or globalization help explain the observable differences across the member states of the European Union (EU) or the Organization for Economic Co-operation and Development (OECD) with respect to the use of policy instruments such as public entrepreneurship, regulation, taxation, or subsidization in the decades after World War II (e.g., Leibfried et al. 2015; Obinger et al. 2016).

Similar studies exist for the less-developed world and/or for country groups including countries with “not-so-democratic” political systems. There is, for example, a politico-economic literature that describes and explains different aspects within the relationship between politics and the economy in autocratic regimes (e.g., Acemoglu and Robinson 2012; Leibfried et al. 2015, parts IV/V). Moreover, there are many studies entitled “The Political Economy of XY” which means that the particular study analyzes the interplay between political and economic factors in the specific context under investigation, for example, the political economy of migration, foreign aid, higher education, terrorism, and so on.

Normative Political Economy: Recommending What Ought to Be

PE research may be done not only in the form of empirical or “positive” analyses (as defined above) but also in the form of a *normative* analysis. This means that a specific area in the economic system is analyzed in order to come to conclusions as to what “the state” should (not) do in the area under investigation. Should the government intervene in a particular sector of the economy or a particular market by means of regulations or other policy tools? Should public bureaucrats be allowed to control certain activities of private sector firms and households? Should

regulatory agencies be mandated to supervise competition in particular sectors and markets? Such questions are addressed in the fundamental and ongoing PE debate over the proper role of the state in the economy (see Boettke and Leeson 2015, for a survey). Contributions to this debate are based, more or less explicitly, on the following major schools of thought.

Varieties of Economic Liberalism

Political economists in the tradition of Adam Smith (1723–1790), whose seminal book on “the Wealth of Nations” (Smith 1776/1981) is the “bible” for those advocating “economic liberalism” and “market liberalism,” basically argue that the state should leave the economy alone. It is assumed that there is some kind of natural tendency to equilibria in markets. That is, if there is an excess demand or excess supply, then such disequilibrium will only persist for a short period of time. According to the economic laws of demand and supply, markets will find a “market-clearing price” at which demand equals supply. In other words, direct governmental interventions into markets are perceived to be unnecessary (or even harmful) as specific markets and the economy as a whole possess “self-healing powers” in the form of the “market forces”: that is, the interplay of demand and supply coordinated via the price mechanism.

However, it should be mentioned that Smith (1776/1981) and other advocates of economic/market liberalism such as Friedrich August von Hayek (1899–1992) and Milton Friedman (1912–2006) acknowledge that society may not be left to markets alone – but that the state has to perform at least some tasks to make markets and society work. For example, political economists in the tradition of Smith (1776/1981), Hayek (1960), and Friedman (1962) consider it to be a government task to ensure that there is a functioning legal system (rule of law, laws, courts, judges, etc.) that can be used, among other things, for enforcing (i) property rights and (ii) the contracts signed by market participants. By contrast, *libertarian* political economists, who consider the possibility of a stateless society, go a step further: they argue that private governance mechanisms

(reputation, nongovernmental courts, etc.) are sufficient to enforce property rights and contractual agreements (see, e.g., Friedman 2014; Leeson 2014; Stringham 2015).

Furthermore, there are two specific variants of economic liberalism which were developed some decades ago but are still influential in the current politico-economic discourse: ordoliberalism and constitutional political economy. Ordoliberals in the tradition of the German economist Walter Eucken (1891–1950) criticize that Smith (1776/1981) and other advocates of classical economic liberalism and its *laissez-faire* approach have neglected that a market economy does not automatically increase the wealth of a nation. For example, individual markets or whole sectors of the economy may suffer from anticompetitive practices by private and/or public companies (market-entry barriers, cartelization, price collusion, competition-distorting state aid, government monopoly, and so on). Consequently, for ordoliberals it is essential that the state creates and enforces a legal order and institutions (e.g., a politically independent competition authority) that try to prevent private and governmental restraints of competition and market forces as far as possible (Eucken 1952/2004; Vanberg 2015).

In a similar vein, constitutional political economists in the tradition of James M. Buchanan (1919–2013) emphasize that markets, competition, and the economy as a whole need “rules of the game” – a “constitution” – which channel the individual self-interests of consumers, firms, and other actors (for more details, see Buchanan 1987; Vanberg 2005). Moreover, constitutional PE points out that creating and enforcing such “rules of the game” is far from trivial. Reading Eucken (1952/2004) one gets the impression that he takes it for granted that there is a benevolent government which realizes that the rules recommended by ordoliberals are beneficial for society and implements these rules. In contrast, constitutional PE assumes that not only firms and consumers but also politicians and public bureaucrats are self-interested actors. Under these conditions, not only powerful firms and interest groups but also politicians and public bureaucrats may impede the implementation of rules which could be beneficial

for citizen-consumers and society as a whole (e.g., the abolition of government monopolies, the abolition of special privileges for state-owned enterprises, better regulations for public utilities, and so on). However, what ordoliberalism and constitutional PE have in common is that both prefer a rule-based economic policy over discretionary government interventions in the economy and market processes.

For the sake of completeness, it should be mentioned that constitutional PE is part of the broader research program entitled “economic theories of politics” or “public choice theory” established by Downs (1957) and others (see Mueller 2003, for a survey). Public choice theory breaks with the welfare-economic assumption of benevolent governments working in the public interest. Instead, it is assumed that politicians and public bureaucrats (i) are primarily interested in maximizing their individual utility and (ii) “act solely in order to attain the income, prestige, and power which come from being in office” (Downs 1957, p. 28). In other words, it is theoretically assumed that political decision-makers are self-interested not only when they make private choices (as consumers, investors, landlords, and so on) but also when they make public choices in government, parliamentary committees, and other political contexts. As “older” schools of PE (e.g., classical economic liberalism in the tradition of Smith) did not pay much attention to the motivations of public sector actors and implicitly assumed that the government is primarily interested in maximizing the wealth of a nation, in the politico-economic literature, public choice theory is often denoted as the “New Political Economy” (Frey 1999).

The next step in this area has been taken by scholars working in the field of “behavioral political economy.” Therein, it is taken into account that real-world actors often do not behave as rationally and with the self-interest that economists’ traditional *homo-economicus* model predicts; this may lead to other policy implications regarding the “optimal” design of the incentive structures under which certain types of consumers, investors, policymakers, and other individuals make their more or less informed and more or less

selfish decisions (see Schnellenbach and Schubert 2015, for a survey).

Keynesianism and Other State-Interventionist Approaches

If the economy slips into recession, then hard-core economic liberals may argue that such an economic crisis may have painful consequences for firms and individuals (a drop in orders, bankruptcy, unemployment, poverty, and so on) but does not require government intervention – because thanks to its “self-healing powers” the economy will recover on its own after some time. By contrast, political economists in the tradition of John Maynard Keynes (1883–1946), whose seminal book “The General Theory of Employment, Interest and Money” (Keynes 1936) belongs to the most influential critiques of the laissez-faire approach of economic liberalism, consider it to be a government responsibility to stimulate the economy in times of economic slump (i.e., increasing government spending, reducing taxes, and so on). If the government lacks the necessary financial resources to implement an economic stimulus package, then Keynesians recommend (a) government borrowing (so-called deficit spending) and (b) repaying the debts after the crisis when government tax revenues increase due to economic growth and rising employment.

Critics of deficit-spending object that step (b) is often not conducted by government, which is one reason for the high levels of public debt observable in many countries these days. Keynesians usually respond to such criticism by arguing that costly state interventions to “stimulate,” “stabilize,” and “steer” the economy are necessary and legitimate as long as there is unemployment in a society (Krugman 2013). Complementary to fiscal stimulus packages, Keynesians propose measures of monetary policy to stimulate the economy (e.g., lower interest rates). If there is a politically independent central bank, then monetary policy is not a tool of government (i.e., politicians have no access to the tools of monetary policy).

While Keynesianism offers a macroeconomic justification for state intervention in the economy, the so-called market failure theory (for a survey,

see Stiglitz and Rosengard 2015) has demonstrated that the laissez-faire approach of economic liberalism ignores the fact that different types of market failures offer a potential justification for government action. For example, the behavior of certain firms and consumers (e.g., environmental pollution by coal-fired power plants) may create *negative externalities* for other society members. The government may implement measures (law, regulations, etc.) that force polluters to reduce or even stop producing negative externalities. Moreover, it can be expected that many society members will not pay for certain goods and services if they can consume these goods and services free of charge. However, if free riding is possible, then private actors have a low or no incentive to supply such goods and services. To secure the provision of *public goods* in the sense that no one in society can be excluded from consuming such goods, the government may step in: for example, the public good argument offers an economic argument to justify the national defense being provided by the government and financed by taxes (i.e., society members, as potential free riders, are forced to pay for national defense).

Informational asymmetries constitute another type of potential market failure. If suppliers are better informed about certain characteristics of products and services (e.g., the quality of used cars) than potential buyers, then the markets for these products and services may not function well: because it can be expected that many consumers under these circumstances would hesitate to enter a market transaction as they would fear being exploited by the better-informed sellers (e.g., low-quality, high-price products). It is also possible that buyers are the better-informed market party. Imagine, for example, insurance companies that do not know the true health status of people seeking to buy a health insurance. In situations with informational asymmetries, the government may implement measures (governmental provision of quality information, disclosure laws, governmental regulation of product quality, etc.) to mitigate these informational problems and facilitate market transactions. Moreover, the market-failure framework considers *natural monopolies* to be a potential justification for government

action. Such monopolies occur if for efficiency reasons in certain sectors or markets of the economy only one firm is doing business (e.g., the provider of a rail network, a power supply line or a water line). To avoid allowing this provider to exploit its *monopoly power* (high prices, bad quality, and so on), the government may regulate this natural monopoly (price regulation, quality regulation, etc.). And, as mentioned above in the context of ordoliberalism, the government may also intervene in some way to tackle the problem that markets and competition do not work properly due to “ordinary” monopolies and other anticompetitive practices.

It should be emphasized that the existence of a market failure does not automatically imply that the government has to solve the problem. For instance, there may be private third parties (e.g., private certification agencies) and market-based mechanisms (e.g., reputation, brand-name capital) that help market participants to overcome their informational problems, so that buyers and sellers are able to enter into mutually beneficial market transactions. In other words, in the politico-economic literature, it is not only discussed (a) whether a certain market or sector of the economy really suffers from “market failures” and “allocative inefficiencies” but also (b) which governmental or private governance mechanisms (or a mixture of both) seem to be the most suitable to solve the problem at hand (Ostrom 2010). Moreover, political economists stress that all of these mechanisms are imperfect solutions that work more or less well depending on the specific real-world context in which they are used (Wolf 1993). And it may be the case that government action to solve a market-failure problem may create new problems (for an overview of the politico-economic debate on “government failure,” see Keech and Munger 2015).

A normative yardstick that is often used by economic liberals to assess whether state activity is necessary to mitigate a certain type of market failure is the so-called subsidiarity principle. According to this principle, government action is only necessary if private market solutions and private governance mechanisms fail. A brief and oft-cited summary of this principle can be found

in the book “Principles of Economic Policy” by Eucken (1952/2004, p. 348): “The structure of society should follow a bottom-up approach. What the individuals or the groups can autonomously accomplish should be done on their own initiative and to the best of their abilities. And the state should only intervene in those cases in which its assistance is indispensable” (own translation, K.M.). By contrast, political economists that have a less individualistic and more state-centered view of economy and society may start from the paternalistic, state-interventionist assumption that the state is automatically responsible for solving market-failure problems (for a survey of the politico-economic literature on paternalistic government, see Le Grand and New 2015). In democratic societies, the ultimate decision-maker in this context is the government in power – and this decision-maker is certainly free to ignore the normative (and often conflicting) policy recommendations made by political economists and other experts.

Marxism and the Social Question

Economic liberalism and its belief in markets and competition have always been the target of criticism. Karl Marx (1818–1883) and Friedrich Engels (1820–1895) have argued that it is a basic feature of capitalist market economies that the “working class” (the so-called proletariat) is exploited by business firms and their owners (the “capitalists”; see, e.g., Marx and Engels 1848/2002). The state is seen as an agent of the so-called bourgeoisie (including the capitalists) which constitutes the ruling class in society. Marx and Engels predicted that capitalism will be overthrown through a “proletarian revolution” that leads to socialism and, eventually, to communism (including a classless society). It is beyond the scope of this paper to critically review everything that has been written by Marx, Engels, and their followers under the label “Marxian Political Economy” about imagined and real existing types of capitalism, socialism, and communism (for a survey, see Fine and Saad-Filho 2012). Nor do we discuss the many problems of “command economies” or “centrally planned economies.” However, Marx and his followers have repeatedly

pointed out a serious problem which many capitalist market economies still have to cope with: it may be the case that an economy consists of a system of well-functioning, efficient markets, but this system produces social problems.

For example, in many countries we observe income and wealth inequality among society members. While hard-core economic liberals may argue that such inequalities have to be accepted and simply reflect individual differences in performance and success on markets, other political economists argue that the state in the name of “social justice” has to tackle distributional problems via redistribution (see Piketty 2014, for an overview of this debate). And even economic liberals that are skeptical of state interventions, such as Hayek (1960) and Friedman (1962), take it for granted that those society members who, for whatever reason (e.g., disease, disability), are not able to earn money in the labor market should receive publicly financed welfare benefit payments ensuring a minimum income needed to exist. Whatever political economists from different schools may think about social problems and their solution – in the end, however, in democratic societies the scope and structure of the welfare state are determined in the political process.

Political Economy Meets Law and Economics

Last but not the least, we have to address the question of what law and economics (LE) as a research program has to do with PE as a multidisciplinary endeavor. First of all, we can observe that terms, concepts, and tools from the toolkit of PE (market failure, externalities, public goods, efficiency, utility, welfare, constitutional PE, behavioral PE, and so on) are used by LE scholars and in LE textbooks as well (see, e.g., Towfigh and Petersen 2015). In this context, it should also be mentioned that economists belonging to the LE movement have made important contributions to the research field of PE as sketched above. See, for example, the studies on externalities and public goods by Ronald H. Coase or the contributions by George J. Stigler, Richard A. Posner, and Samuel

Peltzman to the economic theory of regulation (see the bibliographies in Den Hertog 2000; Boettke and Leeson 2015). In other words, many of the concepts presented above under the label PE, which is mainly used by political scientists, economists, and “political economists,” are presented in LE publications under the label LE, which is mainly used by legal scholars, economists, and supporters of the LE movement. And while some may classify Coase, Stigler, Posner, and Peltzman as economists or LE scholars, others may classify them as political economists.

Likewise, Persson and Tabellini (2003), La Porta et al. (2008), and similar studies investigating the interplay of legal institutions and the economy are oft-cited in the PE as well as in the LE literature. In any case, it should be clear now that political economists and LE scholars who are interested in analyzing different aspects of the interplay between the political and economic sphere of society share a common terminology. This offers opportunities for research cooperation and interdisciplinary research – but does not mean that the disciplines participating in the “joint ventures” labeled PE and LE would have lost their idiosyncrasies and specific strengths. For example, as noted above, in the politico-economic works of Hayek, Friedman, Eucken, and Buchanan it is argued that “the state” should provide a legal framework which ensures that markets and competition work well; however, these thinkers do not say much about the fundamental issue of how exactly the specific legal framework for a specific market or economic sector in a particular real-world society should be designed and enforced (contract law, competition law, capital market law, energy law, environmental law, and so on). That is, legal experts are necessary to bridge the gap between normative politico-economic theories of the proper role of the state and practical public policy.

Moreover, it can be observed that in the positive, empirical branch of PE and LE, there seems to be a methodological convergence in the sense that scholars doing empirical research in this area basically use the same toolkit, consisting of various quantitative and qualitative methods (statistical techniques, document analysis, field

research, etc.; Towfigh and Petersen 2015). As we have seen above, however, such consensus cannot be observed in the normative branch of PE. Looking through the theoretical – some would say “ideological” – lenses of different schools of PE in many cases brings us to different conclusions regarding the question of what the state should do (or not do) in the particular area of the economy under investigation.

Cross-References

- [Capitalism](#)
- [Constitutional Political Economy](#)
- [Government Failure](#)
- [Market Failure: Analysis](#)
- [Market Failure: History](#)
- [Ordoliberalism](#)
- [Public Choice: The Virginia School](#)

References

- Acemoglu D, Robinson J (2012) *Why nations fail: the origins of power, prosperity, and poverty*. Crown Business, New York
- Boettke PJ, Leeson PT (2015) *The economic role of the state*. Edward Elgar, Cheltenham
- Buchanan JM (1987) The constitution of economic policy. *Am Econ Rev* 77:243–250
- Den Hertog JA (2000) General theories of regulation. In: Bouckaert B, De Geest G (eds) *Encyclopedia of law and economics, The regulation of contracts*, vol III. Edward Elgar, Cheltenham, pp 223–270
- Downs A (1957) *An economic theory of democracy*. Harper & Row, New York
- Eucken W (1952/2004) *Grundsätze der Wirtschaftspolitik*, 7th edn. Mohr Siebeck, Tübingen
- Fine B, Saad-Filho A (eds) (2012) *The Elgar companion to Marxist economics*. Edward Elgar, Cheltenham
- Frey BS (1999) *Economics as a science of human behaviour: towards a new social science paradigm*, 2nd edn. Springer, Dordrecht
- Friedman M (1962) *Capitalism and freedom*. University of Chicago Press, Chicago
- Friedman DD (2014) *The machinery of freedom: guide to a radical capitalism*, 3rd edn. Open Court, Chicago
- Hall PA, Soskice D (eds) (2001) *Varieties of capitalism: the institutional foundations of comparative advantage*. Oxford University Press, Oxford
- Hayek FA (1960) *The constitution of liberty*. University of Chicago Press, Chicago
- Keech WR, Munger MC (2015) The anatomy of government failure. *Public Choice* 164:1–42
- Keynes JM (1936) *The general theory of employment, interest and money*. Macmillan, London
- Krugman P (2013) *End this depression now!* W.W. Norton & Company, New York
- La Porta R, Lopez-de-Silanes F, Shleifer A (2008) The economic consequences of legal origins. *J Econ Lit* 46:285–332
- Le Grand J, New B (2015) *Government paternalism: nanny state or helpful friend?* Princeton University Press, Princeton
- Leeson PT (2014) Pirates, prisoners, and preliterate: anarchic context and the private enforcement of law. *Eur J Law Econ* 37:365–379
- Leibfried S, Huber E, Lange M, Levy JD, Nullmeier F, Stephens JD (eds) (2015) *The Oxford handbook of transformations of the state*. Oxford University Press, Oxford
- Lewis-Beck MS, Stegmaier M (2013) The VP-function revisited: a survey of the literature on vote and popularity functions after over 40 years. *Public Choice* 157:367–385
- Marx K, Engels F (1848/2002) *The communist manifesto*. Penguin Books, London
- Mueller DC (2003) *Public choice III*. Cambridge University Press, Cambridge
- Obinger H, Schmitt C, Traub S (2016) *The political economy of privatization in rich democracies*. Oxford University Press, Oxford
- Ostrom E (2010) Beyond markets and states: polycentric governance of complex economic systems. *Am Econ Rev* 100:641–672
- Persson T, Tabellini G (2003) *The economic effects of constitutions*. MIT Press, Cambridge
- Piketty T (2014) *Capital in the twenty-first century*. Harvard University Press, Cambridge
- Ravenhill J (ed) (2016) *Global political economy*, 5th edn. Oxford University Press, Oxford
- Schnellenbach J, Schubert C (2015) Behavioral political economy: a survey. *Eur J Polit Econ* 40:395–417
- Smith A (1776/1981) *An inquiry into the nature and causes of the wealth of nations*. Liberty Fund, Indianapolis
- Stiglitz JE, Rosengard JK (2015) *Economics of the public sector*, 4th edn. W.W. Norton & Company, New York
- Stringham EP (2015) *Private governance: creating order in economic and social life*. Oxford University Press, Oxford
- Towfigh EV, Petersen N (eds) (2015) *Economic methods for lawyers*. Edward Elgar, Cheltenham
- Vanberg VJ (2005) Market and state: the perspective of constitutional political economy. *J Inst Econ* 1: 23–49
- Vanberg VJ (2015) Ordoliberalism, Ordnungspolitik, and the reason of rules. *Eur Rev Int Stud* 2:27–36
- Wolf C (1993) *Markets or governments: choosing between imperfect alternatives*, 2nd edn. MIT Press, Cambridge, MA

Political Violence

► Terrorism

Politicians

Benny Geys^{1,2} and Karsten Mause³

¹Department of Economics, Norwegian Business School BI, Oslo, Norway

²Vrije Universiteit Brussel, Ixelles, Belgium

³Department of Political Science (IfPol), University of Münster, Münster, Germany

Definition

Politicians are persons involved in the process of public policymaking in their role as members of governments, parliaments, political parties, and other political bodies at the (sub)national level (e.g., local government, state legislature, national parliament, etc.) as well as within the supranational political arena (e.g., United Nations Security Council, European Union institutions, and so on). Many politicians get into office through a democratic election, while others are selected or appointed to a public office. For many politicians, politics is a full-time job, while for others, it remains an activity in addition to a main job as a lawyer, teacher, or entrepreneur. This entry takes stock of what law and economics (LE) scholars have contributed to the large social science literature on politicians.

The Law and Economics of Politicians

Analyzing what politicians do belongs to the “core business” of political scientists. However, the law and economics discipline has extensively contributed to this multidisciplinary research field as well. Given the large number of articles and books, it is beyond the scope of this entry to review everything that LE scholars have written about “politicians.” Rather, we focus on the

following aspects: (1) the motivations of politicians, (2) politicians as lawmakers, (3) politicians and interest groups, and (4) the design and effects of the (legal) rules under which politicians act.

What Do Politicians Maximize?

LE scholars including, for instance, George J. Stigler and Richard A. Posner are also active contributors to the field of “public choice.” Public choice theory in the tradition of Downs (1957) breaks with the welfare-economic assumption of benevolent governments working in the public interest and maximizing the wealth of the nation. Instead, it assumes that politicians and public bureaucrats (i) are primarily interested in maximizing their individual utility and (ii) “act solely in order to attain the income, prestige, and power which come from being in office” (Downs 1957, p. 28). In other words, it is theoretically assumed that political decision-makers are self-interested not only when they make private choices (as consumers, investors, landlords, and so on) but also when they make choices in government, parliamentary committees, and other political contexts (Mueller 2003; Posner 2014, chap. 20).

Whether real-world politicians really behave in line with predictions derived from the *homo-oeconomicus* assumption used in economic theories is, of course, an empirical question – and presumably depends on the specific context within which the individual politician acts. In democracies, an individual politician’s private utility maximization is usually constrained by a number of mechanisms that may channel his/her self-interest and act as a disciplining device, e.g., his/her political party, political competition, press freedom, an independent judiciary, nongovernmental watchdog organizations, a politically interested electorate, and so on. However, in the absence of such constraints, a public choice perspective would lead to the expectation that politicians could have a low or even no incentive to pursue the public interest (Mueller 2003; Bovens et al. 2016).

Within the vast LE literature, there are a number of empirical studies indicating that many politicians “shirk” their duties in their last period

in office, that is, once citizen-voters are no longer able to punish them at the next election (see Geys and Mause 2016, for a survey of the “shirking” literature). Moreover, there is a large literature on “political corruption” demonstrating that politicians at times illegally abuse their power for their private gain (money, privileges, jobs, and so on). LE researchers have made numerous contributions to this literature, discussing in particular the question of whether changes in the incentive structures under which politicians act (e.g., more severe punishment, transparency laws, better pay, etc.) help to reduce the probability of corruptive practices occurring (Rose-Ackerman and Palifka 2016).

Politicians as Lawmakers

Otto von Bismarck (1815–1898), the first Chancellor of the German Empire lasting from 1871 to 1918, once said: “Laws are like sausages. It is better not to see them being made.” Obviously many social scientists have ignored Bismarck’s advice since there is a large literature analyzing different aspects of the role of politicians as “makers” of public policy and “producers” of constitutions, laws, regulations, and other outputs of the political process (for surveys, see McCubbins et al. 2007; Parisi 2011; Posner 2014, chap. 20). LE scholarship has contributed to this multidisciplinary research field especially by analyzing and rethinking the system of checks and balances in which policymakers in different political systems are embedded. This literature suggests that it makes a difference whether policymakers “produce” laws and regulations in a representative democracy, a parliamentary democracy with a strong second chamber, a presidential democracy, a direct democracy with a powerful electorate, and so on (Mueller 2003; Persson and Tabellini 2004). Moreover, supranational actors – such as the European Union (EU), the North Atlantic Treaty Organization (NATO), and the United Nations (UN) – have been found to influence (sub)national policymakers’ decisions (Dreher and Lang 2016).

An interesting phenomenon in this context is that some political systems allow politicians to

make laws and regulations with direct implications for themselves. For example, politicians active as lawyers occasionally become involved in making laws that favor the legal profession (Matter and Stutzer 2015). Likewise, the desire of political parties to establish and maintain electoral thresholds may be interpreted as an anti-competitive instrument used to erect a barrier to entry for new parties (Wohlgemuth 1999). A closely related issue is whether politicians should legally be allowed to set their own salaries and other parts of their compensation package – as is currently often the case (Mause 2014). As noted above, problems might occur if politicians’ self-interest and the public interest collide. If citizens get the impression that politics is some kind of “self-service store” for the “political class,” this might have an effect on voter turnout, election results, and trust in politicians and the political system as a whole.

Politicians and Interest Groups

Within any democratic society, interest groups (such as firms, trade unions, employers’ associations, taxpayers’ associations, churches, rabbit breeders associations, rifle associations, and so on) have a strong incentive to lobby for their interests and achieve favorable regulations, government subsidies, protection of monopolies, and other things. Imagine, for example, the case of a powerful association of business enterprises that would prefer to keep its market closed to competitors. Any market entry barriers benefit the incumbent firm(s) as they reduce the level of competition in that particular sector of the economy. This is likely to result in higher prices (and lower consumer surplus) compared to a situation with more competition, but in higher profits for the firms active in this market. Clearly, this business association would have a strong incentive to ensure – via various lobbying activities – that policymakers do not open the market. In the public choice and LE literatures, the phenomenon sketched above is denoted as “rent-seeking” (see Congleton and Hillman 2015, for a survey of the voluminous theoretical and empirical rent-seeking literature).

Politicians are a natural objective for lobbying activities, which can take the form of, for example, information campaigns, dinner invitations, or political (campaign) donations. While certain lobbying activities can be beneficial by improving the information available to policymakers, they can also become a serious societal problem if they influence legislators, legislation, and/or the allocation of public funds in favor of specific groups (Hodler and Raschky 2014). Such political favoritism indeed distorts spending allocations away from the normative principles that ideally drive them, which from a theoretical perspective “lowers aggregate social welfare, [and] creates inequality across social groups” (Bramoullé and Goyal 2016, p. 23). As a result, a vast literature has investigated the link between lobbying activities and political outcomes. For example, studies linking firms’ political donations to congressional voting patterns and public procurement contracts (see, e.g., Gherghina and Volintiru 2017, including a literature review).

As noted above, it is not only business firms and other special interest groups that may try to influence policymakers to make decisions favorable to their particular firm or industry. In this context, a specific type of “rent-seeking” by politicians occurs when the latter use their influence and power to obtain and defend special privileges that “ordinary citizens” do not have (Mause 2014).

Getting Incentives Right: Regulating Politicians

Scholars working at the intersection of LE and constitutional political economy in the tradition of Buchanan (2008) and others tend to maintain a strong focus on the “rules of the game” for politicians. There is, for instance, a substantial literature analyzing how the design of the working conditions in the political system (e.g., wages, term limits, etc.) incentivize individuals to stand for election – and drive politicians to make policies in the public interest. For example, the fact that candidates for a public office have to disclose parts of their income and/or wealth to the public is argued to have an effect

on the incentive to run for office (Djankov et al. 2010; van Aaken and Voigt 2011; Braendle 2016). This also touches on the issue of accountability regarding “shirking” politicians and how to deal appropriately with such behavior (Geys and Mause 2016). Given the potential (self-) selection effects this induces, should there be compulsory attendance for politicians at parliamentary debates, roll-call votes, or committee meetings (possibly with penalties for nonattendance)?

Likewise, there remains an intense debate as to whether existing regulations with respect to donations to politicians and political parties, politicians’ sideline jobs, partisan patronage and favoritism, and political corruption are sufficient (see Rose-Ackerman and Palifka 2016, for an introduction to this research field). The rules of the political game may be respected by many politicians, but there will most likely always be at least some “black sheep” that break the rules and behave “opportunistically” in the sense of Williamson (1985, p. 47): “By opportunism I mean self-interest seeking with guile. [...] More generally, opportunism refers to the incomplete or distorted disclosure of information, especially to calculated efforts to mislead, distort, disguise, obfuscate, or otherwise confuse.”

Cross-References

- [Political Competition](#)
- [Political Corruption](#)
- [Public Choice: The Virginia School](#)
- [Rent Seeking](#)

References

- Bovens M, Goodin RE, Schillemans T (eds) (2016) *The Oxford handbook of public accountability*. Oxford University Press, Oxford
- Braendle T (2016) Do institutions affect citizens’ selection into politics? *J Econ Surv* 30:205–227
- Bramoullé Y, Goyal S (2016) Favoritism. *J Dev Econ* 122:16–27
- Buchanan JM (2008) Same players, different game: how better rules make better politics. *Constit Polit Econ* 19:171–179

- Congleton RD, Hillman AL (eds) (2015) *Companion to the political economy of rent seeking*. Edward Elgar, Cheltenham
- Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A (2010) Disclosure by politicians. *Am Econ J App Econ* 2:179–209
- Downs A (1957) *An economic theory of democracy*. Harper & Row, New York
- Dreher A, Lang VF (2016) The political economy of international organizations. CESifo Working paper no. 6077
- Geys B, Mause K (2016) The limits of electoral control: evidence from last-term politicians. *Legis Stud Q* 41:873–898
- Gherghina S, Volintiru C (2017) A new model of clientelism: political parties, public resources, and private contributors. *Eur Polit Sci Rev*. 9(1):115–137. <https://doi.org/10.1017/S1755773915000326>
- Hodler R, Raschky PA (2014) Regional favoritism. *Q J Econ* 129:995–1033
- Matter U, Stutzer A (2015) The role of lawyer-legislators in shaping the law: evidence from voting on tort reforms. *J Law Econ* 58:357–384
- Mause K (2014) Self-serving legislators? An analysis of the salary-setting institutions of 27 EU parliaments. *Constit Polit Econ* 25:154–176
- McCubbins MD, Noll RG, Weingast BR (2007) The political economy of law. In: Polinsky AM, Shavel S (eds) *Handbook of law and economics*, vol 2. North-Holland Publishing, Amsterdam, pp 1651–1738
- Mueller DC (2003) *Public choice III*. Cambridge University Press, Cambridge
- Parisi F (ed) (2011) *Production of legal rules*. Vol. 7 of *encyclopedia of law and economics*, 2nd edn. Cheltenham, Edward Elgar
- Persson T, Tabellini G (2004) Constitutions and economic policy. *J Econ Perspect* 18:75–98
- Posner RA (2014) *Economic analysis of law*, 9th edn. Wolters Kluwer Law & Business, New York
- Rose-Ackerman S, Palifka BJ (2016) *Corruption and government: causes, consequences, and reform*, 2nd edn. Cambridge University Press, Cambridge
- van Aaken A, Voigt S (2011) Do individual disclosure rules for parliamentarians improve government effectiveness? *Econ Gov* 12:301–324
- Williamson OE (1985) *The economic institutions of capitalism. Firms, markets, relational contracting*. The Free Press, New York
- Wohlgemuth M (1999) Entry barriers in politics, or: why politics, like natural monopoly, is not organised as an ongoing market-process. *Rev Austrian Econ* 12:175–200

Polycentric Law

► Customary Law

Porn

► Pornography

Pornography

Samuel Cameron

Division of Economics, University of Bradford,
Bradford, UK

Abstract

This entry discusses the problems of definition of pornography and the problems of defining an optimal regulatory strategy.

Synonyms

Porn

Judge Richard A. Posner has repeatedly applied orthodox free market Chicagoan law and economic thinking to porn in several of his books (*Sex and Reason*, *Frontiers of Legal Theory*, *Economic Analysis of Law*). Heterodox economist David George (2001) briefly discusses porn as an instance of what he calls “preference pollution.” In this scenario, free markets impose weakness of will upon us leading to the “wrong choice” of consumption basket. We end up less well-off than we might otherwise have been because we are weak and fall prey to market pressures to consume porn. As we are now over a decade further on in the Internet world, his point would presumably be made even more forcefully now.

Regulation of porn, by laws, requires a definition of the activity so that it can be actioned.

This is a long-running problem. Everyone thinks they know what the word “pornography” means, yet legal definitions have been difficult to agree upon. This problem increases as laws come

Samuel Cameron has retired.

into force relying on gradation into extreme or hard-core porn as illegal and lesser types as not. The original meaning was as a description of the life and activities of prostitutes. This derives from the Greek, *pornographos* (*porne* = prostitutes and *graphein* = to write). It acquired its present meaning in the nineteenth century and seems to have two key elements, one being the obscenity of depiction of sexual acts and the other, the exploitation of workers in the production of the material and/or the consumers of the material. There are thus two areas of legislation. The first is to deal with protection of consumers, in their own best interests, and the second to deal with the workers in the industry. The latter has been particularly important in the American economy and less developed in some other territories. A notable instance of this is to be found in the 1995 Child Protection and Obscenity Enforcement Act, passed in the USA following the 1986 report of the Meese Commission. This required porn filmmakers to keep detailed records proving that no underage performers were used.

Much legislation has tended to be under the rubric of "obscenity." Porn has frequently been treated in legal matters as material "with a tendency to deprave or offend" where the decision on how depraving or offensive the matter is being left to a judge and/or jury. American Supreme Court decisions have illustrated the difficulties of Justice Potter Stewart's claim that he would know pornography "when he saw it." Lurking in the Supreme Court decisions has been the notion of an "average Justice Brennan's so-called 'limp dick' test became entrenched as a dividing line in movie censorship person" standard, that is, the judges are agents representing the median preferences of the population.

In terms of welfare economics, the simplest case for control is that porn may be viewed, as mentioned earlier, as a form of "social pollution." Controls can be aimed at the supply or at the demand. As with general crime, the expectation is that punishment will be proportional to harm. The particular dividing line in this respect is with respect to child pornography. The era of more liberal attitudes toward this in some areas of Europe (particularly in Denmark) has now passed.

Most of Europe makes child porn illegal but many parts of the world do not have explicit laws against it. Part of the more liberal attitude, in the past, related to deterrence and safety valve effects. If punishment does not deter then it will occur costs which bring no benefit other than the public satisfaction with retribution and the possible enhanced safety of taking the offenders out of circulation. The latter effect will be negligible for child (or any porn) if there is no complementarity between porn consumption and other offenses. If the offenses are a substitute good for porn consumption, then society can potentially have net benefit from liberal treatment of offenders. This was the core point of the criminologist Berl Kutchinsky about the Danish experience (see Kutchinsky (1985); Kutchinsky and Snare (1999)).

Various empirical studies have been carried out on the possible relationships between availability of pornography and the frequency of sexual crimes. Ad hoc inferences have also been made about cross-country differences in such offending between repressive and more liberal regimes. Due to problems of data accuracy and consistency and difficulties in causality testing, there is no conclusive evidence from these studies that pornography does have a harmful effect. In any case, legislation and policy proceed largely parallel to evidence. The chief force for change in regulation is the continued evolution of Internet markets.

Supply-side regulation is inherently difficult in Internet markets. Thus demand-side controls in Internet markets have been growing following the style of those for circulating media. For printed magazines, we have seen requirements for shrink-wrapping in an opaque cover and/or location in a position in a shop where it is not easily seen, by accident, by children or others likely to take offense. In the case of broadcast porn, we have seen restriction to channels that are not "free to air" (FTA) and can therefore be blocked to inappropriate consumers. However there has been a proliferation, in many countries, of FTA "Babestation" channels which grew up under the rubric of being nonsexually

offensive chat/social entertainment sites. Frequent complaints against breach of this in the UK have been lodged with relatively few fines being levied by the broadcast regulators. These stations continue to broadcast on FTA television material which regular channels would not be allowed to provide. Such channels operate also on the Internet where they provide more and more graphic content. This is of course material that would be legal as porn per se. It would not be an offense to possess copies for private use.

Prevention of access to Internet porn requires that the would-be consumer is unable to access the site. This can be accomplished by specific software that refuses connection to a site. Such software has been mainly aimed at parents wishing to prevent children seeing porn. ISP (Internet service provider) can also lock certain sites as unsuitable.

A fundamental difficulty with expanded regulation is the potential loss of revenue to ISP if blocking is used especially if “opting in to porn” on user sign-up is invoked. This does not come from sales of porn as such but from the degree of Internet traffic it generates which proportionally raises the advertising-based revenue streams. ISP may not be keen on surveillance but its use to target offenders is of much less revenue threat and is very difficult to resist both technically and also politically due to the threats of terrorism.

From a legal point of view, there is a change in some recent results of police seizure of people’s computers and Internet history records. That is, intent is coming to the fore as a ground for prosecution, as opposed to mere possession. In 2014 arrests began from a global initiative involving 19 countries which explored data from around 10,000 Internet users had accessed more than 30 websites carrying pedophilia. For practical reasons, the number of suspects was narrowed down to just over 400 suspects of distributing pedophile images. The 19 countries where warrants were executed were Australia, Belgium, Canada, France, Germany, Israel, Italy, Japan, Korea, the Netherlands, New Zealand, Portugal, Russia, Spain, Sweden, Taiwan, Turkey, the UK and the USA. One of the cases in the 2014 trial of

entertainer, Rolf Harris, in the UK for sex offenses was brought on grounds of accessing child porn. His Internet searches were offered as evidence by the prosecution on the grounds that he had clearly used words intended to find sexual images of children. His defense pointed out that the models, on such sites, were simply posing as underage given that compliance with employment directives was observed by these sites. Hence, on a consumption-based charge, he was innocent. This was upheld but we see here serious grounds being given for a charge based on intent.

The above measures are intrinsically based on restricting the quantity of a good which has “bad” effects. They incur costs and thus it is possible that there could be “too much” prevention of porn in the sense that aggregate social benefit, of so doing, exceeds aggregate social cost. The costs may include wrongful arrest and prosecution.

Traditional economic models might suggest that price regulation is likely to be more efficient than quantity regulation. A porn tax might seem attractive (analogously to a Pigouvian pollution tax), but in terms of more modern welfare economics, the option of selling permits, or licenses, to porn merchants would seem a better price-based solution. While the sale value of the permits could be determined at a level that is social welfare maximizing, both approaches face the difficulty of being seen as dangerous due to making porn seem like a legitimate business activity.

References

- George D (2001) *Preference pollution: how markets create the desires we dislike*. The University of Michigan Press, Ann Arbor
- Kutchinsky B, Snare A (ed) (1999) *Law, pornography and crime – the Danish experience*, Scandinavian studies in criminology, vol 16. Pax Forlag, Norway.. www.krim.ku.dk/Jesperkrim.ku.dk/Jesper
- Kutchinsky B (1985) *Pornography and its effects in Denmark and the United States: a rejoinder and beyond*. Comparative Social Research, JAI Press, Greenwich, CT

Positional Goods and Legal Orderings

Ugo Pagano^{1,2} and Massimiliano Vatterio³

¹University of Siena, Siena, Italy

²Central European University, Budapest, Hungary

³“Brenno Galli” Chair of Law and Economics, Institute of Law (IDUSI), Università della Svizzera italiana (USI), Lugano, Switzerland

Upon passing by a small village of barbarians,

Julius Caesar asserted

“[f]or my part, I had rather be the first man among these fellows than the second man in Rome”.

(Plutarch)

Abstract

People consume because others consume, maintained Veblen in 1899. More recently, theoretical, empirical, and experimental articles have argued that people constantly compare themselves to their environments and care greatly about their relative positions.

Given that competition for positions may produce social costs, we adopt a *Law and Economics* approach (i) to suggest legal remedies for positional competition and (ii) to argue that, because legal relations are characterized in turn by positional characteristics, such legal remedies do not represent “free lunches.”

The Issue

A positional good is an economic good that depends largely on comparison of one’s own consumption with that of others (Hirsch 1976; McAdams 1992; Pagano 1999; Hopkins and Kornienko 2004; Schneider 2007; Vatterio 2009; Fiorito and Vatterio 2013). Positional concerns refer to the fact that individuals consider their rank, relative standing, or position when they evaluate their situation and act upon this evaluation.

For instance, in the above quotation from Plutarch, Julius Caesar (Roman emperor) admitted his willingness to renounce the larger and better

private and public goods that he could consume in Rome (e.g., *panem et circenses*) to gain the position of sovereign (which is a positional good) in a relatively poorer community. Likewise, in John Milton’s *Paradise Lost*, Satan stated that it is “[b]etter to *reign* in Hell, than *serve* in Heaven” (emphasis added). In other words, individuals like Caesar and angels like Satan prefer to be the first or the leader in a reference group, even if it is a second-rate one (e.g., a barbarian village or hell). The recent experimental literature corroborates this “position matters” argument (e.g., Solnick and Hemenway 1998, 2005).

Similarly, Hopkins and Kornienko (2004) state, “it is not just that the car is big [enough for needs] but that it is bigger than those owned by the neighbours that also matters” (Hopkins and Kornienko 2004:1087–1088). Carlsson et al. (2007) test this size-matters hypothesis with an experiment and find that people prefer cars bigger than the average size in their society.

A further example is provided by San Gimignano, a small Italian town close to Florence and Siena. In the past, San Gimignano was characterized by about 80 towers (today there are “only” 20 of those towers remaining). The families of San Gimignano did not build towers to live in them or for military defense (because they were unfit for habitation or for fortification) but rather to display their power, affluence, wealth, and status to the rest of the community. Similarly to the “size-matters” argument in Hopkins and Kornienko, in the case of San Gimignano’s towers, tallness matters.

Positional concerns are pervasive in numerous socioeconomic domains (see Solnick and Hemenway 2005) and play a pivotal role in people’s happiness (e.g., Frey and Stutzer 2002; Clark et al. 2008). This entry wants to deal with negative effects of positional competition and institutional remedies.

Some consequences of competition for positions are illustrated by the following example concerning a labor relationship (see Frank 2012). Consider two types of employment contract, distinguished between wage and safety at work. The former type of contract is characterized by a wage relatively higher *but* a level of workplace safety

relatively lower than those of the latter type. Denoting with w_L , w^H , and R , respectively, the low(er) wage, the high(er) wage, and disutility for unsafe work (e.g., risks of injuries) assume that

$$w_L > w^H - R \tag{1}$$

That is, the disutility due to an unsafe workplace is relatively higher than the increase in wage. In other words, the choice of the contract with unsafe work may produce social costs (i.e., higher health expenditures).

In this game, the Nash equilibrium (*safe work and low wage* for both parties) is efficient (see Fig. 1).

Now assume that positions matter. For instance, a worker with a higher wage can acquire a relatively larg(er) car: that is, there is a positive utility POS deriving from position (e.g., being the worker with the highest wage in the reference group) as well as a corresponding negative utility – POS (for being the worker with the lowest wage in the reference group). The choice of the worker changes as in Fig. 2.

In particular, if $w_L - \text{POS} < w^H - R < w_L < w^H - R + \text{POS}$, then the game becomes a prisoner’s dilemma: the Nash equilibrium (both

parties choose *unsafe work and high wage*) is Pareto inefficient.

Indeed, even if the second type of contract (lower wage *but* a higher level of workplace safety) is efficient in terms of social welfare (because health expenditures with lower workplace safety are relatively higher than the increase in wage, cf. Eq. 1), each worker prefers the first type of contract because positional competition (on the wage) is important. This means that agents may substitute a non-positional good (i.e., safety at work) with a positional good (wealth), thus causing social costs (i.e., health expenditures). Moreover, if every employee chooses the employment contract with a high wage, then no agent will enjoy a positional advantage in wealth. In equilibrium, no one will consume a positional good (i.e., the biggest car), but all workers will “consume” a lower level of safety at work.

Hence, the consumption of positional goods may produce the following social costs:

1. Agents could substitute a non-positional good (e.g., a private and/or public good) with a positional good, leading to suboptimal equilibria (see also Frank 1985a, b, 2008).

Positional Goods and Legal Orderings, Fig. 1 The trade-off between safety (or health) and wage

		Friday	
		Safe work and low wage	Unsafe work and high wage
Robinson	Safe work and low wage	$w_L; w_L$	$w_L; w^H - R$
	Unsafe work and high wage	$w^H - R; w_L$	$w^H - R; w^H - R$

		Friday	
		Safe work and low wage	Unsafe work and high wage
Robinson	Safe work and low wage	$w_L; w_L$	$w_L - \text{POS}; w^H - R + \text{POS}$
	Unsafe work and high wage	$w^H - R + \text{POS}; w_L - \text{POS}$	$w^H - R; w^H - R$

Positional Goods and Legal Orderings, Fig. 2 The choice when positions do matter

2. Because the “parallel investments” in obtaining positional goods may lead to the situation in which no one consumes positional goods, positional competition may waste the economic resources of agents (see Pagano 1999).

How should institutions regulate the competition for positions? What are institutional remedies? A first group, say of *Public Economics*, involves Pigouvian remedies, i.e., a progressive consumption tax on luxury/positional goods (see Frank 1985a, b, 2008). A second group, say of *Law and Economics* (see definition of Marciano 2016, ► “[Economic Analysis of Law](#)”), may consider rules as means to reduce social costs due to positional competition. This second group is the focus of the next two sections.

Institutional Remedies

How could norms reduce social costs due to positional competition? We investigate three types of *Law and Economics* remedies: norms which (i) restrain/punish the consumption of positional goods, (ii) make the consumption of non-positional goods compulsory, and (iii) encourage cooperation by agents competing in positions.

(i) *Restraining the consumption of positional goods*

Because of the social costs deriving from the consumption of positional goods, the lawmaker may penalize or prohibit the consumption of positional goods. An example is the so-called sumptuary law – Black’s Law Dictionary defines such a law as made for the purpose of restraining luxury or extravagance, particularly against inordinate expenditures in the matter of apparel, food, furniture, etc. Sumptuary laws were enacted in Ancient Greece and Rome, from the Middle Ages onward in France and England, and again, in the seventeenth century, in the American Colonies and in feudal Japan (cf. Dari-Mattiacci and Plisecka 2012).

For instance, in ancient Rome, a series of laws governed the materials of which garments could be made and the number of guests at entertainments. In France, Philip IV issued regulations governing the dress and the entertainments of the various social orders. Under later French kings, the use of gold and silver embroidery, silk fabrics, and fine linen was restricted. In 1433, an act of the Scottish Parliament prescribed the lifestyle in Scotland, even going so far as to limit the consumption of pies and baked meats. In feudal Japan, an imperial edict regulated the size of houses and imposed restrictions on the materials that could be used in their construction. Rules in the Tokugawa period in Japan specified the sorts of toys that parents could give their children. For the Aztecs, *macehualtin* – members of the laboring class who displayed finery and precious objects – could be put to death. An interesting and “romantic” view on sumptuary law is in *Dei Sepolcri* by Ugo Foscolo, where the author condemned the Napoleon edict of Saint-Cloud on *positional* expenditures for funerals.

However, attitudes on sumptuary law changed with the Enlightenment, industrial mass production, and consumer-oriented societies. For this reason, there are today limited examples in which legal norms discourage or punish the consumption of a positional good – an exception could be represented by the case of the case of “power”: Power is a positional good (Pagano 1999; Vatiéro 2009) whose consumption in liberal countries is legally and formally limited, e.g., by antitrust law, labor law, and public law (see Vatiéro 2009), which may represent modern forms of sumptuary norms.

Although in liberal society there are not clear examples of legal norms which punish the consumption of positional competition, social norms may do so. This is the case of the tenth commandment “You shall not covet your neighbor’s house; you shall not covet your neighbor’s wife or his servant or his ox or his donkey or anything that belongs to your neighbor.” Because the commandment implies punishment in the case of *envious* choices and conducts, it should affect the behaviors of agents, at least for Christians and Hebrews, and may mitigate positional competition.

(ii) *Making the consumption of non-positional goods compulsory*

Because of the social costs due to the consumption of positional goods, the lawmaker may render (a minimum level of) the consumption of non-positional goods compulsory. That is, the legal system may provide norms which reduce the substitution between positional goods and non-positional goods by defining a non-renegotiable minimum consumption of non-positional goods.

Indeed, labor law establishes non-renegotiable minimum conditions on safety at work. For instance, the employer must provide and maintain a working environment that is safe and without risk to the health of the workers; and workers must use safety equipment with care and act according to prescribed instructions to safeguard their health and protect themselves against injury.

Accordingly, most labor law in the Western countries forbids the renegotiation of these conditions on safety, even against the worker's will. Indeed, the worker may prefer a higher wage at the expense of safety at work. From the perspective of positional competition, this prohibition on renegotiating minimum safety conditions at work is efficient because it reduces the emergence of social costs related to positional competition.

(iii) *Encouraging cooperation among positional competitors*

Because of the social costs due to the consumption of positional goods, the policy maker may encourage cooperation and collaboration among agents (i.e., positional competitors). That is, the wasteful competition for positional goods represents a problem of coordination among agents. For instance, in the case of employment contracts, workers' choices lead to a Pareto-inefficient equilibrium with a lower level of safety at work and nobody consuming a positive level of positional goods. Workers may coordinate to move to a Pareto-superior equilibrium (where nobody still consumes a positive level of positional goods, but all parties have higher levels of safety at work). This implies that cooperation among agents (e.g., workers) may improve the efficiency

of individuals' choices (e.g., in terms of labor safety).

In other words, the labor laws which sustain worker unions and collective bargaining could encourage coordination/cooperation among workers and, because unions take the negative effects of positional competition into account, reduce the substitution of non-positional goods for positional goods.

Legal Positions as Positional Goods

While in the preceding section we argued that legal norms may diminish inefficiencies due to positional competition, in this section we illustrate how legal norms in turn create positional concerns.

According to legal theorists such as Wesley N. Hohfeld and old institutionalists like John Commons, each jural relation is linked to a jural correlative (Pagano 2000; Vatiéro 2010; Fiorito and Vatiéro 2011). The above figure displays the eight fundamental conceptions with which all legal problems may be stated: claim, duty, liberty, no right, power, liability, immunity, and disability (Fig. 3).

Claim/duty relation. In a simplified situation with two agents, say a Dominus and a Servus, a *claim* means that the Dominus has a state-sanctioned assurance that the Servus will behave in a certain way toward Dominus. However, this occurs if and only if the Servus has the *duty* to engage in such behavior with respect to Dominus. That is, a duty is the legal position of the Servus, who is commanded by society to act for the benefit of Dominus, and who will be penalized by society for disobedience. Hence, the correlative of a claim is a duty.

Liberty/no-right relation. *Liberty* stands for one's freedom from the claim of someone else.

	Jural correlatives			
<i>Dominus</i>	Claim	Liberty	Power	Immunity
<i>Servus</i>	Duty	No-right	Liability	Disability

Positional Goods and Legal Orderings,
Fig. 3 Fundamental legal conceptions

Similar to the claim-duty relationship, the Dominus has a liberty to behave in a certain way toward Servus if and only if the Servus has *no-right* toward the Dominus to prevent the Dominus from behaving in a certain way. No-right is therefore the legal correlative of a liberty of another party.

Power/liability relation. *Power* is the legal ability to do certain acts that alter legal relations. The Dominus’s power is when the Dominus’s own voluntary act will cause new legal relations between the Dominus and the Servus, against the Servus’s will. This implies that, whenever a power exists, there is at least one other human being whose legal relation will be altered when the power is exercised. The person whose legal relation will be altered is under a *liability*.

Immunity/disability relation. Finally, *immunity* is any legal situation in which a given relation vested in one person cannot be changed by acts of another person. Correlatively, the one who lacks the legal ability to alter the other individual’s legal relations is said to be under a *disability*.

The correlative nature of legal positions means that each legal position is available to an individual if and only if a corresponding legal position is occupied by some other individual. In particular, legal positions are adversarial in nature (see also Vatiéro 2013a). For instance, claims of one individual imply, at the same time, duties for some another individual and vice versa. That is, the set of actions that defines the claims of the Dominus imposes duties on some individual(s), e.g., the Servus. This brings about the consumption of legal positions with opposite signs: claim by the Dominus, as his/her desired output, and duty by the Servus, as his/her costly input. In a similar manner, the power-liability relationship consists of Dominus’s benefit (i.e., the power) as well as Servus’s cost (i.e., liability). Hence, any utility deriving from rights and powers must jointly relate with disutility deriving from duties and liabilities (see Fig. 4).

Unlike traditional economic goods, jural positions inevitably involve consumptions and utilities with opposite signs. *Everyone* cannot

Claim/Power	Duty/Disability
<i>Benefits</i>	<i>Costs</i>
<i>Desired output</i>	<i>Costly input</i>
<i>Utility</i>	<i>Disutility</i>

Positional Goods and Legal Orderings, Fig. 4 The adversarial nature of legal relations

consume claims, liberties, powers, and immunities; for some individuals, the exercise of these jural positions must imply the exercise of “unfavorable” correlative jural relations (i.e., duties, no rights, liabilities, and disabilities). Because of their adversarial nature, we can represent jural positions as positional goods (see also Pagano 2000; Vatiéro 2013a).

This means, moreover, that the definition of rights, duties, powers, etc. with the purpose of mitigating positional concerns, as illustrated in the previous section, can in turn create positional concerns because judicial positions are positional goods. Following Coasean main contribution, no institution is a “free lunch” (Grillo 1995; Pagano 2012; Vatiéro 2013b; Pagano and Vatiéro 2015). Thus, in the case of positional goods one should take into account benefits deriving from an institutional arrangement able to mitigate positional competition and reduce related social costs but also the costs involved in that institutional arrangement which is in turn characterized, owing to the adversarial nature of legal relations, by positional concerns.

Conclusions

Positions matter in the choices and happiness of agents. People constantly compare themselves to their environments and care greatly about their relative positions, which impact on their choices.

The consumption of positional goods may produce social costs. On the one hand, agents may substitute a non-positional good with a positional good; on the other hand, as an arms race, positional competition may waste the economic resources of positional competitors. Both consequences lead to suboptimal equilibria.

Law and Economics scholars should investigate legal remedies for social costs due to competition for positions. They should take serious account of the fact that legal remedies are not free lunches. In this regard, future research could investigate the costs and benefits that would derive from the definition of a Coasean market (Cf. Medema 2014, ► [“Coase Theorem”](#); Bertrand 2015, ► [“Coase and Property Rights”](#) in this encyclopedia) for positional goods.

Cross-References

- [Coase and Property Rights](#)
- [Coase Theorem](#)
- [Economic Analysis of Law](#)
- [Externalities](#)
- [Institutional Economics](#)

References

- Carlsson F, Johansson-Stenman O, Martinsson P (2007) Do you enjoy having more than others? Survey evidence of positional goods. *Economica* 74:586–598
- Clark AE, Frijters P, Shields MA (2008) Relative income, happiness, and utility: an explanation for the Easterlin paradox and other puzzles. *J Econ Lit* 46(2):425–467
- Dari-Mattiacci G, Plisecka AE (2012) Luxury in ancient Rome. An economic analysis of the scope, timing and enforcement of sumptuary laws. *Leg Roots, Int J Roman Law, Leg Hist Comp Law* 1:189–216
- Fiorito L, Vatterio M (2011) Beyond legal relations: Wesley Newcomb Hohfeld’s influence on American institutionalism. *J Econ Issues* 45(1):199–222
- Fiorito L, Vatterio M (2013) A joint reading of positional and relational goods. *Econ Politica – J Anal Inst Econ* 30(1):87–96
- Frank RH (1985a) The demand for unobservable and other nonpositional goods. *Am Econ Rev* 75(1):101–116
- Frank RH (1985b) Choosing the right pond: human behavior and the quest for status. Oxford University Press, New York
- Frank RH (2008) Should public policy respond to positional externalities? *J Public Econ* 92:1777–1786
- Frank RH (2012) *The Darwin economy*. Princeton University Press, Princeton
- Frey BS, Stutzer A (2002) What can economists learn from happiness research? *J Econ Lit* 40:402–435
- Grillo M (1995) Introduzione. In: Grillo M (ed) *Impresa, Mercato e Diritto*. Bologna, Mulino
- Hirsch F (1976) *The social limits to growth*. Harvard University Press, Cambridge
- Hopkins E, Kornienko T (2004) Running to keep in the same place: consumer choice as a game of status. *Am Econ Rev* 94(4):1085–1107
- McAdams RH (1992) Relative preferences. *Yale Law J* 102(1):1–104
- Pagano U (1999) Is power an economic good? Notes on social scarcity and the economics of positional goods. In: Bowles S, Franzini M, Pagano U (eds) *The politics and the economics of power*. Routledge, London, pp 116–145
- Pagano U (2000) Public markets, private orderings and corporate governance. *Int Rev Law Econ* 20(4):453–477
- Pagano U (2012) No institution is a free lunch: a reconstruction of Ronald Coase. *Int Rev Econ* 59(2):189–200
- Pagano U, Vatterio M (2015) Costly institutions as substitutes: novelty and limits of the Coasian approach. *J Inst Econ* 11(2):265–281. Coase memorial issue
- Schneider M (2007) The nature, history and significance of the concept of positional goods. *Hist Econ Rev* 45:60–81
- Solnick SJ, Hemenway D (1998) Is more always better? A survey on positional concerns. *J Econ Behav Organ* 37(3):373–383
- Solnick SJ, Hemenway D (2005) Are positional concerns stronger in some domains than in others? *Am Econ Rev* 95(2):147–151
- Vatterio M (2009) Understanding power. A law and economics’ approach. VDM-Verlag, Saarbrücken
- Vatterio M (2010) From W. N. Hohfeld to J. R. Commons, and beyond? *Am J Econ Sociol* 69(2):840–866
- Vatterio M (2013a) Positional goods and Robert Lee Hale’s legal economics. *J Inst Econ* 9(3):351–362
- Vatterio M (2013b) Alla ricerca di regole (e istituzioni) efficienti. *Riv Crit Dirit Priv* 31(1):123–138
- Veblen T (1899) *The theory of the leisure class*. MacMillan, New York

Posner, Richard

Alain Marciano

MRE and University of Montpellier, Montpellier, France

Faculté d’Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France

Abstract

Posner is one of the main contributors to what is known as “economic analysis of law.” In this entry, we restrict our presentation to a few controversial claims he made (efficiency, wealth-maximization, Hicks-Kaldor, judicial decision making).

Biography

Richard A. Posner was born in New York city January 11, 1939; he graduated from Harvard Law School and taught at Stanford University Law School and at the University of Chicago Law School; he is judge of the US Court of Appeals for the Seventh Circuit where he was appointed in 1981 and for which he was chief judge from 1993 to 2000.

General Presentation

The first point that could be noted is that Posner is one of the most important judges of all times in the USA. Christopher McCurdy and Ryan Thompson wrote in 2011, he “can arguably be called the most influential judge currently on the bench” (p. 50). Indeed, Posner is “a staple in legal casebooks . . . his name continues to show up in public discourse and peer judge interviews” (2011, p. 3). They cite a few figures that impressively witness of Posner’s influence: between 1998 and 2000, Judge Posner was cited 1,406 times (figure computed by Choi and Mitu Gulati 2004); between 1989 and 1991, he ranked third on the list of the “Top Twenty-Five Prestige Scores (Klein and Morrisroe 1999) and finally 118 Judge Posner opinions appeared in casebooks used in the 1999–2000 school year – ten times more than 90% of federal circuit judges. Yet, the American Bar Association – on a scale going from “exceptionally well qualified,” “well qualified,” “qualified,” to “not qualified” – gave him the “lowest possible ratings, ‘qualified/not qualified’” (Lott 2006). This is also impressive and reveals that, as important as he might be, Posner remains controversial in his judicial decision making.

Second, Posner is and has always been a public intellectual (Fleury and Marciano 2013). Writing for nonacademic audiences has never been secondary for Posner. It became more and more important these last 10 years, after 2004, when he launched the Becker-Posner blog (becker-posner-blog.com/), with economist and 1992 Nobel prize winner Gary Becker. What Posner posts on these blogs might seem a bit far-fetched

sometimes. It would have been useful and interesting to look at these writings in detail. For a lack of place, we left it aside.

Third, and obviously a major aspect of Posner’s work: his academic and scientific writing. Posner has been so prolific that it is impossible to summarize his ideas – one can find his list of publications on the website of the University of Chicago (see <http://www.law.uchicago.edu/faculty/posner-r>). One can nonetheless emphasize a few important ideas he put forward.

Innovative and Original Aspects

Economic Analysis of Law

In 1973 was published the first edition of *Economic Analysis of Law*. At the time, the book was reviewed as a contribution in *law and economics*: the book could “serve very well for a law and economics course” (Diamond 1974, p. 294) and indeed is a “coursebook in law-and-economics” (Krier 1974, p. 1697). At best, it was noted that “[w]ith the publication of Richard A. Posner’s economic analysis of law, that field of learning known as ‘Law and Economics’ [had] reached a stage of extended explicitness” (Leff 1974, p. 451). What was not viewed was that Posner had launched a new field. Economic analysis of law is not simply another name for from law and economics (see Harnay and Marciano 2009). By contrast with “law and economics” that defines economics by its subject matter, an economic analysis of law assumes that economics is a method, an approach or a set of tools that is used to understand nonmarket phenomena. It emerged in the early 1970s under the influence of Gary Becker (Fleury 2015). In this perspective, the law became an object that economists could analyze. Becker (1968), William Landes (1967, 1968, 1971) or Isaac Ehrlich (1970) proposed the first economic analyses of legal problems. Posner was not only the first one to name the field, with the title of his book, he was also the first to systematically adopt such an approach (see criticisms, in particular, in Backhaus 2017 and Malecka 2017).

Efficiency, as Wealth Maximization

One of the consequences of an economic analysis of law lies in the possibility of using the concept of efficiency to assess legal systems: rules, behaviors or judicial decisions are legitimate when or because they are efficient, which, to Posner, means when or because they maximize society's wealth.

Let us start by saying that Posner defines wealth in monetary terms exclusively, as

the value in dollars or dollar equivalents (...) of everything in society. It is measured by what people are willing to pay for something or, if they already own it, what they demand in money to give it up. The only kind of preference that counts in a system of wealth maximization is thus one that is backed up by money – in other words, that is registered in a market

(1979, p. 119). Or, “the sum of all tangible and intangible goods and services, weighted by prices of two sorts: offer prices (what people are willing to pay for goods they do not already own); and asking prices (what people demand to sell what they own)” (Posner 1995, p. 356).

Such definition of wealth is not as narrow as might appear at first sight. Indeed, Posner adopts a rather extended conception of markets. First, he does not restrict to explicit markets. But he considers implicit markets, in which services are sold that can be monetized by reference to substitute services sold in explicit markets. This enables Posner to use the wealth maximization criterion to assess the pecuniary value of a wide set of goods and services. Second, he also considers hypothetical markets, defined as those markets in which high transaction costs prevent individuals to transact with each other voluntarily and, therefore, efficient transactions to occur – for instance, individuals may be forced into involuntary transactions in the case of accidents, when a victim is obliged to transact with the injurer and the victim and the injurer cannot agree on a common price. Then, using a third party (e.g., a court) within a hypothetical market to determine the price of the transaction is a way to have the transaction occur actually and to achieve an efficient allocation of resources. Indeed, agents will consent to the accident-as-a-transaction as long as it is

wealth-maximizing, i.e., when it is less costly to let the accident occur than to prevent it. On the contrary, they will let the accident occur when it is wealth-enhancing. Then, transactions within hypothetical markets reconcile the possibility of involuntary transactions with the idea of individual consent that lies at the core of the wealth maximization criterion.

Indeed, Posner equates efficiency with wealth maximization. Thus, from this perspective, “[r]esources are efficiently allocated in a system of wealth maximization when there is no reallocation that would increase the wealth of society” (Posner 1980a, p. 243). That is, Posner remains in the standard maximizing framework of utilitarianism but substitutes wealth for “utility.”

Using wealth, instead of utility, as most economists usually do, allows Posner to solve the vexed problems of how to measure and to compare individual utilities. In a system based on “wealth maximization,” interpersonal comparison of wealth is possible, making it possible to compare the gains of one individual or a group of individuals to the losses incurred by another individual or group. Therefore, a transaction will be considered as legitimate and desirable when it increases the wealth of the society. Posner gives the following example:

I offer you \$5 for a bag of oranges, you accept, and the exchange is consummated. We can be confident that the wealth of the society has been increased. Before the transaction you had a bag of oranges worth less than \$5 to you and I had \$5; after the transaction you have \$5 and I have a bag of oranges worth more than \$5 to me. We are both richer, as measured by the money value we attach to the goods in question.

(1979, p. 120). The transaction is then legitimate and even desirable.

Another example is given by Posner: “[c]onsider an accident that inflicts a cost of \$100 with a probability of .01 and that would have cost \$3 to avoid. The accident is said to be a wealth-maximizing ‘transaction’ [...] because the expected accident cost (\$1) is less than the cost of avoidance” (1995, p. 358). Thus, clearly, what this accident costs to the society is less than what it would have costed to avoid it. An

individual who does not spend the \$3 to avoid a \$1 cost is not negligent. Conversely, if the cost of avoidance is lower than the accident cost, not only the accident is not wealth-maximizing but also an individual who would have not tried to avoid the accident would have been negligent. This is precisely why Posner also praised judge Learned Hand for his 1942 decision in a liability case: it was a wealth maximizing – and therefore efficient – decision (e.g., Posner 1972, p. 32, or 1999, p. 91).

Wealth and Capacity to Pay

Posner considers that wealth must be determined by using individuals' willingness to pay for a good rather than by its price. This is understandable for two reasons: first, on implicit and hypothetical markets, there are no prices and therefore the only way to value a transaction is to know what individuals would like to pay for a good; and second, the real value of a good for an individual is not represented by the price but by his or her willingness to pay. Wealth relates to consumers' surplus.

But then, one must also note that individuals may not only value a good they desire. They should be able to pay for it. Posner meant capacity to pay as a means to determine wealth:

a desire not backed by ability to pay has no standing—such a desire is neither an offer price nor an asking price. I may desperately desire a BMW, but if I am unwilling or unable to pay its purchase price, society's wealth would not be increased by transferring the BMW from its present owner to me. (1995, p. 357)

or,

that wanting something very much, but not being able to pay more for it than its owner or competing demanders, does not establish a claim to a good in a system of wealth maximization, although it might do so in a system of utility maximization. Wealth maximization thus excludes claims based on pure desire-claims not backed up by willingness (implying ability) to pay. (1980a, p. 243)

One would say that for a criterion that Posner also wanted to be “ethical” – he “argued that “wealth maximization” provides an ethically attractive norm for social and political choices, such as those made by courts asked to determine whether

negligence or strict liability should be the rule for deciding whether an injurer must compensate his victim” (1985, p. 85) – grounding “wealth” on individual's “capacity to pay” is problematic. Posner replied mainly by saying that one should distinguish efficiency from justice and the allocation from the distribution of resources. If a rule, a legal decision or a policy increases and maximizes wealth, then it directly benefits to certain individuals and not to others. Nothing prevents the use of redistributive policy to compensate the effects of such a wealth-increasing rule, decision, or policy.

This relates to another and quite important aspect of Posner's conception of wealth maximization: the Kaldor-Hicks criterion.

Kaldor-Hicks and Wealth Maximization

In economics, efficiency is measured in terms of Pareto optimality. However, Posner insisted that “Pareto superiority is not a necessary condition to be wealth maximizing” (Posner 1995, p. 357). Posner distinguished wealth from utility maximization à la Pareto. To him, a transaction, decision, or policy may be “wealth maximizing even if the victim is *not* compensated” (Posner 1995, p. 358; emphasis added). Precisely, to know if an allocation of resources is efficient without the Pareto criterion, which implies effective compensation, one may use the Kaldor-Hicks test, which rests on potential compensation. In other words, effective compensation is not required to characterize a move from one situation to another one: “[u]nder the Kaldor-Hicks definition of efficiency, which is also widely used by economists, a reallocation of resources is efficient if it enables the gainers to compensate the losers, whether or not they actually do so. This is equivalent to wealth maximization” (Posner 1980a, p. 244). Or, to put it in the more figurative terms Posner himself uses, an efficient decision in the Kaldor-Hicks sense consists in “making the pie larger without worrying about how the relative size of the slices changes” (2000, p. 1155).

Interestingly, this latter formulation reveals one important feature of Kaldor-Hicks efficiency and, therefore, of Posner's wealth maximization criterion: the exclusion of distributional issues and “normative considerations of distributive justice”

as the Kaldor-Hicks criterion “treats a dollar as worth the same to everyone.” (2000, p. 1154). However:

to the extent that distributive justice can be shown to be the proper business of some other branch of government or policy instrument (for example, redistributive taxation and spending) and that ignoring distributive considerations in the particular domain of decision making that is under consideration will not have systematic and substantive distributive consequences, it is possible to set distributive considerations to one side and use the Kaldor-Hicks approach with a good conscience. This assumes, as I have said, that efficiency in the Kaldor-Hicks sense . . . is a social value. (Posner 2000, pp. 1154–1155)

From that perspective, Posner adds, “efficiency in the Kaldor-Hicks sense is accepted as a social value, albeit not the only social value.” (Posner 2000, p. 1154). In addition, “even if Kaldor-Hicks efficiency has no social value”, it is “a convenient instrument” (Posner 2000, p. 1156) that can be used to decide whether or not one should make a decision. This is where Posner links the “Kaldor-Hicks concept of efficiency” to cost-benefit analysis: decisions can be made, rules modified, policy measures taken by comparing costs and benefits without the winners being obliged to compensate the losers. Then, if the benefits are greater than the costs, wealth is maximized; on the contrary, if costs exceed benefits, wealth is not maximized. Thus, the legitimacy of CBA comes from the use of a Kaldor-Hicks definition of efficiency. Reciprocally, Posner’s defense of CBA is also a defense of the Kaldor-Hicks principle.

Common Law and Efficiency

A major positive claim in the law and economics literature that was first formulated by Posner is that the common law is efficient. Indeed, the development of common law is mainly driven by judicial decisions issued in response to the needs of the society through cases. When facing new cases for which no precedent is available, judges make decisions relying on the economic efficiency criterion, defined as wealth maximization or Kaldor-Hicks efficiency. Thus, they are assumed to have a preference for efficiency over other moral values (such as income redistribution)

and to share that definition of efficiency as a desirable social goal with the society. Further, even if they prefer other values, they issue decisions promoting efficiency because the limitations of the judicial process constrain them in their ability to pursue other objectives. In particular, unlike legislators, they are poorly equipped with tools allowing them to engage in effective wealth redistribution or other unattainable values (see for instance Posner 1973, 1979, 2010). As a consequence, judges having internalized efficiency as a socially desirable goal, even unconsciously, adopt an economic logic when they make their decisions. Indeed,

[T]he character of common law litigation forces a confrontation with economic issues. The typical common law case involves a dispute between two parties over which one should bear a loss. In searching for a reasonably objective and impartial standard, as the traditions of the bench require him to do, the judge can hardly fail to consider whether the loss was the product of wasteful, uneconomical resource use. In a culture of scarcity, this is an urgent, an inescapable question. And at least an approximation to the answer is in most cases reasonably accessible to intuition and common sense. (Posner 1973, p. 99)

Parties to litigation consent to efficient decisions made by courts because their self-interest is promoted by supporting such decisions. Namely, “by doing so they increase the wealth of the society; they will get a share of that increased wealth; and there is no alternative norm that would yield a larger share” (Posner 1980b, p. 505). Furthermore, independent judges are insulated from interest group pressures and personal factors (Posner 1993) and can therefore promote social efficiency.

Resting on the examination of particular common law fields, much of Posner’s works are forceful attempts to determine what legal rules would be efficient and to examine how they correspond to the legal rules that exist. Along this line, he analyzes various fields of common law, including contract law, liability rules, property law, criminal law, family law, and sex law, using the lens of economic efficiency. Then, common law appears to differ from a collection of multiple and conflicting precedents whose coherence may at

first sight be uneasy to grasp. But it may rather be best understood as shaped by an economic – albeit most of the time implicit – economic logic underlying all judicial decisions and intending to maximize social wealth. Posner’s theory of the efficiency of common law helps thus rationalize common law within a single – economic – framework. In his own words (Posner 2011, pp. 315–316),

economics is the deep structure of the common law, and the doctrines of that law are the surface structure. The doctrines, understood in economic terms, form a coherent system for inducing people to behave efficiently, not only in explicit markets but across the whole range of social interactions.

Posner’s claim that the common law is efficient has fueled an important literature, either criticizing or agreeing with him. One of the most debated issues deals with the mechanisms likely to explain the evolution of law towards efficiency. A substantial body of literature has raised doubts about the existence of such mechanisms, for various reasons. First, the judicial preference for efficiency over other moral values and preferences (for leisure, reputation) has been a matter of considerable debate, whereas the apparent disinterest of judges in efficiency, at least as expressed in common-law court opinions, has been emphasized. Thus, in response to Posner’s seminal analysis, various demand-side models have been developed, calling attention to forces other than judicial preferences in explaining the law and suggesting that selective litigation of rules by parties may actually drive legal evolution and help promote efficient legal common law rules, regardless of judicial preferences, as inefficient rules may be challenged before courts and therefore overturned more often than efficient ones (see for instance Goodman 1978; Landes and Posner 1979; Priest 1977; Rubin 1977). Also in response to Posner’s theory of the evolution of common law, other works have cast doubt on the very fact that common law is efficient. Some authors point out that common law adjudication may rather lead to cycles of efficiency and inefficiency (Cooter and Kornhauser 1980). Others argue that the excessive amount of information needed for judges to decide cases may prevent them from

making efficient decisions (Aranson 1992; Rizzo 1980). It has also been argued that common law may lead to inefficient lock-in and path-dependency when it preserves obsolescent rules. Furthermore, Tullock (1980) has claimed that the English common law adjudication is less efficient than legislation – public choice critics mainly focusing on the idea that judicial processes are subject to the same sorts of interest-group pressures as are legislatures. More recently, the legal origins literature has revived the debates on the common law efficiency through its attempts to demonstrate the superior economic performance of common-law legal systems over other systems.

Judges (Judicial Preferences and Behavior)

Posner devoted a lot of work analyzing Judges and judicial behavior. It was one of the first areas that he viewed as important when he started to use economics to analyze the law. In 1971, he thus wrote that judges seem to display a certain propensity towards efficiency – they “are guided by concern with economic efficiency” (1971, p. 223) and “think in economic terms” (1971, p. 224). Or, in 1973, he insisted that one can “*assume* that judges make their decision in accordance with the criterion of efficiency” (1973, p. 325; emphasis added). What was important to Posner was to explain the efficiency of the Common Law. To be more precise, there were two versions of this theory. First, the positive version holds that rational judges having a preference for efficiency do actually issue decisions that maximize social wealth; hence, common law tends towards efficiency. Second, the normative view argues that judicial decisions should help common law to evolve towards efficiency, as rational judges should make their decisions with the wealth maximizing criterion (Posner 1980a).

Eventually, Posner criticized the traditional and idealized view on judicial behavior according to which public interest may be a judicial goal – assuming this is actually the case would amount to acknowledging that judges do not behave as ordinary human beings. To him, there is no reason to argue that judges are different from other individuals. Hence, it can be assumed that they are self-

interested utility maximizers – they maximize “the same thing as everybody else does” (1993). Their utility function may contain both monetary and nonmonetary arguments, such as leisure, popularity, prestige, and reputation. Judicial decisions may also have some pure consumption value for judges who may therefore derive some kind of positive utility from their decisions close to that derived by “artists or craftsmen who value aesthetic excellence in their field” (Posner 2008).

One of the consequences of this view is that judges may not follow the precedent and may be led to innovate. This means that this is consistent with the idea that judges create the law, that they acting as *interstitial legislators* and significant policy-makers.

Over time, Posner himself has also continued documenting the key role played by judges from both a theoretical and empirical standpoint, regularly publishing contributions on adjudication and judicial behavior – the room continuously dedicated to that issue in the successive editions of his textbook *Economic Analysis of Law*, as well as his most recent publications (2008, or 2013 with Epstein and Landes), obviously express his continuing interest for the topic.

Legacy

It is impossible to summarize Posner’s legacy. His impact on economic analyses of law is huge. It is probably not exaggerate to claim that all contributors to law and economics/economic analysis of law are in one way or another “Posnerians”. Either because they agree or because they disagree with him. Posner has radically transformed law and economics and also probably the legal system.

Cross-References

- [Becker, Gary S.](#)
- [Coase, Ronald](#)
- [Economic Analysis of Law](#)
- [Wealth Maximization: Efficiency and Equality Considerations](#)

References

- Aranson PH (1992) The common law as central economic planning. *Constit Polit Econ* 3(3): 289–317
- Backhaus JG (2017) Lawyers’ economics versus economic analysis of law: a critique of professor Posner’s “economic” approach to law by reference to a case concerning damages for loss of earning capacity. *Eur J Law Econ* 43(3):517–534
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Choi SJ, Mitu Gulati G (2004) Choosing the next supreme court justice: an empirical ranking of judge performance. *Calif Law Rev* 78(1):23–117
- Cooter RD, Kornhauser LA (1980) Can litigation improve the law without the help of judges? *J Leg Stud*, 9:139–163
- Diamond PA (1974) Review of economic analysis of law, by Richard A. Posner. *Bell J of Econ and Management Science* 5(1):294–300
- Domnarski W (2016) Richard Posner. Oxford University Press, New York
- Ehrlich I (1970) Participation in illegitimate activities: an economic analysis. PhD dissertation, Chicago
- Epstein L, Landes WM, Posner RA (2013) The behavior of federal judges: a theoretical and empirical study of rational choice. Harvard University Press, Cambridge
- Fleury J-B (2015) Massive influence with scarce contributions: the rationalizing economist Gary S. Becker, 1930–2014. *Eur J Law Econ* 39:3–9
- Goodman JC (1978) An economic theory of the evolution of common law. *J Leg Stud* 7(2):393–406
- Harnay S, Marciano A (2009) Economics and the law: from ‘law and economics’ to an economic analysis of law. *J Hist Econ Thought* 31(2):215–232
- Fleury J-B, Marciano A (2013) Becker and Posner: freedom of speech and public Intellectualship. *Hist Polit Econ* 45(Suppl):254–278
- Klein D, Morrisroe D (1999). The prestige and influence of individual judges on the U.S. courts of appeals. *J Leg Stud* 28(2), 371–391
- Krier JE (1974) Review of economic analysis of law by Richard A. Posner. *Univ Pa Law Rev* 122(6): 1664–1705
- Landes WM (1967) The effect of state fair employment laws on the economic position of nonwhites. *Am Econ Rev* 57(2):578–590
- Landes WM (1968) The economics of fair employment law. *J Polit Econ* 76(4):507–552
- Landes WM (1971) An economic analysis of the courts. *J Law Econ* 14(1):61–107
- Landes WM, Posner RA (1979) Adjudication as a private good. *J Leg Stud* 8(2):235–284
- Leff AA (1974) Economic analysis of law: some realism about nominalism. *Virginia Law Review* 60(3): 451–82

- Lott J (2006) Pulling rank. *New York Times*, 25 Jan. http://www.nytimes.com/2006/01/25/opinion/25Lott.html?_r=2&
- Malecka M (2017) Posner versus Kelsen: the challenges for scientific analysis of law. *Eur J Law Econ* 43(3):495–516
- McCurdy CC, Thompson RP (2011) The power of Posner: a study of prestige and influence in the federal judiciary. *Idaho Law Rev* 48(1):49–71
- Medema SG (2010) Richard A. Posner. in *The Elgar Companion to the Chicago School of Economics*, Edward Elgar, Cheltenham, 306–310
- Posner RA (1971) Killing or wounding to protect a property interest. *J Law Econ* 14(1):201–232
- Posner RA (1972) A theory of negligence. *J Leg Stud* 1(1):29–96
- Posner RA (1973) *Economic analysis of law*, 1st edn. Aspen Publishers, Boston
- Posner RA (1979) Utilitarianism, economics, and legal theory. *J Leg Stud* 8(1):103–140
- Posner RA (1980a) The value of wealth: a comment on Dworkin and Kronman. *J Leg Stud* 9(2):243–252
- Posner RA (1980b) The ethical and political basis if the efficiency norm in common law adjudication. *Hofstra Law Rev* 8:487–507
- Posner RA (1985) Wealth maximization revisited. *Notre Dame J Law Ethics Public Policy* 2:85–105
- Posner RA (1993) What do judges and justices maximize? (the same thing everybody else does). *Supreme Court Econ Rev* 3:1–41
- Posner RA (1995) *The problem of jurisprudence*. Harvard University Press, Cambridge
- Posner RA (1999) The law and economics of the economic expert witness. *J Econ Perspect* 13(2):91–99
- Posner RA (2000) Cost-benefit analysis: definition, justification, and comment on conference papers. *J Leg Stud* 29(2):1153–1177
- Posner RA (2008) *How judges think*. Harvard University Press, Cambridge
- Posner RA (2010) Some realism about judges: a reply to Edwards and Livermore. *Duke Law J* 59(3):1177–1186
- Posner RA (2011) *Economic analysis of law*, 8th edn. Aspen Publishers, Boston
- Priest GL (1977) The common law process and the selection of efficient rules. *J Leg Stud* 6(1):65–82
- Rizzo MJ (1980) The mirage of efficiency. *Hofstra Law Rev* 8:641–658
- Rubin PH (1977) Why is the common law efficient? *J Leg Stud* 6(1):51–63
- Tullock G (1980) *Trials on trial, the pure theory of legal procedure*. Columbia University Press, New York

Post-grant Patent Review

► Patent Opposition

Power Indices

Dennis Leech

University of Warwick, Coventry, UK

Abstract

Many voting bodies are constituted on a principle of accountability whereby a member's influence is intended to be a reflection of a measure of size such as financial contribution or population. Examples include the joint stock company, the US Electoral College, the IMF executive and the EU council of ministers.

Power indices are a tool for addressing the (often ignored) problem inherent in this: that the constitution defines the voting rules and not the effective voting powers they imply. Indices due to Penrose-Banzhaf and Shapley-Shubik are often used. That they ignore preferences is often cited as a limitation with regard to positivistic analysis of existing constitutions but an advantage when they are used to address the normative problem of designing voting rules. Power indices can reveal hidden properties of voting rules that are not obvious at a superficial level. The issue of how to construct behavioral power indices that do take account of preferences remains an important research dimension. Power indices can also help us understand multi-tiered governance structures such as federal constitutions or corporate networks, an area where there is need for further research.

Power Indices

Many decision-making bodies are designed with voting rules that assign different voting powers to different participants. Examples include intergovernmental organizations such as the United Nations security council, the Bretton Woods institutions, the Council of Ministers of the European Union, and so on. Many are formally constituted to embody a principle of accountability whereby a member's influence through voting is intended to

be a reflection of financial contribution. An example is the joint stock company whereby shareholders' voting powers are directly proportional to the number of shares they own.

Power indices (more properly perhaps known as voting power indices because power is a much broader concept than is simply defined by formal voting rules) are a tool for addressing the important problem inherent in this: that the constitution defines the voting rules and not the voting powers they imply. For example a shareholder who owns 40% of a company's voting shares could have nearly 100% of the voting power in the sense of the ability to determine a decision taken by voting if the other 60% of shares are widely held. But on the other hand, it could be worth less than 40% of the decision-making power if there are other large bloc shareholders. This problem is rarely acknowledged by policy makers and even more rarely addressed. The best reference on voting power is Felsenthal and Machover (1998).

Power indices are quantitative measures of the influence of each voter in terms of the voter's ability to affect the outcome whenever a hypothetical vote is taken. The basis of the definition of the index is an examination of each possible "yes/no" vote that is theoretically possible. A voter has voting power whenever he (or she or it if the voter is an institution) can change the outcome of a vote from "no" to "yes" by changing the way he casts his vote. The power index for any voter is then the proportion of voting outcomes – taking into account all the possible ways all the voters can cast their votes – for which this is possible.

A real-world example of why this approach is important is provided by the history of the original European Economic Community when it was founded in 1958. The council of ministers adopted a system of qualified majority voting with weighted voting whereby the three largest countries (France, Italy, and West Germany) all had four votes each, the smaller countries (Belgium and Netherlands) two, and tiny Luxembourg one. This was intended to ensure that Luxembourg got due recognition as a sovereign state beyond what its tiny population size might warrant if representation was to be strictly proportional to population. So the intention of the framers of the

constitution of the EEC was that voting weights be less than proportional to population so that the larger countries would be underrepresented relative to their populations. However an examination of the voting rule against possible voting outcomes indicates that this assumption is far from the truth. The voting rule was that a qualified majority decision required a threshold of 12 or more weighted votes to pass. Therefore, in order for the vote of Luxembourg to be able to make a difference, that is, for Luxembourg have voting power, there would have to be 11 "yes" votes from the other countries. But the constitution guaranteed that was impossible. So Luxembourg had zero voting power although it had almost 6% of the voting weight. The subject of the design of the system of qualified majority voting for the council of ministers, particularly in the Nice and Lisbon treaties, has given rise to a substantial literature on voting power indices.

All this is straightforward, but turning the idea of measuring voting power into a practical voting power index is not. How voting outcomes should be counted – whether orderings or divisions of voters – is not obvious, and this has given rise to different indices that give different results in general. There are two so-called classical indices, often referred to as the Penrose-Banzhaf index (PBI) and the Shapley-Shubik index (SSI) (though nomenclature varies between authors). The SSI is based on the idea that voting outcomes are essentially orderings of voters; each ordering counts equally as a different possible outcome of the voting rules. Each voter has power to the extent they can be the pivot or swing voter able to change the decision from "no" to "yes." On the other hand, the PBI treats each possible division of voters into two groups voting "for" and "against" a decision. A power index is then defined for each voter as the proportion of outcomes for which that voter is the pivot or swing (Banzhaf 1964; Shapley and Shubik 1954; Penrose 1946).

These definitions lend themselves to interpretation as probabilities, each voting outcome being deemed to occur randomly with equal probability, and are often expressed in such terms. This interpretation has led to criticism by scholars who

object to the idea of random voting as a description of real-world behavior. However this probabilistic interpretation is not essential, and a simple interpretation in terms of relative frequencies of outcomes avoids the need to assume anything about behavior. Nevertheless, because direct computation of power indices from basic definitions requires consideration of all outcomes, this can be difficult when the number of voters is large; then algorithms based formally on probability theory are useful (Leech 2003).

Much of the literature fails to get to grips with the issue of which of the two classical power indices is better as a measure of voting power. There tends to be a preponderance in applied work for the PBI, probably because the idea of an outcome of a vote as a division has more intuitive appeal than an ordering, but many studies report the two indices side by side, leaving it to the reader to decide between them. The case against the SSI was first made by Coleman (1971) who argued strongly against the SSI on the basis that treating all orderings of voters is theoretically implausible and the weighting the formula requires is not justified. See Leech (2002), which also showed that results from applying the SSI to British company shareholdings gave results at variance with empirical evidence about company control.

A second major issue, besides the choice of which coalition model to use, relates to what sort of analysis is being done using the voting power indices. Is a voting power index meant to tell us about how voters actually behave in reality or about nothing more than theoretical possibilities? The failure of many writers to recognize this point has led to a lot of – often strongly expressed – discussion. In fact it is important to distinguish between two types of analysis for which power indices are useful: normative and positive.

Normative analysis is concerned with the design of voting rules. It is essential that this is done “a priori” without regard to the preferences or likely voting behavior of the voters. The voting rule is seen as being designed behind a Rawlsian veil of ignorance that attributes equal importance to all voting outcomes that can theoretically occur. It would clearly be morally wrong, for instance,

when considering the enlargement of a voting entity such as the European Union, to allocate votes to a new member on the basis of its likely voting behavior: to give it different weighted votes as to whether it was likely to support or oppose a particular grouping. Nothing other than an objective measure such as population would be reasonable as a basis for choosing its voting weight. Therefore for this normative analysis, it would be wrong to take account of preferences (Braham and Holler 2005).

This point has been ignored by positivist writers who have tended to be more interested in the behavior of voters with respect to a *given* voting rule than the *choice* of voting rule. From a positivist point of view, a priori or “constitutional” power indices, the PBI and SSI, have been criticized for not taking account of preferences which would assign different probabilities to different outcomes. In the EU council, for example, a positivist description of voting behavior would give a higher probability to outcomes where France and Germany vote on the same side than where the UK and France or Germany are in agreement. The PBI and SSI are clearly limited measures of voting power from this point of view.

Voter	Weight	BPI	Weight	BPI	Weight	BPI
1	9	0.0	9	5.0	9	10.0
2	17	20.0	16	15.0	9	10.0
3	17	20.0	17	18.3	17	16.7
4	18	20.0	17	18.3	18	16.7
5	19	20.0	17	18.3	19	16.7
6	20	20.0	24	25.0	27	30.0

Percentages. Threshold 51. Normalized BPI. Computations using algorithm at [Leech and Leech](#)

There is a need for much more work on developing “a posteriori” or “behavioral” power indices which do take account of preferences and are able to give an empirical account of the distribution of voting power.

One aspect of the normative use of power indices is in enabling the discovery of hidden properties of constitutional designs. In the EEC example above, it was fairly easy to spot that there was a voter with positive weight but zero power. But more complicated voting rules require voting

power analysis. The table below shows some possibilities that can occur with six voters. Fairly innocuous-looking changes to the voting weights assigned to other voters can make a profound difference to the share of voting power enjoyed by the voter number 1 with 9% of the votes, from 0% to 10%. This illustrates a particularly useful application of voting power analysis in being able to find that a voter with positive weight has in fact zero power. Such a voter is known in the literature as dummy. This feature of voting power analysis is especially useful because it is independent both of the index used and also whether or not preferences are taken into account.

The main open research question is the choice of which of the two so-called “classical” power indices, BPI or SSI, is better. So far there is no consensus on the best index to measure voting power. Some scholars prefer the mathematical rigor of the SSI, while others argue that the assumptions underlying the BPI contain more realism. There is also the challenge of developing behavioral power indices.

Much discussion recently has centered on the properties of voting power indices in *large* voting bodies, that is, when the number of voters is large in some sense. It has been argued that there is a general asymptotic tendency, as the number of voters increases without limit, for relative voting powers to become proportional to relative voting weight. Lindner and Machover (2004) have come up with sufficient conditions for this, which they called “Penrose’s limit theorem.” Leech (2013) however has shown that this is a special case, and there is no tendency toward proportionality of voting power indices and voting weight in general.

An important extension of the voting power indices method as an analytical tool is to multi-level voting rules such as voting power in federal systems (e.g., the US electoral college, see Miller (2013)), company groups with complex ownership structures Crama and Leruth (2007), and many other applications.

Voting power remains an active research field. Two recently published edited volumes containing applications are Holler and Nurmi (2013) and Fara et al. (2014).

References

- Banzhaf JF III (1964) Weighted voting doesn’t work: a mathematical analysis. *Rutgers Law Rev* 19:317
- Braham M, Holler MJ (2005) The impossibility of a preference-based power index. *J Theor Polit* 17(1):137–157
- Coleman JS (1971) Control of collectivities and the power of a collectivity to act. *Soc Choice*:269–300
- Crama Y, Leruth L (2007) Control and voting power in corporate networks: concepts and computational aspects. *Eur J Oper Res* 178(3):879–893
- Fara R, Leech D, Salles M (2014) *Voting power and procedures*, Springer
- Felsenthal D, Machover M (1998) *The measurement of voting power*. Edward Elgar, Cheltenham
- Holler MJ, Nurmi H (2013) *Power, voting, and voting power: 30 years after*. Springer, Berlin
- Leech D (2002) An empirical comparison of the performance of classical power indices. *Political Stud* 50(1):1–22
- Leech D (2003) Computing power indices for large voting games. *Manag Sci* 49:831–838
- Leech D (2013) Power indices in large voting bodies. *Public Choice* 155(1–2):61–79
- Leech D, Leech R Algorithms for voting power indices: www.ecaa.ac.uk
- Lindner I, Machover M (2004) LS Penrose’s limit theorem: proof of some special cases. *Math Soc Sci* 47(1):37–49
- Miller NR (2013) A priori voting power and the US Electoral College. In: Holler MJ, Nurmi H (eds) *Power, voting, and voting power: 30 years after*. Springer, Berlin, pp 411–442
- Penrose LS (1946) The elementary statistics of majority voting. *J R Stat Soc* 109:53–57
- Shapley LS, Shubik M (1954) A method for evaluating the distribution of power in a committee system. *Am Polit Sci Rev* 48(9):787–792

Preferential Tariffs

Laura Huici Sancho

Department of International Law and Economics,
University of Barcelona, Barcelona, Spain

Abstract

Tariff preferences refer to measures that involve tariff reductions which benefit certain goods originating in certain countries. Historically tariff preferences were agreed on a reciprocal basis to recognize special relations between states. The multilateralization of

trade implied the generalization of preferences, which become an instrument for development cooperation. Tariff preferences might be accorded, unilaterally or through regional agreement schemes, to developing countries. However the progressive reduction of tariffs in international trade limits the role of tariff preferences as an instrument to agree a special treatment to certain kind of countries. Besides the increasing diversity of states requires of flexible instruments that allow special and differential treatment to better adapt to development, financial and trade needs.

This unequal treatment, which is characteristic of preferences, has turned them into a major instrument of development cooperation, and they are thus used for building relations between developed and developing countries. Nevertheless, the increasing use of the MFN in trade agreements, the multilateralization of such with the signing of GATT, the progress made in liberalizing trade as a result of successive rounds of trade negotiations, the creation of the WTO, and the expansion of its Member States are factors that have led to a declining role of tariff *preferences* as tariff reductions have gradually become more widespread.

Synonyms

Commercial benefits; Commercial concessions; Tariff benefits; Tariff concessions

Definition

Preferences are concessions which involve special treatment in trade relations between two or more countries, economic integration organizations, or regional groupings. Generally speaking, a benefit stemming from a treaty that is not extended to all states, whether they are party to it or not, may be regarded as a *preference*. However, in international economic law, the term *preferences* is used more specifically to refer to measures that involve tariff reductions which benefit only goods originating in certain countries. With this meaning, preferences aim to liberalize trade relations, since they involve reductions or even the elimination of tariffs. Yet they upset the principle of no discrimination because, by definition, they do not apply equally to all states. So being contrary to the Most Favoured Nation Clause (MFN), in the GATT/WTO system, preferences can find justification in article XXIV, the Enabling Clause, or requesting a specific waiver (GATT article XXV:5 – WTO article IX:3).

Preferences may be reserved only for certain types of products originating in certain countries; they may be limited qualitatively to certain goods and/or be subjected to quantitative restrictions.

An Historical Overview of Preferences: An Evolving Institution

The notion of preferences may be defined through their historical evolution with emphasis placed on their main features and most controversial points. With this objective in mind, four stages may be distinguished: first, preferences are incorporated into trade agreements on a reciprocal basis and respond primarily to factors concerning particular links between states; second, as part of the multilateralization of trade relations, these tend to favor the generalization of preferences limited to relations between developed and developing countries, thus eliminating reciprocity; third, preferences tend to place more conditions on compliance with certain requirements, particularly environmental and labor protection; and fourth, a crisis of the unilateral tariff preferences model has arisen, raising doubts concerning their interest and continuity in the near future, while regional preferential trade agreements are growing and new models of preferential treatment have appeared.

Reciprocal and Special Preferences

Traditionally, the establishment of a *preferential* relationship between parties was the subject of trade agreements in which states made mutual concessions. Often, the establishment of these preferences responded to the existence of cultural, historical, or geographical ties among the affected countries which sought to encourage and enhance their trade relations with a reciprocal recognition

of preferences. It should be noted that preferential and reciprocal tariff treatment among members is the basis of the free trade areas and customs unions which constitute, in this sense, specific examples of special preferential arrangements.

Granting preferential tariffs became common in relations between the metropolis and its colonies once these had become independent states. One example is the *imperial preferences* that were first established by the United Kingdom among its colonies and later continued with the Commonwealth countries. Despite initial opposition, “at the Imperial Economic Conference of 1932 in Ottawa the United Kingdom finally accepted the extension of imperial preference throughout the Empire. Eleven bilateral agreements were signed and the whole system of Empire preferences was augmented and co-ordinated” (Clute and Wilson 1958, p. 464). As regards the terms of these agreements, “the imperial preference system is a maze of fiscal arrangements, by which each participant admits into its markets certain exports of the others at lower tariff rates than are imposed upon the same products when originating elsewhere” (Feis 1945–1946, p. 663). This was a unique exclusive system of special preferences for Commonwealth members. If a member of the Commonwealth ever offered the same preferential treatment to a nonmember state, it would have to obtain a *waiver* that would allow it to do so. This unequal treatment raised tensions in relations with nonparticipating countries, in particular the USA, which led the multilateralization process of trade relations on the basis of MFN. At the signing of the GATT in 1947, the compatibility of imperial preferences with the MFN clause was recognized but limiting their future expansion (Clute and Wilson 1958, p. 467).

Another example of preferences granted to former colonies are Association Agreements signed by the European Economic Community and the African, Caribbean and Pacific Countries (ACP). The need to provide a special relationship with states that had formerly been colonies of a Member State is explicitly recognized in the Treaty of Rome signed in 1957, which established the European Economic Community (EEC). Based on this, the EEC granted preferences to several

countries that were mainly former French and Belgian colonies. In the Yaoundé Convention, held in 1963, the EEC implemented a progressive reduction of tariffs on products originating in the 18 African and Malagasy States which participated in this Convention (Article 2 of the Convention), and tariffs were completely eliminated for certain tropical products. In return, these countries undertook to eliminate any differential treatment between the six Member States of the EEC and to reduce tariffs on their products, although they did allow some flexibility at this point to protect their domestic industry (Article 3 of the Convention). In 1969, a second Yaoundé Convention was signed extending these reciprocal preferences and three more African countries were included.

The situation evolved with the signing of the Lomé Convention in 1975. The UK's entry into the EEC led to this Convention being expanded to the Association of Caribbean and Pacific States, reaching a total of 46 members. The United Kingdom was forced to leave its Commonwealth model behind and adapt to the EEC Association model. However, countries such as Canada, Australia, New Zealand, and India were not included in the statute of association and were forced to seek other means of cooperation such as trade agreements adopted on the basis of Article 113 of the Treaty of Rome. As regards the system of preferences, the Lomé I Convention signed in 1975 preserved the tariff preferences of the Yaoundé Conventions but reciprocity was reduced. Products from the ACP States benefited from a reduction or elimination of tariffs in the EEC (articles 2 and 3 of the Convention), while European products only received the MFN treatment among the Associated States (article 7 of the Convention). These general features of the system of preferences were continued in the successive Lomé II Convention signed in 1979 and Lomé III Convention signed in 1984. Even after the EEC's Generalized System of Preferences (GSP) was implemented, the Associated States benefited from better preferences than other developing countries (Grilli 1994, p. 148).

Along with the agreements with the ACP countries, the EEC also signed trade and cooperation

agreements with other states. This group included, among others, agreements with the Maghreb, Masreq, and other countries from around the Mediterranean together with several Asian and Latin American countries, which also recognized a system of preferences for them, not as high as the ACP, but still prioritized them (Grilli 1994, p. 151). In these agreements, preferences were not subject to providing direct reciprocity but were significantly limited to protect the EEC's *sensitive* products or sectors.

The nonreciprocity of preferences as well as the abolition of special regimes or reverse preferences and their replacement by a GSP was repeatedly called for by several states at the Second United Nations Conference on Trade and Development (UNCTAD) held in New Delhi in 1968 (UNCTAD doc. TD/97, Vol. I, 1968). The USA demanded that states receiving these preferences should give them up if they wished to receive preferential treatment from them. Opposing this, the EEC, the United Kingdom, and the states benefitting from these special preferences insisted that the generalization of preferences should not harm them. However, all participants agreed to eliminate the need for reciprocity and turn preferences into an instrument for development cooperation which would benefit all developing countries and serve to increase exports and aid access to international trade relations.

The Generalization of Preferences

The signing of the GATT Agreement in 1947 established the application of the MFN clause in trade relations, thus prioritizing the principle of nondiscrimination between the participating states. Although the GATT was not unaware of the existence of special regimes which, as pointed out in the previous section, prioritized relations between some participating states, these were found to be compatible with the system. Along with this, following World War II, the idea spread that aid should be given by developed countries to developing countries. The United Nations Charter linked the need for cooperation among states to reduce economic and social differences in order to ensure international peace and security (Chapter IX of the UN Charter).

It was the developing countries that would eventually introduce into the debate the need to regulate international trade to address their specific needs. As early as 1961, India called for the establishment of general tariff preferences in the GATT. In a joint declaration, titled Appendix to Resolution 1897 (XVIII) of the United Nations General Assembly, 11 November 1963, several developing countries stated that “international trade could become a more powerful instrument and vehicle of economic development not only through the expansion of the traditional exports of the developing countries, but also through the development of markets for their new products.” This was the aim of the First UNCTAD Conference held in 1964, which supported the idea of establishing a GSP to favor developing countries in order to increase and diversify their commercial relations (Final act, doc. E/CONF/46/139, 1964).

In the Charter of Algiers, approved at the Ministerial Meeting of the Group of 77 on 24 October 1967, urgent immediate measures were called for to encourage trade with developing countries and a series of principles was also passed for a general system of tariff preferences that “should provide for unrestricted and duty-free access to the markets of all the developed countries for all manufactures and semi-manufactures from all developing countries.” This model was debated in the Second UNCTAD Conference which ended with the unanimous adoption of Resolution 21 (II), *on preferential or free entry of exports of manufactures and semimanufactures of developing countries to developed countries*, whereby an acceptable mutual and Generalized System of Preferences was agreed and implemented, with no reciprocity or discrimination which would be advantageous for developing countries. In sum, UNCTAD adopted the idea of a “non-discriminatory” system of discrimination in favor of all less developed countries” (Metzger 1967, p. 770). However, the developed countries’ position on this issue varied widely.

Not all developed countries were willing to grant a GSP with the same conditions and features. As pointed out in the document addressed to UNCTAD by the OECD in 1969, “each donor country nonetheless reserves the right of taking

action appropriate to its own possibilities and taking account of certain features of its particular circumstances” (doc. OECD C(69)142). Thus developed countries believed that a GSP must be voluntary. Furthermore, donor countries failed to agree on the possibility of establishing quantitative restrictions, identifying beneficiary states, or limiting the types of products that may benefit from preferences. So, far from achieving the establishment of a single GSP for all developing countries, a proliferation of GSP arose.

The GSP approved by Australia in 1966 provided “non-reciprocal preferences on selected manufactures and semi-manufactures from developing countries up to the levels of specified quotas for each commodity.” The scheme was aimed at less developed countries which could seek to benefit from certain preferential annual import quotas for certain products, unless they were already competitive by themselves (GATT Doc L/2627, April 4, 1966 on Australia). This first scheme which was adopted for 1 year has since continued to be extended and modified, successively amending fees and/or scope of preferences.

In 1971, the EEC approved its first GSP providing tariff preferences for developing countries. The preferences were limited to a series of regulations that established criteria to identify which products were regarded as originating in developing countries and therefore most likely to benefit from the preferences. It was a complex and seemingly rigid system which aimed to recognize substantial advantages for transformed agricultural products and semimanufactured or manufactured goods from developing countries while preserving the interests of the industries of the Member States and Associated States (Hoffmann, 1971 p. 650). The preferences were highly limited in the case of agricultural products, subject to quantitative restrictions, and mainly aimed at the Group of 77 Member States and Hong Kong. The EEC would also gradually modify this first scheme and has renewed its GSP ever since.

Throughout 1971 and early 1972, other developed countries implemented their respective GSP. However, just like the previous two, all included safeguard mechanisms and largely excluded products that might be of interest for

developing countries, which significantly limited, at least in the short term, the effects of GSP on exports (Hoffmann 1971, p. 660). However, differential treatment included in the GSP contradicted the obligations stemming from the MFN under Article I of the GATT, and so it was necessary to obtain a waiver to implement it and provide continuity. From this perspective, the need soon arose for a provision that would generally uphold this special and differential treatment received by developing countries.

The continuity of differential treatment and tariff preferences for developing countries was considered from the outset of the Tokyo Round negotiations. The Enabling Clause was part of the resolutions adopted at the end of the negotiations in 1979 and aimed to accommodate differential and more favorable treatment for developing countries. This clause was based on the compatibility of “preferential tariff treatment accorded by developed contracting parties to products originating in developing countries in accordance with the Generalized System of Preferences” (Article 2 (a), GATT doc. L/4903). The problem was how to interpret the terms of the Enabling Clause in relation to the MFN clause and particularly, whether it allowed the interpretation made by developed countries as regards distinguishing developing countries.

There is no absolute answer to this question. For one, neither the GATT nor the WTO provides a definition or list of developing countries, while this group is becoming increasingly more diverse. Above and beyond the group of least developing countries (LDC) that are identified by the UN, or the orientations that may arise from the World Bank’s classifications or the human development index, it is states that grant preferences that unilaterally decide which they regard as developing countries. Additionally, the Enabling Clause states that “any differential and more favorable treatment provided under this clause shall (...) be designed and, if necessary, modified, to respond positively to the development, financial and trade needs of developing countries” (article 3 (c) GATT doc. L4903). These terms may be interpreted as providing different treatment according to different needs which may be

consistent with the system. The different interpretations regarding this were highlighted by the WTO's Dispute Settlement Body (DSB) in the case lodged by India against the EU due to its GSP in 2002. As will be seen in the following section, India questioned whether the special arrangement to combat drug production and trafficking should not be open to any developing country which might benefit from the European Union's GSP. The Panel's report concluded that "the term non-discriminatory (...) requires that identical tariff preferences under the GSP be provided to all Developing Countries without differentiation" (Doc. WT/DS246/R, above no. 6). Later, the Appellate Body (AB) overruled this opinion stating that "the term 'developing countries' in paragraph 2(a) should not be read to mean 'all' Developing Countries" (WTO doc. WT/DS246/AB/R, above no 25). Differential treatment might be acceptable as it would be defined according to "financial and trade needs of developing countries."

In short, although GSP have indeed led to a generalization of preferences for the beneficiary states, they have not done so in an absolute sense, but rather, this has meant that potentially any state regarded as a developing country by countries granting the preferences can benefit from the same if it proves it meets the requirements of such a scheme. The Decision on Implementation-Related Issues and Concerns adopted in 2001 as part of the Doha Round states that "preferences should be non-discriminatory" but are not specifically required (WTO doc. WT/MIN (01)/17, 41 LL.M. 757 (2002)). This relative notion of a generalized treatment is exacerbated when features of conditions unrelated to trade are added to the granting of preferences.

The Conditionality of Preferences: A New Way to Reverse Preferences?

The Enabling Clause specifies that "the developed countries do not expect reciprocity for commitments made by them in trade negotiations to reduce or remove tariffs and other barriers to the trade of developing countries, i.e., developed countries do not expect the developing countries, in the course of trade negotiations, to make

contributions which are inconsistent with their individual development, financial and trade needs" (article 5, GATT doc. L/4903). This section consolidates the nonreciprocity of preferences in relations between developed and developing countries. However, once again the reference made to the "development, financial and trade needs of developing countries" opens up the gate used by developed countries to condition the preferences granted to developing countries.

In its preamble, the Lomé IV Convention signed between the European Community (EC) and the ACP countries in 1990 emphasized the close relationship between promoting and respecting human rights and environmental protection for developing states. The aim of the association is to promote sustainable development and does not depend solely or primarily on an increase in trade relations nor, particularly, exports of these countries. With the review of the Convention in 1995, aid to the Associated States was linked to fulfilling the objectives of respecting human rights and the principles of good governance since noncompliance may justify withdrawal of aid or preferential treatment. The Cotonou Agreement, after 2000, has maintained this link of association with the aim of sustainable development and good governance. The erosion of preferences as a result of the reduction of general tariffs prioritizes measures of positive cooperation by consolidating the requirements of preferential trade treatment on these other obligations which are regarded as intrinsic to the effective development of countries.

In the same vein, in the early 1990s, the EC would also establish additional special tariff preferences under its GSP for products originating in the Andean and Central American states with the aim of aiding the fight against drug production and trafficking. Combating the production and trafficking of drugs was seen to be essential for the development of these countries, and in this sense, the special scheme seemed justified. Special incentive arrangements were introduced to protect labor rights and the environment a few years later, in 1994 and 1996. Again the EC granted additional preferences to developing

countries for adopting and implementing, on a domestic level, international environmental standards for agriculture, the substance of the standards laid down by the ITTO relating to the sustainable management of forests and the International Labor Organization Conventions numbers 87 and 98 concerning the application of the principles of the right to organize and to bargain collectively, and number 138 concerning the minimum age for admission to employment. In these cases, protecting the environment and recognizing labor rights were presented as “special development, financial and trade needs of developing countries.”

In 2001, the Council Regulation 2501/2001, which approved the GSP scheme for the period 2002–2004, finally grouped into one single legislative instrument the provisions relative to the general preference scheme and four special regimes for least developed countries – the everything but arms initiative – to combat drug production and trafficking; for the protection of labor rights; and for the protection of the environment. The first two were aimed only at certain developing countries. This included the LDC, identified as such by the United Nations, whose particular situation is explicitly recognized in the Enabling Clause and the waiver agreed in 1999, renewed in 2009 (doc. WT/L/304, 1990 & WT/L/759, 2009). The regime to combat drug production and trafficking was only aimed at states which the EU included in their list of beneficiaries without giving objective criteria. However, the two other special arrangements are open to any developing country that already participates in the GSP and is shown to comply with the additional conditions imposed on them by such. According to the EU, all these special regimes respond to special development, financial and trade needs of developing countries, and are, therefore, based on the Enabling Clause.

Given this interpretation, India argued in its application to the WTO’s DSB that the special arrangement to combat drug production and trafficking was discriminatory because the list of beneficiary states was unilaterally set by the EU without providing any criteria which, objectively, has led to some developing countries and not

others to be included as beneficiaries of the EU’s GSP. As mentioned above, the panel and the AB disagreed on this point yet, although the AB found that differential treatment for developing countries is, in principle, compatible with the Enabling Clause (since it must adapt to the specific needs of each country), it concluded that the EU’s special drugs regime is contrary to it, as the criteria that led to defining the list of beneficiary states (WTO doc. WT/DS246/AB/R, above 25) were not justified. However, in this case no questions were raised over the conditionality of the preferences contained in special schemes, but how countries benefitting from the preferences in the special scheme were chosen, which meant these preferences would not be available to all developing countries that might have the same needs. In short, it is not discriminatory to provide differential treatment to a different situation, but it is if the situation is the same or very similar.

Following this decision, the EU revised its GSP and Regulation 980/2005 established the GSP plus, which included three regimes of preferences: a general arrangement, a special incentive arrangement for sustainable development and good governance, and a special arrangement for LDC (article 1 of the Regulation). Apart from the LDC scheme, the other special scheme addresses countries ratifying or undertaking to ratify the agreements specified in Annex III of the Regulation. This lists what the EU considers the major universal international treaties for the protection of human rights, labor rights, environmental protection, and those protecting principles of good governance. Countries that can prove they meet the conditions may request to benefit from additional preferences under the special regime and the European Commission must ensure they comply with the requirements. Regulations 732/2008 and 978/2012 maintain three distinct regimes, without ever questioning their compatibility with WTO rules. Against this, some authors argue that the choice of a closed list of international treaties, such as the sections in Annex III of the EU Regulation, is not flexible and cannot adapt to the needs and reality of each developing country (Bartels 2007, p. 884; Turksen 2009, p.965; Wardhaugh 2013, p 839).

USA's GSP shares the carrot-and-stick approach used by the EU and also establishes differences between developing countries submitted to complying with certain conditions or criteria. Instead of setting out special regimes as part of the GSP, the USA has established specific regional preference programs and has obtained waivers for the Caribbean Basin Economic Recovery Act, the Andean Trade Preference Act, and the African Growth and Opportunity Act. All these programs establish more benefits than the US GSP, but they are restricted to some specific countries and submitted to US foreign policy interests (Kennedy 2012). Besides the USA maintain their GSP. To identify the beneficiary developing countries in the GSP, the conditions are, for instance, that they may not be communist countries; they must respect labor rights and must wipe out child labor, cooperate in the fight against terrorism, respect intellectual property rights, and apply international arbitration decisions. These conditions are established and applied unilaterally and discretionally by the USA. For example, more than 20 years ago, the USA decided to reduce up to 50% the benefits accorded to Argentina through the GSP because they had estimated that national legislation had not granted enough protection to intellectual property rights (Lavopa and Dalle 2012).

Conditionality may seem like a form of reciprocity as it requires a certain action to be taken by developing countries to benefit from a preference (Kishore 2011, p. 895). Subjecting preferences to respecting certain human rights, protecting the environment or the democratic principles of government introduces issues as factors of development that are not directly related to trade. In the Doha Round, developing countries were able to keep these issues out of the negotiations. The unilateral basis which defined the GSP and other preferential trade relations has allowed developed countries to reintroduce them into the debate by conditioning the preferences. Indeed, it is hard to deny that these issues are key to the development of a country, but the way that these are presented seem to respond more to the needs of developed countries than developing countries.

Erosion of Preferences, but Not Their Disappearance

The impact of preferences on developing countries has differed depending on whether they have been channeled through special regional agreements, such as the Association Agreements between the EU and the ACP States or the US Caribbean Basin Initiative, or by establishing a GSP. Developing countries participating in regional preferential agreements enjoy a special and privileged status compared to the rest. The concessions under these agreements are larger, tariffs are reduced or even eliminated, and the product range is wider. On this point the GSP face three problems. First, they are also of a unilaterally voluntary nature which makes them respond more to considerations imposed by the donor states than the most urgent needs of the beneficiary countries. Second, they are open to a larger group of states and are therefore more diverse, and not every country can always benefit from the same levels of preferences. Third, implementing GSP preferences involves a difficult process which leads to an increase in costs, sometimes making it unattractive to the potential beneficiary. It is up to the developing countries to apply for these preferences and demonstrate that their products comply with the regulations of origin set out in GSP, which may involve additional costs and reduce the final benefit. From this perspective, many authors criticize the results achieved by the GSP.

The general reduction of tariffs in international trade has also progressively limited the role of tariff preferences as an instrument for development cooperation. The issue of preference erosion has been widely debated in by authors for years. In regional preferential agreements, this loss has been offset by other means of cooperation, such as financial or technical assistance. But the GSP are the result of a different logic and have also been stretched by the existence and proliferation of other preferential trade agreements which provide greater preferences than those general schemes and due to the loss of margin for preferential treatment as a result of successive negotiation rounds.

Furthermore, the temporary nature that characterizes the special and differential treatment for

developing countries in the WTO creates some uncertainty. It seeks to promote development so that once the goal is achieved *exceptions* to the principle of nondiscrimination and the MFN governing multilateral trade relations will become meaningless. This feeling of uncertainty is greater in the case of a strictly unilateral GSP. A GSP depends on the willingness of the states to establish such, and they may eventually decide not to continue it. However, it is worth noting that the main GSP still persist today. The EU, the USA, and Russia have renewed their GSP in 2013; Australia did so in 2012; and Japan in 2010. But, for example, the EU has severely reduced the list of beneficiary states excluding from 2015 those states that have been classified by the World Bank as upper-middle-income countries since 2011 (European Commission delegated Regulation 1421/2013). Ways remain to be explored that can empower the role of tariff preferences as tools for development. Agreements that facilitate the management of trade in goods subject to tariff preferences would represent a qualitative edge in response to an open, general, debate in the WTO (Bartels and Häberli 2010; Persson 2012).

Before concluding it should be recalled that a more general definition of the term “preference” means any treatment that is more beneficial for a state. Other agreements also provide preferences with this broader meaning that have been of great interest to the beneficiaries. Of special interest is the decision referring to preferential treatment to services and service suppliers of LDC, 2011, which allows members to grant “preferential treatment to services and service suppliers of Least Developed Countries without according the same treatment to like services and service suppliers of all other Members” (doc. WT/L/847). The terms of this document are quite similar to those of the Enabling Clause, because preferential treatment is justified in order to increase the participation of LDCs in trade in services, which are expected to become generalized preferences for all LDC, but of a unilateral nature from the perspective of the donor state, and set temporarily.

Nevertheless, there is an increasing diversity of states in different levels of social and economic development. Today it is difficult to maintain the division between developed and developing countries as clearly separate and stable groups in different areas of international cooperation (Pauwelyn 2013). This is reflected by developing countries acting less in common, which may weaken their ability to negotiate, but also that of developed countries, some of which face serious domestic problems with a much lower growth rate than in other so-called developing countries. A system based on a generalization of preferences does not adapt well to this reality than more than ever requires ample flexibility to adapt to the “development, financial and trade needs” of the beneficiary states.

References

Monographs

Grilli ER (1994) The European community and the developing countries. Cambridge University Press, Cambridge

Journal Articles

- Bartels L (2007) The WTO legality of the EU's GSP+ arrangement. *J Int Econ Law* 10(4):869–886
- Bartels L, Häberli C (2010) Binding tariff preferences for developing countries under Article II GATT. *J Int Econ Law* 13(4):969–995
- Clute RE, Wilson RR (1958) The commonwealth and the favored-nation usage. *Am J Int Law* 52: 455–468
- Feis H (1945–1946) The future of British imperial preferences. *Foreign Aff* 661: 661–674
- Kennedy KC (2012) The generalized system of preferences after four decades: conditionality and the shrinking margin or preference. *Mich State Int Law Rev* 20:521–668
- Kishore P (2011) Conditionalities in the generalized system of preferences as instruments of global economic governance. *Int Lawyer* 45:895–902
- Lavopa F, Dalle D (2012) “¿Hay vida después del SGP? Implicancias de la posible exclusión de Argentina de los sistemas generalizados de preferencias de Estados Unidos y la Unión Europea”, Cátedra OMC FLACSO Argentina, August 2012. Available at: http://catedraomc.flacso.org.ar/wp-content/uploads/2012/08/Anexo-I_SGP.pdf
- Persson M (2012) From trade preferences to trade facilitation: taking stock of the issues. *Econ Open Access-Open Assess E-J* 6:1–33

- Turksen U (2009) The WTO law and the EC's GSP+ arrangement. *J World Trade* 43(5):927–968
- Wardhaugh B (2013) GSP+ and human rights: is the EU's approach the right one? *J Int Econ Law* 16:827–846

Further Reading

Monographs

- Hoekman B, Özden Ç (2006) Trade preferences and differential treatment of developing countries. Edward Elgar Publishing Limited, Northampton

Journal Articles

- Bartels L (2003) The WTO enabling clause and positive conditionality in the European community's GSP program. *J Int Econ Law* 6(2):507–532
- Breda Dos Santos N, Farias R, Cunha R (2005) Generalized system of preferences in General Agreement on Tariffs and Trade/World Trade Organization: history and current issues. *J World Trade* 39(4):637–670
- Charnovitz S, Bartels L, Howse R, Bradley J, Pauwelyn J, Regan D (2004) The Appellate Body's GSP decision. *World Trade Rev* 3(2):239–265
- Fernandez-Pons X, Lavopa F (2014) ¿Ojo por ojo, diente por diente? Trazando los límites de las contramedidas comerciales en el marco del Derecho de la OMC. In: Linares D, Luciano M. (Coord.): *Controversias en la Organización Mundial del Comercio. Protagonismo y estrategias de los países en desarrollo*. Editorial FLACSO Argentina – Teseo, Buenos Aires, pp 233–299
- Grossman GM, Sykes AO (2005) A preference for development: the law and economics of GSP. *World Trade Rev* 4(1):41–67
- Hofmann G (1971) Les préférences tarifaires généralisées. *Cah de Dr Eur* 6:641–661
- Howse R (2003) India's WTO challenge to drug enforcement conditions in the European community generalized system of preferences: a little known case with major repercussions for 'Political' conditionality in US trade policy. *Chic J Int Law* 4:385
- Huici-Sancho L (2006) What kind of 'Generalized' systems of preferences? *Eur J Law Econ* 21:267–283
- Inama S (2011) The reform of the EC GSP rules of origin: per aspera ad astra? *J World Trade* 45(3):577–603
- Inama S (2003) Trade preferences and the WTO negotiations on market access – battling for compensation of erosion of GSP, ACP and other trade preferences or assessing and improving their utilization and value by addressing rules of origin and graduation? *J World Trade* 37(5):959–976
- Irish M (2007) GSP tariffs and conditionality: a comment on EC–Preferences. *J World Trade* 41(4):683–698
- Khorana S, Yeung MT, Kerr WA, Perdakis N (2012) The battle over the EU's proposed humanitarian trade preferences for Pakistan: a case study in multifaceted protectionism. *J World Trade* 46(1):33–59
- López-Jurado C (2011) La oferta comercial preferencial de la Unión europea a los países en vías de desarrollo:

modalidades e interacciones. *Revista de Derecho Comunitario Europeo* 39:443–483

- Manero-Salvador A (2007) El SGP+ como elemento de la cooperación entre Europa y América Latina. *Revista Electrónica Iberoamericana* 1(1). Available at: http://www.urjc.es/ceib/investigacion/publicaciones/REIB_01_A_Manero_Salvador.pdf
- Metzger SD (1967) UNCTAD, developments in the law and institutions of international economic relations. *Am J Int Law* 61:756–775
- Ochieng C, Milton O (2007) The EU–ACP economic partnership agreements and the 'Development Question': constraints and opportunities posed by article XXIV and special and differential treatment provisions of the WTO. *J Int Econ Law* 10(2):363–395
- Pauwelyn J (2013) The end of differential treatment for developing countries? Lessons from the trade and climate change regimes. *Rev Eur Int Environ Law* 22:29–41
- Shaffer G, Apea Y (2005) Institutional choice in the generalized system of preferences case: who decides the conditions for trade preferences? The law and politics of rights. *J World Trade* 39(6):977–1008
- Yusuf AA (1980) Differential and more favourable treatment: the GATT enabling clause. *J World Trade Law* 14(6):488–507

Prisoner's Dilemma

Wolfgang Buchholz and Michael Eichenseer
University of Regensburg, Regensburg, Germany

Abstract

The Prisoner's Dilemma (PD) is probably the most famous two-person game in which a fundamental divergence between individual and collective rationalities arises: If the agents play noncooperatively, an equilibrium is achieved which, however, does not constitute the best available solution. Such a PD situation characterizes many situations of voluntary cooperation, e.g., the provision of the global public good climate protection. But in reality agents are – despite the predictions of the PD game – often willing to cooperate voluntarily to some degree which has been confirmed by experimental economics. Furthermore there are a lot of institutional devices which help overcome the cooperation dilemma in a PD situation.

The Standard Prisoner Story

The Prisoner's Dilemma (PD) is probably one of the most famous games analyzed in game theory. Its name refers to a rather artificial story conceived by the American mathematician Albert Tucker in order to convey the structure of the PD game to a broader public: Two prisoners – A(ndrea) and B(ritney) – who have jointly committed crimes are interrogated by the police separately without any chance of communicating with each other. Actually the police cannot take enough evidence to convict them of a serious crime (what the police clearly would like to do) but only of a smaller crime which would send both A and B to prison for 1 year. Without further provisions neither A nor B will have an incentive to tell the truth about the serious crime. Anticipating this discretion, the police offer a principal witness regulation: If one of the two prisoners betrays the other, she will be set free immediately while the thus convicted perpetrator will have to go to jail for 3 years. This regulation, however, only applies if just one of the two prisoners confesses and sends her accomplice to the doom. Otherwise, i.e., when both prisoners give evidence, both A and B will have to serve 2 years.

Even though this story has some drawbacks (e.g., as it is not obvious why in the case of a unilateral confession the betrayed prisoner should get a harder punishment), it highlights the cooperation dilemma that occurs in a PD situation: When acting in complete isolation from one another, the best each prisoner can do is to admit to a more serious crime – independent of what she expects her co-perpetrator to do. Such a behavior follows from a rational calculation which runs as follows: If my counterpart confesses, I will be imprisoned for 3 years if I remain silent but only for 2 years if I confess, too. Likewise, if the other one remains silent, my best strategy is again to confess because in this case I can benefit from the principal witness regulation and avoid staying in jail any longer.

In technical terms this means that in a PD, confessing is the dominant strategy for player A as well as for player B. Thus the game's outcome will be that each prisoner confesses and

goes to jail for 2 years. This, however, does not seem to be a satisfactory solution for either of them: If both remained silent instead, they could improve their situation since this would save each of them 1 year in prison. However, the problem is that the personal incentives which are guiding the agents' behavior when they act independently and in complete isolation will prevent the attainment of the more favorable outcome. Thus for players A and B, a dilemma between individual and collective rationalities arises which constitutes the distinctive feature of the PD game.

In game theory the incentive structure of the PD game and the determination of the equilibrium solution are usually described by the normal form representation as depicted in Figure 1. Here the numbers 1, 2, 3, and 4 indicate the ranking of each player's personal utilities (or "payoffs") depending on the four possible strategy pairs chosen by A and B: Player A's rank number appears in the upper and B's rank number in the lower part of each cell of the matrix. For example, number 4 in the lower left cell indicates that A reaches her most favorable position if she confesses while B does not. The stars behind the numbers indicate the optimal (Nash) reactions of each player which are obtained for player A by comparing her rank numbers in each row and, analogously, for player B by comparing her rank numbers in each column. The unique Nash equilibrium (confess, confess) of the PD game is situated in the upper left cell of the normal form where the stars of A and B coincide, i.e., where the optimal reactions of both players are consistent. A comparison between the upper left and the lower right cell (don't confess, don't confess) shows the inferiority of the noncooperative outcome relative to the hypothetical cooperative solution (Fig. 1).

Public Good Provision

The incentive structure underlying the PD game characterizes many situations of mutual cooperation between two or more agents. A very prominent example for such a cooperation problem is the voluntary provision of a public good which – particularly due to the perception of

Prisoner's Dilemma,
Fig. 1 Normal form payoff
matrix for the PD

B \ A	Confess	Don't confess
Confess	2*, 2*	1, 4*
Don't confess	4*, 1	3, 3

climate protection as a global public good – has received a lot of attention over the past few years. How a PD situation arises in the case of public good provision is even easier to explain than with Tucker's standard story about the two prisoners.

For the public good scenario, we assume that the players A and B are two countries that share the same environmental medium (water or air) and thus suffer from the same environmental damage. Each country has two options, either to take an abatement measure to curb its pollution which entails costs (c) for this country or not to abate and to carry on as usual. If none of the countries abate, no environmental damage is avoided and the environmental benefit for each country is $B(0) = 0$. If instead one single country abates, both countries will enjoy the environmental benefit $B(1) > 0$, as there are spillovers of pollution between the two countries. In other words, there is neither rivalry in nor excludability from the benefits of an abatement measure, i.e., environmental quality has the well-known properties of a public good. If not only one but both countries make an abatement effort, then each country gets an environmental benefit equal to $B(2) > B(1)$.

Concerning the relation between environmental benefits and abatement costs, it is now supposed that:

- $B(2) - c > 0$ or equivalently $B(2) > c$ which means that collective abatement efforts are beneficial for both countries.
- $B(2) - B(1) < c$ and $B(1) < c$ which says that making an own abatement effort is never worthwhile for a country irrespective of whether the other country does abate or not.

Given these conditions we obtain the same payoff structure as depicted in Figure 1 by identifying the strategy “abatement” (or “cooperation”) with “not confess” and “nonabatement” (or “noncooperation”) with “confess” in the prisoners’ story. Voluntary public good contributions made by payoff-maximizing countries non-cooperatively lead to an undersupply of the public good, i.e., to insufficient abatement activity. At the same time, there are lower net benefits for both countries compared to the cooperative outcome with collective action on public good provision. The best solution for each country is to be a free rider on the abatement activities of the other country.

In the two-country case, it is well plausible that game types which are different from PD may arise (see Osborne 2009). So if $B(1) > c$ is assumed while $B(2) - B(1) < c$ still holds, a chicken game with two asymmetric Nash equilibria results which are characterized by unilateral abatement measures of just one country. In this case unilateral abatement pays for each country (e.g., to avoid an environmental catastrophe) but not additional abatement when abatement has already been done by the other country.

Notwithstanding that, the two-agent PD serves as the prototype for a more general public goods game with many agents, where each agent affects the total public good supply not much through his own costly contributions but benefits considerably from the contributions of the many other agents. This extension to many agents further emphasizes the empirical relevance of the PD game.

Ways Out of the Dilemma

Given the unsatisfactory outcome of a non-cooperatively played PD game for its participants, naturally the question arises how to implement the cooperative solution and thus to overcome the cooperation dilemma. In order to achieve such a Pareto improvement over the noncooperative Nash equilibrium, several approaches appear to be viable.

Agents can take up negotiations and enter into an agreement in which they commit to performing the cooperative action, i.e., remaining silent or abating as in the two examples above. The difficulty of this approach is, however, that the cooperation problem may only be shifted from the level of participation to the level of internal stability of a cooperative arrangement. This is due to the fact that in a PD each agent will not want to keep her promise and implement the cooperative action after the agreement has been concluded. Defection incentives may be avoided and stability may be ensured by introducing sanctioning mechanisms. At the national level the government punishes tax evaders who do not pay their share in the cost of public goods. In line with that, PD-like scenarios already have played a major role in the traditional contract theories of the state as developed by the philosophers Th. Hobbes and J. Locke centuries ago.

However, if no central authority with sanctioning power exists (as it is typically the case for global public goods), other forms of punishment may help to solve the dilemma and to bring about mutual cooperation. If the PD is not a one-shot game but rather iterated over many periods, it becomes possible for an agent to punish her partner's noncooperation in an earlier period by refusing to cooperate in subsequent periods. The largest part of the many punishment strategies that have been conceived by game theorists are variations of Rapoport and Chammah's (1965) simple tit-for-tat TFT strategy where each player directly reciprocates by repeating the other player's preceding action. Refinements of TFT mainly differ with respect to the other player's history of actions which

trigger punishment, the length of the punishment period, and the integration of random elements in the punishment strategy. Beginning with Axelrod (1984), the success of these strategies has been tested by the use of computer tournaments. In the meantime, new strategies based on recognition and learning ("Pavlov") or gradual punishment mechanisms have proved to be even more successful in maximizing individual payoffs (Kendall et al. 2007). Note, however, that given strictly rational agents, punishment strategies only work as desired if the PD game is (at least potentially) repeated over infinitely many periods and future payoffs are not too heavily discounted.

In real-life situations punishment may also occur outside the PD by some kind of "issue linkage" which, e.g., could mean that cooperation on the provision of international public goods is fostered by the threat of trade sanctions. Moreover, if formal governmental institutions are absent, people may organize themselves and establish rules and mechanisms for an improved public good provision. In the context of local and regional commons (as fisheries or grazing lands) as specific types of public goods, the Nobel laureate E. Ostrom (1990) has demonstrated the efficiency of such self-organized governance systems as well as their potential limitations in a lot of field studies.

The basic assumptions underlying most theoretical treatments of PD are complete individual rationality and material self-interest as the players' sole motivation. These requirements are clearly not satisfied for real people in many situations (see Fehr and Schmidt 2006 for a review). Thus it does not come as a surprise that a lot of experimental studies have shown that noncooperation is not the standard outcome: In a one-shot public goods game subjects usually contribute between 40% and 60% of their endowment (Ostrom 2000), and cooperation rates in one-shot as well as in repeated PD games turn out to be significantly higher than zero ranging between 5% and 96.9% with an average of about 45% (Sally 1995; Jones 2008). The experimental setting, however, has a large impact on cooperation rates: Framing a PD as a "Stock Market Game",

for example, reduces cooperation rates while cooperation is encouraged by calling it a “Community Game” (Ellingsen et al. 2012). Moreover, individual behavior in PD games seems to be influenced by gender (Croson and Gneezy 2009) and the cultural and ethnical background of the players (Cox et al. 1991).

Recently, Khadjavi and Lange (2013) have provided evidence on behavior in a PD game played by a subject group of female inmates who after all did cooperate in 55% of all cases in the one-shot PD scenario. Ultimately the PD has thus returned to the place where it virtually started: The prison.

References

- Axelrod R (1984) The evolution of cooperation. Basic Books, New York
- Cox TH, Lobel S, McLeod PL (1991) Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Acad Manage J* 34: 827–847
- Croson R, Gneezy U (2009) Gender differences in preferences. *J Econ Lit* 47:1–27
- Ellingsen T, Johannesson M, Mollerstrom J, Munkhammar S (2012) Social framing effects: preferences or beliefs? *Games Econ Behav* 76:117–130
- Fehr E, Schmidt KM (2006) The economics of fairness, reciprocity and altruism – Experimental evidence and new theories. In: Kolm SC (ed) *Handbook of the economics of giving, altruism and reciprocity*. Elsevier, Amsterdam, pp 615–691
- Jones G (2008) Are smarter groups more cooperative? Evidence from prisoner’s dilemma experiments, 1959–2003. *J Econ Behav Organ* 68:489–497
- Kendall G, Yao X, Chong SY (2007) The iterated prisoner’s dilemma: 20 years on. *Advances in natural computation*. World Scientific Pub Co, Singapore
- Khadjavi M, Lange A (2013) Prisoners and their dilemma. *J Econ Behav Organ* 92:163–175
- Osborne MJ (2009) *An introduction to game theory*. Oxford University Press, New York
- Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge
- Ostrom E (2000) Collective action and the evolution of social norms. *Econ Perspect* 14:137–158
- Rapoport A, Chammah AM (1965) *Prisoner’s dilemma: a study of conflict and cooperation*. Michigan University Press, Ann Arbor
- Sally D (1995) Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Ration Soc* 7:58–92

Prisons

Paolo Buonanno¹ and Juan F. Vargas²

¹Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy

²Department of Economics, Universidad del Rosario, Bogota, Colombia

Synonyms

[Detention](#); [Imprisonment](#); [Incarceration](#)

Definition

Prison (or incarceration) is one of the most important forms of sanction in many modern criminal system. Incarceration may impact the overall level of crime and affect criminal behavior through several channels: incapacitation, deterrence (general deterrence and specific deterrence), rehabilitation. Understanding whether individuals respond to any of the effects outlined above is a crucial aspect for the design of effective crime reduction policies.

Introduction

The most important form of criminal sanction in many modern criminal justice systems is incarceration. The utilitarian theory of punishment developed by Cesare Beccaria (1764) and Jeremy Bentham (1789) defines the extent to which incarceration or imprisonment may be considered beneficial for society as a whole. The utilitarian rationale justifies any form of punishment, under the condition that the associated benefits are higher than the implied costs. In particular, if punishment deters or incapacitates criminals, or facilitates their rehabilitation – thus bringing benefits to most individuals of a society – then punishment is socially “good,” and therefore justifiable.

From a theoretical perspective, prisons have at least five social functions. First, incarceration mechanically incapacitates criminally active individuals from committing other crimes during the length of conviction. This is the **incapacitation effect**. Second, other things equal the threat of imprisonment could deter potential criminals from offending. Indeed, in the economic model of crime, proposed by Becker (1968), potential offenders consider the expected costs of committing a crime, which are related to the probability of being captured and condemned, and to the severity of the punishment. This is the **general deterrence effect**. In turn, the **specific deterrence effect** appears when formerly convicted individuals behave in such way to avoid returning to prison. A fourth social function of prisons arises insofar as these types of facilities could potentially help rehabilitating criminals, thus favoring their social reintegration and preventing them from recommitting crimes. **This is the rehabilitation effect**. On the other hand, however, serving time in a prison may increase the criminal propensity of inmates, for instance, by enhancing their criminal network. This is the **criminogenic effect**.

Understanding whether individuals respond to any of the effects outlined above is a crucial aspect for the design of effective crime reduction policies. Over the last decade or so, the economics literature has made significant progress in studying the empirical relevance of each of the mechanisms through which prisons can help reducing crime. With a few exceptions, most of this literature has focused on the US case, thus limiting somehow the external validity of its findings.

Incapacitation

The incapacitation effect of imprisonment is extremely challenging to estimate. The earlier criminology literature based its estimates on inmate interviews. Prisoners are asked about their criminal activity prior to their last arrest, and researchers use their responses to extrapolate the counterfactual criminal behavior in the absence of imprisonment (Marvell and Moody 1994, review this literature). The resulting

estimates, that show a large variance and thus provide little useful information, have been criticized (see Miles and Ludwig 2007) for their reliance on self-reported offenses and their sensitivity to outliers (prisoners who report an unrealistically large criminal record).

A more credible approach is the one followed by Owens (2009), who exploits a natural experiment that took place in the state of Maryland in 2001. Following a law change, judges stopped considering juvenile criminal records when sentencing adult offenders, so actual sentences went down. The author then examines the crimes committed by released offenders during the time they would have otherwise been incarcerated had the law not changed. The estimated incapacitation effects are much smaller (about one fifth) than the average effect from the research based on inmates' self-reported past offences. A critique of this strategy is, however, that it hardly distinguished between incapacitation and rehabilitation or specific deterrence.

Natural experiments of this sort are, however, less than common. A different empirical approach correlates aggregate crime and imprisonment data, usually exploiting state-level variation for the USA (see Marvell and Moody 1994). This approach has, however, two main limitations. On the one hand, estimates aggregate the incapacitation and the general deterrence effects. On the other, it is challenging to find causal effects at high levels of aggregation. Again, natural experiments can offer solutions. Levitt (1996), for instance, takes advantage of the fact that state-level prison overcrowding lawsuits exogenously reduce incarceration rates. Using an instrumental variables strategy, the author finds that a 1% increase in prison time reduces state-level violent crimes by 0.4%. While this approach solves the latter critique, it cannot disentangle the imprisonment and the general deterrence effects. Other instrumental variables estimates of the joint incapacitation/deterrence effects are Johnson and Raphael (2012) for the case of USA and Barbarino and Mastrobuoni (2014) for the case of Italy.

An interesting theoretical insight of the incapacitation effect is that, as the incarceration rate increases, the average incapacitation decreases.

This is because everything else equal higher incarceration rates imply that the marginal criminal is less dangerous. Liedka et al. (2006) for the case of the USA, Vollaard (2013) for the Dutch case, and Buonanno and Raphael (2013) for the case of Italy find evidence in favor of this hypothesis.

General Deterrence

General deterrence is a function of both prison length and prison conditions (Drago and Galbiati 2012). Regarding sentence length, the available evidence suggests that the severity of punishment does deter potential criminals from engaging in illegal behavior, as predicted by the standard economic model of crime. However, as mentioned above, the standard empirical approach – that compares incarceration rates to aggregate (state-level) crime – tends to confound the effect of incapacitation.

Again, natural experiments can come in handy. Kessler and Levitt (1999), for instance, exploit the exogenous variation in prison length provided by the so-called three strikes laws in the USA. The authors use California's Proposition 8 to show that, consistent with an aggregate deterrence effect, sentence enhancements of persistent offenders cause a 4% aggregate crime reduction immediately after the passage of the law. Likewise, exploiting the exogenous variation that the 2006 Italian Collective Clemency Bill created on the expected future punishment of released offenders if recommitting a crime, Drago et al. (2009) find that larger expected punishments dissuade recidivism.

Prison conditions also factor in the expected punishment and thus have a deterrent role. Using inmates' death rates as a proxy of (bad) conditions, Katz et al. (2003) corroborate that worse facilities are associated with less crime in the USA. The policy implication of this evidence is, however, not clear. On the one hand, prison conditions are more readily manipulatable than sentences length. On the other, it is hard to argue that facilities should be made harsher for crime reduction. Moreover, the size effects obtained by Katz et al. (2003) are rather small, so the cost-

benefit balance of manipulating prison conditions is unclear. It is perhaps more promising to study the specific deterrent effect of prisons for individual who have spent jail time.

Specific Deterrence and Rehabilitation

Prisons vary substantially in terms of the quality of facilities, the services and activities offered, and the average physical space on inmates. All of these aspects potentially affect a released individual's propensity to recidivate. However, the myriad of factors that condition inmates' future behavior, and the difficulty to measure them in a systematic and comparable fashion, make it difficult to empirically investigate specific deterrence and rehabilitation mechanisms of (future) crime prevention. The available evidence is, therefore, scarce and it has limited external validity. Moreover, in the absence of individual-level data, it is hard to empirically separate the specific deterrence effect of prisons from the general deterrence effect.

Recent evidence from Italy (Drago et al. 2011) and Colombia (Tobón 2017) suggest that worse prison conditions (respectively higher mortality and higher overcrowding) increase the likelihood that former prisoners reengage in criminal behavior. In turn, evidence from the USA (Kuziemko 2007) and France (Maurin and Ouss 2009) suggests that experiencing longer sentences reduce recidivism.

Criminogenic Effect

While the evidence above points toward a positive (crime-reducing) social function of prisons, prisons can also breed crime. Indeed, inmates not only interact with a prison's facilities and rehabilitation opportunities but also – and primarily – with other inmates. From their peers, inmates can access information (about criminal markets and illegal lucrative opportunities), be exposed to a specific set of values and social norms, and get acquainted with people with similar preferences and abilities while finding

potential complementarities that reduce the costs of committing crimes.

Understanding the scope of prison-specific peer effects is crucial for the design of policies aiming at improving the effectiveness of the prison system. For instance, there is evidence that prison networks persist after inmates are released. Specifically, using data from the 2006 Italian Collective Clemency Bill, Drago and Galbiati (2012) find that sharing a prison facility with inmates of one's same nationality predicts a correlated former prisoners' post release criminal behavior.

Also, peer effects are stronger for younger individuals, the brain of whom is more malleable, and thus more vulnerable to stimuli from the environment (Giedd 2004). Consistent with this observation, Bayer et al. (2009) estimate large peer effects on subsequent criminal behavior for juvenile offenders of the same correctional facility in Florida.

To conclude, there are several mechanisms through which prisons can influence society's crime rates. While challenging, empirically distinguishing the contribution of each one of them to reduce (or exacerbate) crime, is of foremost importance for public policy.

References

- Barbarino A, Mastrobuoni G (2014) The incapacitation effect of incarceration: evidence from several Italian collective pardons. *Am Econ J Econ Pol* 6(1):1–37
- Bayer P, Hjalmarsen R, Pozen D (2009) Building criminal capital behind bars: Peer effects in juvenile corrections. *Q J Econ* 124(1):105–147
- Beccaria C (1764) On crime and punishments
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Bentham J (1789) An introduction to the principles of morals and legislation
- Buonanno P, Raphael S (2013) Incarceration and incapacitation: evidence from the 2006 Italian collective pardon. *Am Econ Rev* 103(6):2437–2465
- Drago F, Galbiati R (2012) Indirect effects of a policy altering criminal behavior: evidence from the Italian prison experiment. *Am Econ J Appl Econ* 4(2): 199–218
- Drago F, Galbiati R, Vertova P (2009) The deterrent effects of prison: evidence from a natural experiment. *J Polit Econ* 117(2):257–280
- Drago F, Galbiati R, Vertova P (2011) Prison conditions and recidivism. *Am Law Econ Rev* 13(1):103–130
- Giedd JN (2004) Structural magnetic resonance imaging of the adolescent brain. *Ann N Y Acad Sci* 1021(1): 77–85
- Johnson R, Raphael S (2012) How much crime reduction does the marginal prisoner buy? *J Law Econ* 55(2):275–310
- Katz L, Levitt SD, Shustorovich E (2003) Prison conditions, capital punishment, and deterrence. *Am Law Econ Rev* 5:318–343
- Kessler D, Levitt SD (1999) Using sentence enhancements to distinguish between deterrence and incapacitation. *J Law Econ* 42(1):343–364
- Kuziemko I (2007) Going Off Parole: how the elimination of discretionary prison release affects the social cost of crime. Working paper no 13 380 (September), NBER, Cambridge, MA
- Levitt, SD. (1996). The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation. *Q J Econ*, 111(2), pp. 319–51
- Liedka RV, Piehl AM, Useem B (2006) The crime-control effect of incarceration: does scale matter? *Criminol Public Policy* 5:245–276
- Marvell TB, Moody CE Jr (1994) Prison population growth and crime reduction. *J Quant Criminol* 10(2):109–140
- Maurin E, Ouss A (2009) Sentence reductions and recidivism: lessons from the Bastille Day quasi experiment. IZA discussion paper no 3 990 (February), Inst. Study Labor, Bonn
- Miles TJ, Ludwig J (2007) The silence of the lambdas: deterring incapacitation research. *J Quant Criminol* 23(4):287–301
- Owens EG (2009) More time, less crime? Estimating the incapacitative effect of sentence enhancements. *J Law Econ* 52(3):551–579
- Tobón S (2017) Prison conditions, human capital and recidivism: evidence from an expansion in prison capacity, documento CEDE Universidad de Los Andes, no 2017–7
- Vollaard B (2013) Preventing crime through selective incapacitation. *Econ J* 123(567):262–284

Privacy

Ignacio N. Cofone
NYU Information Law Institute,
New York, NY, USA

Synonyms

Data privacy; Information privacy

Definition

There is no universal agreement on what is privacy. Within law and economics, privacy has been modeled as concealment of personal information (Posner 1978, 1981) and as the standard deviation in the probability distribution forming people's perception over our personal information, with privacy loss being a tighter posterior in that distribution (Cofone and Robertson 2018a).

Introduction

The law and economics literature on information privacy revolves around two dialogues, one normative and one empirical. To a large extent, these dialogues have remained separate. The normative dialogue asks whether privacy is worth protecting, to what extent, and how. The empirical dialogue focuses on better understanding consumer behavior through experimental methods.

Normative Privacy

The puzzling aspect of the normative debate is that while much of the normative economic literature argues categorically against privacy protection, most of the legal literature does not address these arguments and moves directly to second order questions of degree and manner: to what extent privacy should be protected and how.

For Posner (1978, 1981), privacy creates an information asymmetry that disadvantages the buyers in the market (e.g., employers in job interviews), thereby re-distributing wealth and creating inefficiencies. There are individuals with good and bad traits, and the individuals with bad traits want to hide them (privacy), while the individuals with good traits want to show them. Privacy allows individual to hide their traits to deceive others and thus reduces the information with which the market allocates resources. Stigler (1980) adds that people who want information about someone protected by privacy will use other less precise (and usually less intrusive and costlier) substitutes as a proxy for the data they

want to acquire, so privacy law is at best ineffective. This is similar to an application of the Coase theorem: as long as transaction costs remain low, whether there is a privacy rule (which allocates the right over information to the person to whom that information refers) or there is no privacy rule (which allocates information to whoever finds it) is irrelevant for the information's final allocation. For that reason, under a no-transaction-cost condition, introducing disturbances in the market such as information asymmetries, which would be welfare decreasing, is unjustified. Later research takes a similar approach for information in the financial market, arguing that allowing firms to collect and sell consumer data leads to increased overall welfare (Kahn et al. 2005; Kim and Wagman 2015).

Some resistance to the idea that privacy is necessarily inefficient is later introduced. Introducing a model where individuals suffer reputational harm from the loss of privacy, Daughety and Reinganum (2010) show that privacy protection allows people to engage in an optimal level of desired activities. Cofone (2017) shows that the conclusion that privacy is always inefficient holds only in a static world, but not in a dynamic world where information production is costly. Gradwohl and Smorodinsky (2017) show how privacy concerns affect people's choice of actions in strategic settings.

In the legal literature, Solove (2006) introduces a taxonomy of six different conceptions of privacy: (i) as the right to be left alone, (ii) as autonomy, (iii) as secrecy or concealment of discreditable information, (iv) as control over one's personal information, (v) as personhood and preservation of one's dignity, and (vi) as intimacy. The two conceptions that use law and economics arguments to approach privacy normatively have been secrecy (the approach taken by Posner and Stigler) and control. The latter mostly involves a discussion on whether a property right protection over personal data is warranted. While some suggest a protection regime similar to that of property (Murphy 1995; Schwartz 1999, 2004), others are skeptical of the idea. They argue that the reasons to protect property or intellectual property and the reasons to protect privacy are different (Litman

2000; Samuelson 2000) and that the debate around privacy should focus not on property but on autonomy (Cohen 2000).

More recent literature addresses privacy harms. Calo (2011) distinguishes two different dimensions of privacy harms: intrinsic privacy harms, which relate to privacy in itself, and extrinsic harms, which relate to material damage that is independent of privacy, such as financial damage resulting from identity theft or annoyances from spam – where the latter have been the focus of the prior privacy literature. Calo (2014) explores extrinsic privacy harms further and shows how people's personal data can be used to influence their decisions. Along the same lines, Conitzer et al. (2012) show that, counterintuitively, both a low and a very high level of privacy lead to price discrimination and produce extrinsic privacy harms by allowing sellers to capture all surplus, while an intermediate level of privacy is best for consumers. Regarding intrinsic privacy harms, Cofone (2016) introduces a method to quantify intrinsic privacy harms in the context of medicine, which presents a unique context in which privacy preferences can be measured with a dollar value. Cofone and Robertson (2018a) introduce a model of privacy harms that takes into account privacy preferences and therefore captures intrinsic privacy harms, separating it from reputational harms, and apply it to tort law and criminal procedure. Cofone and Robertson (2018b) apply a modification of such model to the context of consumer privacy and show that the nonbelief in the law of large numbers can justify consumer privacy regulations.

Empirical Privacy

The empirical law and economics dialogue takes place mainly in the economics of privacy literature, and it has been called “the new economics of privacy” (Acquisti et al. 2015). It consists of insights on how consumers behave regarding their personal information, and contains sometimes unexplored policy consequences. One of its central findings is that privacy decisions are affected by cognitive and behavioral biases, such as

immediate gratification or status quo bias (John et al. 2011). Acquisti et al. (2013), for example, show that consumers exhibit a much higher discrepancy than normal between their willingness to pay and their willingness to accept than in other choices: the value that consumers place in protecting their personal information differs to up to five times depending on the framing of the question. It has also been demonstrated that privacy concerns vary both with context and with personal traits (Acquisti et al. 2015).

Another central finding, and perhaps the central puzzle of this second dialogue, is the “privacy paradox.” The paradox describes the phenomenon in which people declare a high value for their private information in surveys but, in incentivized experiments, they disclose such information for low compensation. Studies on how much people value their privacy show an inconsistency between their declared concern for privacy and their actual behavior online (e.g., Norberg et al. 2007; Beresford et al. 2012). Privacy concerns are a weak predictor of the amount of personal information disclosed. However, consumers are not completely unreactive to privacy. Advertisements that are targeted and obtrusive are more likely to trigger privacy concerns among users than advertisements that do not match the content of a website (obtrusive but not targeted or do not impede visibility (targeted) but not obtrusive) (Goldfarb and Tucker 2011). People's ability to deal with privacy choices changes with their complexity (John et al. 2011). Consumers are also willing to pay to purchase from merchants that protect their privacy when privacy policies are available and their content is salient (Tsai et al. 2011). And, as shown through a series of surveys asking about personal financial information, privacy concerns have increased over time (Goldfarb and Tucker 2012).

Another major finding is the extent to which consumers are unaware of what policies they agree to and what happens with their data – uncertainty that is usually referred to as “consumer unawareness.” Milne and Culnan (2004) show that people rarely understand privacy policies and license agreements. McDonald and Cranor (2008) show that a disproportionate amount of

time and energy would be needed to read all privacy policies one is involved with. Hoofnagle et al. (2012) show that the advertising industry finds new ways of tracking that consumers are often unaware of. Hoofnagle and Urban (2014) show that consumers are unaware of the extent in which they are subject to behavioral targeting and believe that there are implied duties of confidentiality also when these do not exist. There is discussion on whether privacy notices are effective at counteracting this. Some show empirical evidence of their ineffectiveness to increase consumer awareness (Martin 2016). Others find that simplifying such disclosures along standard best-practices also has no effect (Ben-Shahar and Chilton 2016) and that consumers are unreactive of different types of language in privacy policies (Strahilevitz and Kugler 2016).

Cross-References

- [Privacy Regulation](#)
- [Signalling](#)

References

- Acquisti A, John LK, Loewenstein G (2013) What is privacy worth? *Journal of Legal Studies* 42(2):249
- Acquisti A, Brandimarte L, Loewenstein G (2015) Privacy and human behavior in the age of information. *Science* 347(6221):509
- Ben-Shahar O, Chilton A (2016) Simplification of privacy disclosures: an experimental test. *Journal of Legal Studies* 45(S2):41
- Beresford A, Kübler D, Preibusch S (2012) Unwillingness to pay for privacy: a field experiment. *Economic Letters* 117:25
- Calo R (2011) The boundaries of privacy harm. *Indiana Law Journal* 86(3):1131
- Calo R (2014) Digital market manipulation. *George Washington Law Review* 82(4):995
- Cofone IN (2016) A healthy amount of privacy: quantifying privacy concerns in medicine. *Cleveland State Law Review* 65(1):1
- Cofone IN (2017) The dynamic effect of information privacy law. *Minnesota Journal of Law, Science & Technology* 18(2):517
- Cofone IN, Robertson AZ (2018a) Privacy harms. *Hastings Law Journal* 69(4):101
- Cofone IN, Robertson AZ (2018b) Consumer privacy in a behavioral world. *Hastings Law Journal* 69(6):1193
- Cohen J (2000) Examined lives: informational privacy and the subject as object. *Stanford Law Review* 52(5):1373
- Conitzer V, Taylor CR, Wagman L (2012) Hide and seek: costly consumer privacy in a market with repeat purchases. *Marketing Science* 31(2):277
- Daughety AF, Reinganum JF (2010) Public goods, social pressure, and the choice between privacy and publicity. *American Economic Journal: Microeconomics* 2(2):191
- Goldfarb A, Tucker C (2011) Online display advertising: targeting and obtrusiveness. *Marketing Science* 30(3):389
- Goldfarb A, Tucker C (2012) Shifts in privacy concerns. *American Economic Review Papers & Proceedings* 102(3):349
- Gradwohl R, Smorodinsky R (2017) Perception games and privacy. *Games & Economic Behavior* 104:293
- Hoofnagle CJ, Urban JM (2014) Alan Westin's privacy homo economicus. *Wake Forest Law Review* 49(2):261
- Hoofnagle CJA, Soltani A, Good SN, Wambach DJ, Ayenson MD (2012) Behavioral advertising: the offer you can't refuse. *Harvard Law & Policy Review* 6(2):273
- John LK, Acquisti A, Loewenstein G (2011) Strangers on the plane: context-dependent willingness to divulge sensitive information. *Journal of Consumer Research* 37(5):858
- Kahn CM, McAndrews J, Roberds W (2005) Money is privacy. *International Economic Review* 46(2):377
- Kim JH, Wagman L (2015) Screening incentives and privacy protection in financial markets: a theoretical and empirical analysis. *RAND Journal of Economics* 46(1):1
- Litman J (2000) Information privacy/information property. *Stanford Law Review* 52(5):1283
- Martin K (2016) Do privacy notices matter? Comparing the impact of violating formal privacy notices and informal privacy norms on consumer trust online. *Journal of Legal Studies* 45(S2):191
- McDonald AM, Cranor LF (2008) The cost of reading privacy policies. *I/S: Journal of Law and Policy for the Information Society* 4: 543
- Milne GR, Culnan MJ (2004) Strategies for reducing online privacy risks: why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing* 18(3):15
- Murphy R (1995) Property rights in personal information: an economic defense of privacy. *Georgetown Law Journal* 84(7):2381
- Norberg PA, Home DR, Home DA (2007) The privacy paradox: personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs* 41(1):100
- Posner R (1978) The right of privacy. *Georgia Law Review* 12(3):393
- Posner R (1981) The economics of privacy. *American Economic Review Papers & Proceedings* 71(2):405
- Samuelson P (2000) Privacy as intellectual property? *Stanford Law Review* 52(5):1125
- Schwartz P (1999) Privacy and democracy in cyberspace. *Vanderbilt Law Review* 52(6):1607

- Schwartz P (2004) Property, privacy, and personal data. *Harvard Law Review* 117(7):2056
- Solove D (2006) A taxonomy of privacy. *University of Pennsylvania Law Review* 154(3):477
- Stigler G (1980) An introduction to privacy in economics and politics. *Journal of Legal Studies* 9(2):623
- Strahilevitz L, Kugler MB (2016) Is privacy policy language irrelevant to consumers? *Journal of Legal Studies* 45(S2):69
- Tsai JY, Egelman S, Cranor LF, Acquisti A (2011) The effect of online privacy information on purchasing behavior: an experimental study. *Information Systems Research* 22(2):254

Privacy Regulation

Fabrice Rochelandet¹ and Silvio H. T. Tai^{2,3}

¹IRCAV, Université Sorbonne Nouvelle Paris 3, Paris Cedex 05, France

²PPGE, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

³RITM, Université Paris-Sud, Paris, France

Abstract

Privacy regulation is a set of rules and enforcement tools designed to control the collection and use of personal information. Not only these rules aim at protecting privacy, but also reducing the scope of misuses of information including identity theft, higher prices, spam, and effort spent by individuals to protect their data. The spectrum of privacy instruments varies from purely (self-regulating) market-based solutions to regulatory-orientated rules and from ex ante to ex post tools. The protection of personal data involves costs for firms, such as the restriction of available information and detrimental effects on innovation. From the consumer point of view, costs are related to consent and information issues, such as reading or writing privacy charts, complying with privacy standards and adopting privacy-enhancing technologies. Since there are privacy trade-offs arising from the interaction between this regulation and other economic and social issues, the economic impacts of privacy regulation depend to the adequacy of the privacy protection arrangement to the context.

Definition

Privacy regulation is a set of rules and enforcement tools designed to control the collection and use of personal information.

Privacy Regulation

Privacy regulation is a set of rules and enforcement tools designed to control the collection and use of personal information. Not only these rules aim at protecting privacy but also reducing the scope of misuses of information including identity theft, higher prices, spam, and effort spent by individuals to protect their data. It articulates legal rules (US *Privacy Act*, French law “informatique et libertés”), the role of medias and civil society, the use of technologies (encryption, anonymous web browsing), and the actual behavior of organizations and individuals, that are more or less privacy sensitive. The weighting of each of those factors varies among countries: for instance, the role of medias in the USA proves to be stronger than in European countries. As for the public intervention to protect privacy, the USA and the EU have adopted two different approaches. While the USA is limited to ad hoc regulation, the EU puts the spotlight on general regulation based on principles.

Various instruments permit to regulate privacy. There is a spectrum of solutions between purely market-based ones (reductionist approach supporting self-regulation of firms) to regulatory-orientated rules (strict application of notice and consent mechanisms). Many authors focus on the cost-benefits analysis and comparisons of these solutions (Swire 1997; Bennett and Raab 2006; Koops et al. 2006). Privacy instruments can also range from ex ante tools that apply before the damages (e.g. legal prohibition to collect data on minors, privacy by design rules, the fair information practices) to ex post solutions (e.g. a liability rule, the right to be forgotten).

Literature on privacy regulation has emerged since the 1970s with the enactment of the US Privacy Act. In particular, Judge Richard Posner (1981) considers it as a source of market

inefficiency. By hindering the capacity of firms to gather personal data, privacy is supposed to reduce the information available in the markets about consumers, workers, borrowers, and so on. However, the advent of the Internet and digital platforms renew these issues. Moral hazard issues derive from the costs to observe the uses that firms make from the data they gather over individuals. Their well-being is undermined from such a position of imperfect or asymmetric information regarding who collect their data, when, for what purposes, and even with what consequences. At the same time, regulating firms can impose them important legal costs that could have detrimental effects on innovation and business.

The design of an efficient regulation raises many difficulties. One is to assess the optimal rules of the regulation because this requires a large amount of information that regulators rarely know. For instance, what are exactly the actual harms due to privacy infringement and how to evaluate them or who are the “polluters” in a digital environment? It proves very uneasy to know who gather what kind of information and to do what.

Another problem is that privacy protection can be a source of failures (Rubin and Lenard 2001) either because of their implementation costs (e.g., detection of unlawful behavior) or because of their consequences (e.g., restrictions on digital innovations). The costs of implementing privacy regulation might be higher than those rising from privacy abuse. For instance, internet users benefit from free online services, social networking service, and targeted advertising thanks to the exploitation of their online data. Quite often, privacy regulations require a combination of various instruments (institutional design), and thereby they can interact with other regulations and laws. Some authors support self-regulation by considering firms (advertisers, website, and platform operators) as sensitive to the consumer response to their strategies whenever these ones prove intrusive (invasive or too much targeted ads, spam, and unacceptable price discrimination) and irritate consumers (Anderson and Simester 2010). Hence, firms can engage themselves

in pro-privacy competition, for instance, by choosing to actually comply with their privacy commitment (Jamal et al. 2003). However, the efficiency of self-regulation supposes a certain ability of individuals to be informed and the capacity of firms to provide protection of privacy without excessive costs (Swire 1997). Here, the advantage of more strict regulatory solution is to permit individuals to save on costs associated with the need for complex information inherent with various privacy policies and interaction with firms that exploit their data (Milberg et al. 2000).

From an economic point of view, the efficiency of any privacy regulation is to minimize the social cost associated with the exploitation personal data while maximizing its social welfare. Therefore, there are at least three criteria to evaluate the efficiency of a privacy instrument (Rallet and Rochelandet 2011). First, a privacy instrument can be considered as efficient whenever it minimizes the transaction costs associated with its setup and enforcement. Privacy involves costs for the consumers, including the costs of consent when they want to use an online service. More generally, private agents can incur information costs, i.e., when reading or writing privacy charts, complying with privacy standards, adopting privacy-enhancing technologies, and so on. The same prevails for public institutions when they set and enforce privacy rules with the surveillance and detection of unlawful behavior. All these costs are unequally allocated among institutions, organizations, and individuals, and hence the resulting allocation could affect the social benefits of such or such privacy instrument. For instance, a property rule in favor of individuals can be inefficient because these ones are not always in the best position to assess the value of their data, leading to overestimation from their part and then less innovation and socially useful exploitation of their data. Similarly, charging too stringent privacy norms to firms can be costly for them to comply with, leading to underinvestment in innovation and quality of service.

Secondly, in order for privacy regulation to be efficient, rules assigned to every regulated agent must be followed by them. However, incentives as

well as financial or cognitive ability might be insufficient. For instance, psychological biases can lead to overexposure of individuals over digital platforms, even though they have strong preferences for privacy (Acquisti et al. 2015). In addition, firms can benefit from asymmetric information when it appears to be costly to observe their behavior, resulting in an exploitation of resources in the market for personal data (Schwartz 2004). Moreover, privacy rules can lead to firms to adopt inefficient decisions. For instance, they can decide to adopt business models (Cecere and Rochelandet 2013) or to locate their business according to the actual level of privacy (Rochelandet and Tai 2016).

Thirdly, There are privacy trade-offs arising from the interaction between this regulation and other social issues such as innovation, health, antitrust, and public order. How to best protect privacy without harming the positive effects of information sharing and exploitation? To some extent, privacy rules can complement or, conversely, be opposed to other policies through its effects on the functioning of markets. In a market perspective, ex post regulation is held to be more in favor of innovation by letting the market and innovation process work, even if it means subsequent privacy harms. However, such damages might be irremediable because they result from the use of information goods that are nonrival and then hard to seal off. However, too vigorous privacy regulation can be inconsistent with other social issues such as health, public order, freedom of speech, and innovation. For instance, the fact that European Union Data Protection Directive is more restrictive than the US Privacy Act could lead to a reduction in welfare (for instance, in health services, Miller and Tucker 2011) or a reduction in the general level of privacy protection (Rochelandet and Tai 2016 show that stringent privacy laws may lead firms to locate in less stringent countries, what weakens the actual privacy protection).

Possible approaches to privacy protection range from market-based solution to formal regulation. Thus, different frameworks are in-between self-regulation and strict regulatory protection. The economic impacts of privacy regulation

depend to the adequacy of the privacy protection arrangement to the economy context.

References

- Acquisti A, Brandimarte L, Loewenstein G (2015) Privacy and human behavior in the age of information. *Science* 347(6221)
- Anderson E, Simester D (2010) Price stickiness and customer antagonism. *Q J Econ* 125(2):729–765
- Bennett CJ, Raab CD (2006) The governance of privacy: policy instruments in global perspective. MIT Press, Cambridge, MA
- Cecere G, Rochelandet F (2013) Privacy intrusiveness and web audience: empirical evidences. *Telecommun Policy* 37(10):1004–1014
- Jamal K, Maier M, Sunder S (2003) Privacy in e-commerce: development of reporting standards, disclosure, and assurance services in an unregulated market. *J Account Res* 41(2):285–309
- Koops BJ, Prins C, Schellekens M, Lips M (eds) (2006) Starting points for ICT regulation: deconstructing prevalent policy one-liners, Information technology & law series, vol 9. T.M.C. Asser Press, The Hague
- Milberg SJ, Smith HJ, Burke SJ (2000) Information privacy: corporate management and national regulation. *Organ Sci* 11(1):35–57
- Miller AR, Tucker CE (2011) Can health care information technology save babies? *J Polit Econ* 119(2):289–324
- Posner RA (1981) The economics of privacy. *Am Econ Rev* 71(2):405–409
- Rallet A, Rochelandet F (2011) La régulation des données personnelles face au web relationnel : une voie sans issue ? *Réseaux* 29(167):19–47
- Rochelandet R, Tai SHT (2016) Do privacy laws affect the location decisions of internet firms? Evidence for privacy havens. *Eur J Law Econ* 42(2): 339–368
- Rubin PH, Lenard TM (2001) Privacy and the commercial use of personal information. Kluwer Academic Publishers, Norwell, MA, USA
- Schwartz P (2004) Property, privacy, and personal data. *Harv Law Rev* 117(7):2056–2128. <https://doi.org/10.2307/4093335>
- Swire PP (1997) Markets, self-regulation, and government enforcement in the protection of personal information. In: Privacy and self-regulation in the information age. U.S. Department of Commerce, Washington, DC

Private Law

► Customary Law

Private Property: Origins

Enrico Colombatto^{1,2} and Valerio Tavormina³

¹Università di Torino, Turin, Italy

²IREF (Institut de Recherches Économiques et Fiscales), Lyon, France

³Università Cattolica del Sacro Cuore, Milan, Italy

Abstract

This entry focuses on the legitimacy of private property and analyzes the process of first appropriation. In particular, we examine and comment the different views on the origin of private property rights that have emerged through the history of economic and legal thinking, from Democritus to de Jasay. These views have been grouped in two broad categories: consequentialism and fundamental principles. Although consequentialism is now dominant among economists and inchoate in the legal profession, we observe that it is in fact an alibi for discretionary policymaking by the authority. By definition, fundamentalist approaches generate rules that limit discretion. However, we show that some fundamentalist views rest on questionable *a priori* statements. De Jasay's argument based on the presumption of liberty is perhaps the only perspective that escapes this criticism.

Introduction

Property rights play a crucial role in economics: They define the very essence of this discipline, which studies how individuals exchange in order to enhance their welfare, subject to scarcity constraints. If there were no property rights, grabbing and looting would replace exchange, and the time horizon of any economic activity would depend on how effectively and at which cost each individual can protect the goods under his/her control.

One can identify three categories of property rights regimes: common property, centralized property, and private property (see also Waldron

2016). Common property defines a context in which the notion of property is abolished or severely restricted. It is in fact equivalent to the absence of property rights: Every individual in the relevant community/group is free to claim and appropriate everything he/she finds. The underlying idea is that individuals produce for their own immediate consumption or for the community at large (altruistic behavior with an expectation that the others will behave likewise) and that each individual has a right to take and consume whatever he likes. Theft is thus *de facto* legalized, unless it involves physical violence. Actually, one could even argue that theft no longer exists, because there are no legal owners and no member of the community has the right to prevent others from taking.

Not surprisingly, this social arrangement is of limited practical interest. It would quickly lead to the demise of the community (hardly anybody would produce goods and services, except for situations in which they can be manufactured secretly and consumed immediately) or to slavery (the most effective looters would force the rest of the community to produce and surrender their output). Of course, in the latter case, the slave masters would be the actual owners.

By contrast, centralized property corresponds to a system in which property rights are clearly assigned and belong to a central authority. This authority can be an individual, such as a dictator or an absolute sovereign. For example, since the end of the XVI century, absolute monarchies were based on the idea that God gave all the existing resources to one individual, who would then manage them in the interest of the community and in accord with God's design. The central authority can also be a set of individuals chosen through a shared procedure (elections). This set of individuals often operates by majority voting (e.g., parliaments) or assigns to other parties the power to decide on its behalf (e.g., an agency). Centrally planned economies and modern social democracies follow these patterns. In these circumstances, the expression "private property" is not absent from the debate, but it is in fact misleading. For example, in modern democracies, the central authority enjoys full power to encroach upon

somebody's property. This means that in a social democracy, the central authority allows individuals to make use of the goods in their possession within the limits defined by the central authority itself. As a result, the ultimate owner is actually the central authority, rather than the individual. In these cases, therefore, individuals engage in economic activities with other individuals, subject to their expectations about encroachment by the authority on their preferences and (temporary) property. When individuals interact with the government widely understood (policymakers, bureaucrats, agencies), they know that their counterparts are the ultimate judges of their own behavior and therefore enjoy discretionary power to which little opposition is possible. As mentioned above, property is merely temporary and far from absolute.

The recognition and enforcement of private property rights are the founding pillars of a free-market economy. Private property means that individuals have absolute, exclusive, and permanent rights on what they legally own: they can do whatever they like with their property, nobody can interfere with their decisions, and there is nobody to whom these rights must be returned after a given time period. Thus, under this regime, each individual engages in unfettered voluntary exchange, subject to his/her compliance with the freedom-from-coercion principle (no violence and no cheating are allowed) and insofar as he/she respects the private property of the other individuals. Put differently, the legitimacy of private property and the freedom-from-coercion principle specify the moral foundations of a free-market economy. By contrast, the illegitimacy of private property and the limits to private property define the features of the centralized economies, regardless of the political format – dictatorship or social democracy.

The debate on the origin and legitimacy of private property is thus of crucial importance, since it defines the very features of an economic system (institutions), the role of government, and the content of economic policymaking (see also Alchian 1965). Briefly put, the debate on the legitimacy of private property focuses on two areas. One regards property rights on natural

resources, while the other regards property rights on manmade goods, services, and intangibles. Research on the origins of private property analyzes and explains the process through which an individual can legitimately appropriate a resource never previously appropriated by somebody else. Moreover, the literature examines whether the goods, services, and ideas produced by the individual become private property of the producer; or whether they actually belong to somebody else, e.g., the community of which the producer is a member; or whether they are part of the common pool and thus belong to nobody.

The following sections provide a critical analysis of these two agendas by considering the different views developed through the history of economic and legal thinking. In particular, the next section is devoted to the consequentialist, natural right, and religious beliefs before John Locke. Section “[Aquinas and Locke on the Process of Appropriation](#)” explores the notion of property as the result of rightful appropriation. Section “[A Different Natural Right Approach to Private Property: Natural Dominion](#)” focuses on natural dominion, while section “[Recent Free-Market Agendas: From Demsetz to De Jasay](#)” discusses the more recent approaches. The final sections conclude and offer a brief outline of the debate about intellectual property rights.

Private Property Before John Locke: Democritus, Aristotle, and the Etruscan Legacy

The first attempt to justify individual appropriation – as opposed to common property – harks back to Democritus (460–370 BC). In his view, private property is justified because it leads to superior economic results (efficiency). In a sentence, individuals make better use of the resources when they own them (Diels 1903: 55 B, frag. 279). In today's wording, one would say that when private property rights are well defined and enforced, the positive and negative externalities generated by the use of a resource are minimized, and efficiency is enhanced.

According to this approach, therefore, private property is just because it leads to desirable outcomes, efficiency being the criterion that defines desirability. Surely, this view is squarely in the consequentialist camp: morality defines justice, and, in turn, the desirability of the outcomes associated with one action or one institutional arrangement defines morality.

However, considering desirability (indirectly) equivalent to justice presents one major weakness. It fails to specify who should define desirability. This is not a trivial aspect, since it is apparent that different individuals are likely to have different preferences and thus attach different meanings to the notion of “desirable.” For example, one could follow Plato and argue that since each individual should pursue virtue, and since private property encourages greed and social tensions to the detriment of virtue and peaceful coexistence, private property should be outlawed. Likewise, the very notion of efficiency is ambiguous. Efficient is whatever enhances value. Yet, value is subjective, and, therefore, the notion of efficiency can vary across individuals and groups of individuals, depending on preferences, religion, traditions, and culture. Certainly, technical efficiency is by no means enough to measure social happiness, whatever this means.

The upshot is that if one follows this consequentialist viewpoint, the presence, the boundary, and the stability of private property rights are conditional on a notion of desirability that is necessarily arbitrary and subject to change over time. Put differently, by resorting to the notion of desirability in order to legitimize private property, one actually avoids analyzing the foundations of private property and moves instead to comparing different concepts of social desirability. Yet, this comparison is hardly conducive to a useful answer. In order to avoid arbitrary rule, therefore, the notion of social desirability requires that the members of the community unanimously agree on a hypothetical desirability function (social preferences). Moreover, this notion also requires that the community members unanimously agree on the institutional tools that allow a society to obtain the desired set of goals (policymaking): when raising revenues to finance desirable public expenditure,

a tariff on car imports is different than a tax on inheritance. Failure to meet these two social choice constraints – shared goals and shared instruments – ensures that we can say little or nothing about the consequentialist justification of property in a society.

Aristotle (384–322 BC) enriched Democritus’ arguments in favor of private property by mentioning the role of human nature. According to his line of reasoning, since history and factual observation show that individuals like to own resources and goods, one must necessarily conclude that the principle of ownership is indeed part of human nature (see *Politica*: I,8 and II,5). Hence, the institution of private property is part of natural law, and denying or constraining private property would amount to negating the essence of each human being and of the natural order. The Romans also referred to natural law: the use of reason, the observation of reality, and consistency with tradition are the instruments through which cultivated men discover the natural law – “*si in unum sententiae concurrunt, id, quod ita sentiunt, legis vicem optinet*” (Gaii Inst., I § 7). And private property is part of the natural law (Gaii Inst., II §§ 66, 69, 73, 79).

However, and despite its simplicity and consistency, the Aristotelian version is also vulnerable to doubts. By claiming that the foundation of private property lies in natural law, one makes a statement that raises additional questions, rather than answers the original query. Put differently, by claiming that private property is founded on natural law, one wonders whether private property is indeed a component of natural law. One can also ask why natural law is superior to positive law (manmade legislation) and – last but not least – who has the final word on all these questions.

A third and final vision on private property was typical of the Etruscans and very popular with the Romans. As mentioned in Liggio and Chafuen (2004), religion is a family matter: each family’s ancestors are sacred and present a sacred link with the land upon which the family insists. Violating that land and – more generally – the family’s belongings was sacrilegious and usually deserved capital punishment, as the Romulus and Remus legend witnesses. Thus, there could be no religion

(and no family) without private property. Hence, private property is not a matter of economic efficiency, nor is it related to human nature. Rather, it is the necessary connection between the household and the gods. It is part of the essence of classical civilization and an element that will also play a crucial role in the defense of private property put forward by the early Fathers, together with the obligation of sharing with “the poor [...] the fruits of [the owner’s] labour” (Lactantius, *Divine Institutes*, V, 5).

Aquinas and Locke on the Process of Appropriation

St. Thomas Aquinas (1225–1274) produced two original insights, from which a theory of private property legitimized by rightful appropriation emerged. Aquinas agreed with the dominant classical worldview on the desirability/efficiency of private property. Yet, he made a second point, which was new and of great importance. He argued that unused resources do belong to humankind. However, when an individual combines them with his own labor, his claim to the resources trumps all others’ and transforms possession into private property (this is one way of reading *Summa Theologiae*, II-II, 66, 1). A few years later, this argument was forcefully repeated, clarified, and expanded by John of Paris (1250–1306), a student of Aquinas’: “lay property is not granted to the community as a whole. ..., but is acquired by individual people through their own skill, labour and diligence, and individuals, as individuals, have right and power over it and valid lordship; each person may order his own and dispose, administer, hold or alienate it as he wishes, so long as he causes no injury to anyone else since he is lord” (quoted in O’Donovan and O’Donovan 1999: 403; see also Rothbard 1995: 57; and Kilcullen and Robinson 2017).

In modern times, John Locke’s *Second Treatise* (1689, chapter V) popularized and expanded the line of theorizing initially proposed by St. Thomas and John of Paris and later enriched by the Dominican Francisco de Vitoria

(1483–1546), the Levellers, and Earl Shaftesbury (see Lepage 1985: 63; Rothbard 1995: 315–317; see also Vaughn 1980). Locke’s line of thinking can be summarized in six points:

1. God gave natural resources to humankind, and are part of the so-called state of nature.
 2. Every individual is the owner of himself.
 3. By mixing natural resources with his own manual labour, ideas and creative efforts, the individual removes the resources from state of nature and makes them his own.
 4. Property includes the resources initially appropriated as well as their fruit.
 5. Since God would disapprove of wastage or abuse, appropriation is no longer valid when private property involves wastages, or when appropriation prevents other people from satisfying their needs.
- From a normative viewpoint, Locke argues that the preservation of property is the only function of the social contract, which is the origin of government, an institution formed by individuals who agree on finding a peaceful solution to disputes. In fact,
6. Private property pre-dates government and justifies its existence.

Clearly, the notion of the “state of nature” plays a key role in the Lockean context. This concept was actually introduced by Juan de Mariana at the end of the sixteenth century. It is a synonym for the common pool and defines a situation in which resources do not belong to anybody, not even to the community as a whole. Put differently, one could argue that Locke did not theorize private property as an institution inherent in nature or legitimate per se, but rather as a desirable manmade institution that originates from common property (in the state of nature) and ownership of one’s own self. It materializes through an act of appropriation (removal of a resource from the state of nature), is consistent with God’s will, and is subject to constraints dictated by God’s will.

All in all, Locke’s theory remains incomplete and not entirely satisfactory. As observed earlier, his key elements are the emphasis on the role of

self-ownership, which allegedly justifies ownership on what the individual removes from the state of nature (the finders-are-keepers principle already formulated by Gaius, Inst. § 66), and the fact that God is not opposed to private property. However, this is Locke's weakness; it is not evident that self-ownership justifies grabbing from the common pool. Locke's view about the legitimacy of private property is more a description of how private property emerges, rather than an explanation of legitimacy. In other words, the essence of Locke's argument boils down to claiming that God created common property and that He does not object to grabbing for a good purpose (wealth creation). In this light, one may indeed argue that Locke's line of reasoning rests on consequentialism (wealth creation), possibly mixed with an act of faith (God wants people to improve their welfare well beyond what is needed to survive and private property serves that purpose). Hence, the Lockean vision not only presents the limitations typical of all consequentialist approaches but also relies on one's vision about religion and is burdened by the famous proviso which, if taken literally, de facto prevents first appropriation in the presence of scarcity or makes first appropriation conditional on everybody else's consensus.

A Different Natural Right Approach to Private Property: Natural Dominion

Consequentialism and appropriation are surely the most popular arguments in favor of private property. As mentioned earlier, consequentialism harks back to classical Greece and maintains that private property is justified by the desirable results (efficiency) it produces. By contrast, the case for appropriation was put forward in the Middle Ages; it was further developed by Locke and implies acceptance of the dominion thesis: consistent with God's design, men are free to exploit natural resources (including animals), and private property is the outcome of appropriation by an individual or a group of individuals, as long as no other human being is harmed.

However, the Middle Ages not only claimed that private property is desirable and consistent with God's design. Not long before the end of its pontificate, Pope John XXII (1249–1334) argued that private property is just also from a deontological perspective. In his view, and in contrast with other property regimes, private property stems from "pure" natural rights. As argued in the encyclical *Quia vir reprobis* (1329), which the Pope released to condemn Franciscan pauperism, God owns whatever exists, and since man has been created in the image of God, the principle of private property is necessarily embedded in the very nature of each human being. Hence, limiting private property would amount to disputing God's design. Of course, in this case, the adjective "pure" is important, since the pure version of natural rights emphasizes that these rights are embedded in each individual since his/her birth.

This natural-dominion approach differs from the Aristotelian version, according to which natural rights are those revealed by spontaneous behavior and tradition. And it also differs from the rationalist version, which owes a great deal to Grotius (1583–1645), according to whom natural law coincides with what is needed to ensure survival, while natural rights are manmade absolute rights dictated by reason and independent of the circumstances. Within the framework suggested by Grotius, therefore, sociability is the moral benchmark with reference to which all institutions are evaluated, and private property is the operational tool to meet sociability (man's inclination to live in harmony with other human beings).

Of course, the normative consequences of the natural-dominion thesis are important. According to this approach, private property is all but sacred and can never be encroached upon; whereas according to the rationalist perspective, private property ends up being the result of human constructivism and has nothing to do with religion. This also affects the very notion of right. In the former case, a right corresponds to one's freedom to use a resource as he pleases; whereas in the latter (rationalist) case, it corresponds to one's claim to enjoy something, as dictated by the (hopefully) enlightened authority.

Recent Free-Market Agendas: From Demsetz to De Jasay

In recent times, the debate on the origins of private property has subsided, with few exceptions. Indeed, mainstream theorizing has taken an evolutionary turn. Bordering with Benthamite utilitarianism, Demsetz (1967) and Pejovich (1972) pioneered an approach that neglects the moral justifications of private property and tends to treat this institution as the spontaneous result that emerges when groups of individuals face the problem of scarcity. In this context, the term “spontaneous” underscores the fact that private property is the result of human action, but not the result of constructivist design. Rather, it originates from trial and error, so that bad solutions lead to unsatisfactory performance and tend to be discarded in favor of superior institutional answers (see also Barzel 1989: Chap. 6).

In brief, it is claimed that private property emerges when unrestrained access to scarce resources leads to inefficiencies (overexploitation) and when a community finds it appropriate to generate incentives that drive individual behavior in a desirable way. Put differently, private property is a way of granting access only to selected individuals (the legitimate owners) and of managing valuable goods by internalizing potential externalities. Two consequences follow. First, private property is a manmade institution, the features of which evolve according to the current environmental conditions (transaction costs, preferences, technology). Thus, it can be neither absolute nor perpetual. Second, this evolutionary approach raises a normative issue: Who defines the features of a private property right system? This literature draws attention to the institutional entrepreneurs, who devise and experiment alternative property arrangements, and to the judges, who carry out the necessary cost-benefit analyses following which property rights are assigned and reassigned. As a result, good arrangements survive and bad arrangements are corrected, ignored, or discarded. However, when judges, experts, and lawmakers are responsible for the assignment, the recognition, and the enforcement of property rights, the outcome is *de facto* determined by government, and spontaneity

necessarily falls victim to state coercion. Not unlike Jeremy Bentham and John Stuart Mill, therefore, one hopes that coercion pursues the common interest, whatever this means; and one neglects to notice that the interests of the majority end up trumping those of the minority.

More generally, the recent evolutionary approach mentioned above follows the Lockean tradition in that it has little to say about the origins of private property. Similar to Locke, it provides a consequentialist description of how private property emerges. However, Locke based his consequentialist approach on God’s will. By contrast, the law-and-economics tradition to which Demsetz and Pejovich belong bases its consequentialist claim on the utilitarian stamp of an enlightened government. This explains why, and in contrast with the Lockean tradition, the evolutionary approach takes for granted that government precedes private property: legislators create the law in accord with utilitarian principles, and the law defines private property.

Not surprisingly, in order to recover the essence of the debate on the origin of private property, one has to focus on those libertarian scholars who deny the legitimacy of government as a coercive authority, a role that necessarily transforms the state into the source of property. In particular, by drawing on John Locke, these unorthodox authors maintain that private property preexists government and that, therefore, government cannot encroach upon private property.

In order to develop their argument, the libertarians follow two lines of reasoning, which do without God and religion and ignore Locke’s proviso (see Bouillon 2011). The first one starts from the notion of self-ownership and generalizes this principle into an ethics of private property. A second perspective focuses on a different notion of legitimacy, which turns the burden of proving the legitimacy of private property upside down by relying on the notion of Paretian optimality.

Rothbard (1974) is probably the most prominent proponent of the first view, based on an *ex contrario* argument. Rothbard justifies appropriation by observing that we live in a world of scarcity and that a resource is scarce if at least two individuals want to exploit it. Now, individual

A does not need to ask permission to exploit/appropriate the resource if it is a first appropriation. Thus, A becomes the lawful first owner. By contrast, if A must ask permission, it means that B (or somebody else before him) had already become the owner (or the co-owner) through first appropriation. Put differently, the Rothbardian perspective transforms the debate on private property into an analysis of the legitimacy of first appropriation. If one denies the right of first appropriation, all potential first owners should behave as if all resources were in common, including their own selves and the air they breathe (see also Boaz 1997). This would be absurd, since if each individual had to ask permission to act and breathe from billions of other people, the human race would quickly perish.

Rothbard also rejects the possibility that the individuals are at least partially owned by others – e.g., policymakers – who take decisions in the interest of the rest of the community. In his view, this possibility would be immoral and contradictory. It would be immoral, because all moral rules require generality (all individuals must enjoy the same rights). It would be contradictory, because accepting asymmetric ownership would be equivalent to saying that the human race is actually made of humans and non-/subhumans. According to Rothbard, therefore, the first appropriation is a law of nature because it is necessary for survival and establishes a general principle. Hence, the natural origin of private property. This principle is moral and also applies to the first appropriation of what an individual creates with his own labor.

De Jasay (1991, 1998) is an advocate of the second approach, based on the so-called presumption of liberty, i.e., on the idea that an individual can act as he pleases, unless a challenger falsifies the presumption by raising justified objections. Justified objections stem from three situations: (1) when it is apparent that an action conflicts with (spontaneous) conventions, (2) when the actor violates obligations that he had previously and voluntarily assumed, and (3) when somebody else's liberty to act is impaired (harm). The third condition is actually reminiscent of the weaker Lockean proviso, which is how Robert Nozick (1974: 178–182) identified a situation in which

individual A's action prevents others from exploiting opportunities to enhance their well-being. Although situation (3) above remains ambiguous – one could eliminate all ambiguities by understanding the term “harm” as a synonym for physical violence at the expense of the challenger – the core thesis outlined by Nozick and De Jasay is clear: the legitimacy of first appropriation is guaranteed by the lack of opposition by individuals who could claim a previous right over the good/resource. If such opposition had substance, then the finder would not be the first finder and could not keep the good. Another individual is in fact the first finder or the legitimate owner of the goods he obtained from the true first finder.

Put differently, and absent justified opposition, the finders-are-keepers rule is consistent with the presumption of liberty and necessarily represents a Pareto improvement: it makes the finder better off and makes nobody worse off. The second comer would be in a different position, since his claim to the goods appropriated by the finder would violate the presumption of liberty and would not be a Pareto improvement. Of course, the presumption of liberty also applies to private property obtained through exchange or by means of inheritance and to self-ownership.

To summarize, by articulating a presumption of liberty, De Jasay introduces a negative notion of legitimacy: all non-illegitimate actions are legitimate, and the burden of proof lies with the challenger. Since the presumption of liberty is met when nobody has justified cause to object, and since the absence of objection is also the essence of a Pareto improvement, all actions that imply a Pareto improvement are necessarily legitimate. In regard to the legitimacy of private property, therefore, the debate boils down to assessing whether an individual's property is based on a previous act of appropriation that violated the presumption of liberty. Clearly, this would be the case with theft, nationalization, or expropriation.

Preliminary Conclusions

By and large, the arguments in favor of private property have focused on three points:

characterizing property rights in the initial status quo, justifying the legitimacy of private property through consequentialism, and explaining the legitimacy of private property by resorting to fundamental principles.

In regard to the first point, the literature analyzes various possibilities. According to one version, the initial position consists in the state of nature, where resources belong to nobody and can be appropriated by the homesteader (first-user principle), possibly with some qualifications (the Lockean clauses). A second version consists in claiming that the initial position involves common ownership of the resources and that regulating the exploitation of the resources owned by a community is a political issue. In practice, the government decides who can use what and to what extent. Hence, government is in fact the origin of private property. This means that an individual manages what the government assigns to him, subject to the conditions and during the time period established by the authority. Following a third perspective, it has been claimed that the origin of private property is the individual property of one's own self, a concept that defines the very idea of individual and in the absence of which life would be impossible. Finally, according to a fourth (and radical) possibility, it is argued that the initial position is irrelevant and that, given a presumption of liberty based on self-ownership, what matters is the extent to which the various steps that have led to the current status are objectionable. Of course, lack of objections amounts to confirming the legitimacy of homesteading and exchange.

With the partial exception of the view proposed by De Jasay, private property of manufactured goods is considered an extension of the principle of self-ownership. If an individual combines the resources he owns – land, raw materials, and labor – denying his property right over the output would involve an act of aggression and amount to a violation of the freedom-from-coercion principle. On the one hand, since each unit of output is a mix of inputs owned by the producer, each unit of output is necessarily property of the producer. On the other hand, it is manifest that the manufacturer is the first finder of the result of his activity and

that nobody can raise justified objections to the manufacturer's exercise of his liberty (acting as a producer and producing and appropriating the result).

As mentioned earlier, the second set of explanations regarding the origins and foundations of private property focuses on private property as an institutional arrangement justified by efficiency: it ensures better economic outcomes and wards off social tensions. Although this approach is dominant among economists and inchoate in the legal profession, it remains problematic. As a matter of fact, explaining why private property exists is not the same as analyzing why it is legitimate (and thus immune to manmade rule making). In particular, it raises the problem of specifying who decides about efficiency: the homesteader or the ruler? As result, different perspectives produce different answers.

Finally, the fundamentalist approaches try to justify private property by resorting to religion, or to natural rights, or to expedience (which differs from consequentialism). As we observed, religion is problematic, in that it has no universal value. If one justifies private property by appealing to faith, different religious views can lead to radically different conclusions. Natural rights are different, since the various approaches underscore one or more natural traits of all human creatures and give birth to different theories of property in accord with human nature(s). Hence, if one assumes that all individuals share a set of natural rights, then private property is inviolable insofar as it is recognized as a natural right or strictly derived from a natural right. In a similar vein, if freedom from aggression is recognized as a natural right, private property is legitimate as long as it does not involve aggression (homesteading doesn't, by definition) and as long as appropriation by the comers does not imply violence (voluntary exchange doesn't). However, if one has doubts about the "natural" nature of private property, or about the general validity of the freedom-from-coercion principle, then the legitimacy of this institution becomes questionable. By contrast, expedience rests on identifying the presumption of liberty as the simplest way of organizing a community (De Jasay 2005). This presumption

leads to the finders-are-keepers rule and requires the second comer to challenge the incumbent owner. The reader might notice three interesting aspects. According to this vision, (1) the origin of private property rests on a negative (*ex post*) notion of legitimacy; (2) the presumption of liberty is not a moral argument in favor of liberty, but rather an operating principle that minimizes costs; and (3) the negative notion of legitimacy ensures that the burden of proof lies with the challenger, a choice that has no moral content, but enhances social cooperation.

An Extension to Intellectual Property

Theories about the foundations of private property have usually focused on natural resources and the fruits of men's activities. In recent times, however, considerable attention has been devoted to the property rights on intangibles, with particular reference to two areas: information and patents. In all these cases, the research agendas share one common feature and are relatively simple.

In contrast with what happens in the realms of natural resources and material goods, intangibles are frequently characterized by the presence of free riding, i.e., by the possibility that individual A benefits from the activity of individual B, regardless of whether A and B are bound by a contract. Knowledge, information, and ideas are typical examples. Hence, when free riding occurs, A's welfare increases thanks to B's labor and talents, regardless of whether B agrees or feels slighted. This raises a question. Is free riding a problem or perhaps a crime against private property? Or is it just a fact of life, which some people consider harmless and others undesirable?

The research agendas follow the lines of thinking suggested by the questions above, depending on whether one believes that free riding should be restrained or, rather, freely allowed. One approach alludes to fairness/envy. These notions play a significant role when some agents have privileged access to relevant information and exploit this privilege in dealing with allegedly uninformed counterparts. Those who lament that uneven information amounts to unfairness require that

the informed disclose everything they know, except for situations in which they actually bought or researched the information from which they benefit or in which the information is easily accessible to anybody (in which case there would be no privilege). Legislation against insider trading follows this view: it is all right if individual A reads the business press or carries out extensive research about company Z and decides to operate accordingly on the stock exchange. However, A cannot exploit the information he obtains if he enjoys an allegedly privileged position. In the case of insider trading, the privilege consists in being an employee or an administrator of Z.

By contrast, the legislation regarding patents exemplifies other views on property rights. One is based on the claim to self-ownership and one on consequentialism. The argument stemming from the assumption of self-ownership rests on the fact that the mind is part of one's own self and that, therefore, what is produced by one's mind is necessarily an extension of the individual, who homesteads it. Put differently, the second inventor/author/discoverer cannot claim property rights on something that others have already removed from the state of nature, a state of nature that includes knowledge, skills, and artistic concepts that humankind ignores. Although this thesis has merit, it runs into a number of problems, which we shall only mention. Most innovations are based on earlier insights and discoveries. This approach would require that all inventors track – and ask permission to – all the legitimate owners of the knowledge that the current inventor is using. Moreover, the difference between a new invention, a marginal improvement on an idea already existing, and the bare use of an existing idea is frequently far from clear. Finally, the very act of exercising one's intellectual abilities hardly justifies preventing other individuals from using their own intellectual abilities, which include thinking, observing, and possibly reproducing. Of course, this line of reasoning resonates with the presumption of liberty mentioned in the previous section, a presumption often used as an argument against the existence of property rights on intangibles and thus against the legitimacy of patents. In other words, one may argue that the theory based on

the extension of one's self regards the appropriation of the result one obtains by using his talent, his labor, his knowledge, and his imagination. Yet, a result is not a process. Put differently, discovering a process and making use of it do not remove knowledge from the state of nature. It merely makes knowledge accessible to the inventor and to the rest of the community. If this is true, one can thus conclude that enriching the state of nature does not involve homesteading and that involuntary altruism is not a source of rights.

The previous comments help understand why the current legislation on intellectual property rights is founded on consequentialism rather than on philosophical theorizing. As it often happens in law-and-economics debates, the debate presents different views. We list two of them. On the one hand, a large portion of the literature neglects to discuss the nature and foundations of property rights on intellectual property. Rather, it maintains that governments driven by utilitarian principles are justified in enforcing property rights on intangibles in order to compensate their authors for the damage suffered at the hands of free riders. In other words, the origin of intellectual property rights is the government, which assigns them in the common interest. The consequentialist counterargument is that too much protection awarded to a set of inventors may prevent potential competitors from improving on the existing technology and developing new insights. The upshot is that the presence of free riding justifies patents and other barriers to entry. However, these barriers should expire after some time and allow new competitors to enter the scene at a relatively low cost.

Yet, there is also a second and more recent consequentialist perspective. As mentioned at the beginning of section “[Recent Free-Market Agendas: From Demsetz to De Jasay](#),” some eminent scholars argue that private property originates as a response to scarcity. In other words, private property is legitimate because it is an efficient way of exploiting scarce resources: it enhances exchange, and it allows individuals to distribute consumption over time, possibly taking into consideration also the potential benefits enjoyed by future generations. Intangibles,

however, do not present a problem of scarcity. The use of knowledge by one individual does not prevent other individuals from exploiting that very knowledge. Hence, absent scarcity, private property has no reason to exist, and the debate on the origin of private property in the realm of intangibles is moot. From a normative viewpoint, therefore, patents have no legitimacy.

Cross-References

- [Consequentialism](#)
- [Copyright](#)
- [Economic Efficiency](#)
- [Intellectual Property: Economic Justification](#)

References

- Alchian A (1965) Some economics of property rights. *Il Politico* 30(4):816–829
- Barzel (1989) *Economic analysis of property rights*. Cambridge University Press, Cambridge
- Boaz D (1997) *Libertarianism*, chapter 3. The Free Press, New York
- Bouillon H (2011) *Business ethics and the Austrian tradition in economics*, chapter 2. Routledge, London
- de Jasay A (1991) *Choice, contract, consent: a restatement of liberalism*. Institute of Economic Affairs, London
- de Jasay A (1998) *Justice*. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*. MacMillan Reference Limited, London, pp 400–409
- de Jasay A (2005) Freedom from a mainly logical perspective. *Philosophy* 80(4):565–584
- Demsetz H (1967) Toward a theory of property rights. *Am Econ Rev* 57(2):346–359
- Diels H (1903) *Die Fragmente der Vorsokratiker*. Weidmannsche Buchhandlung, Berlin
- Kilcullen J, Robinson J (2017) *Medieval political philosophy*. In: Zalta EN (ed) *The Stanford encyclopedia of philosophy* (Summer 2017 Edition). forthcoming URL = <https://plato.stanford.edu/archives/sum2017/entries/medieval-political/>
- Lepage H (1985) *Pourquoi la propriété*. Hachette, Paris
- Liggio L, Chafuen A (2004) *Cultural and religious foundations of private property*. In: Colombatto E (ed) *The Elgar companion to the economics of property rights*. Edward Elgar, Cheltenham, pp 3–47
- Locke J (1689) In: Hollis T (ed) *Two treatises of government*, London. A Millar et al. 1764 – available at <http://oll.libertyfund.org/titles/222>

- Nozick R (1974) *Anarchy, state and utopia*. Basic Books, New York
- O'Donovan O, O'Donovan JL (eds) (1999) *From Irenaeus to Grotius: a sourcebook in Christian political thought*. Wm. B. Eerdmans Publishing, Grand Rapids
- Pejovich S (1972) Towards an economic theory of the creation and specification of property rights. *Rev Soc Econ* 30(3):309–325
- Rothbard MN (1974) Justice and property rights. In: Blumenfeld S (ed) *Property in a humane economy*. Open Court, La Salle, pp 101–122
- Rothbard MN (1995) *Economic thought before Adam Smith*, vol I. Edward Elgar, Cheltenham
- Vaugh K (1980) John Locke's theory of property: problems of interpretation. *Lit Lib: Rev Contemp Lib Thought* III (1):5–37
- Waldron J (2016) Property and ownership. In: Zalta EN (ed) *The stanford encyclopedia of philosophy* (Winter 2016 Edition). URL = <https://plato.stanford.edu/archives/win2016/entries/property/>

Privatization

Brady P. Gordon
Trinity College, University of Dublin, Dublin,
Ireland

Abstract

Privatization as a tool for economic management emerged in the 1970s and 1980s in response to the decline of Keynesian economics and the collapse of communism and has expanded to become a pillar of public policy in all of the world's major economies. The economic benefits of privatization emerged out of new economic theories which applied long-established principles of market failure to weaknesses in public sector economic governance. It is seen as a useful way of increasing efficiency by introducing private competition to otherwise inefficient, monopolistic, and politicized operations in the public sector. Since the late twentieth century, privatization outcomes have been increasingly recognized as context-dependent and socially contested, but privatization remains a pervasive and useful instrument of government policy in pursuit of a wide range of economic and social aims.

Definition

Privatization describes the transfer of government assets or functions from public to private ownership. Privatization is the opposite of nationalization, in which governments acquire privately owned assets or operations.

Types of Privatization

Privatization encompasses four broad types of transactions:

Denationalization consists of the divestiture of state-owned assets to private ownership, typically through sale or lease. The first large-scale privatization campaigns at state level consisted largely of denationalization. These were led by the UK, the USA, and Chilean governments in the 1980s and saw the denationalization of many public utilities (i.e., industries in the telecommunications, energy, transportation, sewage, and water sectors), traditionally associated with the public sector. Denationalization allows governments to expose what might be inefficient, subsidized and politicized government-owned monopolies to private competition, and increase short-term revenues.

Concessions consist of the granting of a legal right to fund or construct an infrastructure project and/or operate a public service. The most common example of a concession is the right of a private company to construct and operate a toll road or bridge. Concessions allow governments to engage in public works projects which might otherwise be unaffordable or politically infeasible for the government.

Outsourcing or tendering consists of contracting with a private entity to finance or deliver a public service that had hitherto been carried out in-house by a branch of the public service. Tendering processes are often considered to allow governments greater control over privatization outcomes than simple divestment, since it allows the government to set minimum service standards and multiple criteria as the basis for competition. Outsourcing may occur for

entire services or it can be introduced at specific levels of the supply chain in order to open financing, production, or service provision stages to private competition. This may be useful where a government wishes to retain control over the function as a whole. For example, public policy may require government control over the provision of health care, but the operation of specific ambulance services or the construction of hospitals may be outsourced to the private sector without sacrificing that commitment. In the USA, federal and state governments frequently outsource even core state functions such as military and security services.

Deregulation consists of the opening of state-operated or state-controlled activities to private sector competition, allowing the government to retain its ownership but exposing the public entity to increased competition.

In addition to the main categories of privatization, there are also innumerable intermediate forms of privatization. For example, these might consist of public-private partnerships (PPPs) or private finance initiatives (PFIs), wherein an asset is financed entirely or partially by the private sector and the investment is recouped either through reimbursement from the government or through the accrual of profits from the operation of the service.

Historical Development of Privatization

For most of the twentieth century, until the 1970s, public ownership of key industries was a basic tenet of economic policy in each of the world's major economic blocs. In particular, public utilities such as energy, telecommunications, water, sewage, and transport were nearly universally state-owned industries. This began to change in the 1970's and 1980's, when new economic theories began to apply principles of market failure to weaknesses in public management, and the socialist economies of Central and Eastern Europe fell into rapid decline. By the end of the twentieth century, privatization had become a prevalent

feature of public service provision in all of the world's major economies. Privatization is now recognised as a useful economic tool with different costs and benefits depending on the economic and social context.

In the liberal market economies of the USA and Western Europe, Keynesian macroeconomic theory prevailed from the late 1930s until the late 1970s. Keynesian economics emphasized the influence of aggregate demand on economic output and therefore the facilitating potential of government intervention in the economy. Keynesianism reached its zenith at the 1944 Bretton Woods Conference with the establishment of the global monetary system, the International Monetary Fund (IMF), and the World Bank. Following the "total war" state mobilizations of World War II, most major manufacturing and utilities industries in western economies were state-owned, and it was widely accepted that key industries should remain so. There were some important instances of privatization during this period, including the privatization of the British steel industry in the 1950s, and coordinated divestitures by the West German government in the 1960s (e.g., the sale of a majority state in Volkswagen in 1961). For the most part, however, public utilities and other important industries remained state-owned. This consensus was entrenched further as increasing prosperity after World War II led to the expansion of welfare states and social-market economies around the world. Keynesian economics informed the management of western market economies for much of the postwar economic expansion.

In Eastern and Central Europe, the rise of socialism sometimes led to near-total appropriation of enterprise and centralised planning of economies by the state during the postwar period. The constitutions of socialist countries based on the 1936 USSR Constitution, for example, stipulated that all productive assets (including firms) were property of "all the people" (i.e., the state). Private markets were severely constricted in eastern-bloc countries.

In South America, Asia, and Africa, encouraged by the new global Keynesian institutions and governments of both socialist and capitalist

countries, developing nations equally saw state control and investment as the best way to achieve fast “catch up” in modernization and industrialization (Parker 2000).

Until the 1970s, economic theorists in all the world’s major economic blocs were therefore primarily concerned with identifying imperfections in private markets and making adjustments through public intervention. The reign of Keynesianism in the west and socialism in the east meant that economic thought to the 1980s was dominated by the literature on market failure.

Privatization only began to take root in developed countries during the ascendancy of neoliberal economic theory in the late 1970s and early 1980s. Applying lessons of the market-failure literature to the economic performance of states, economists began to identify failures in the government’s ability to perceive public wants, and further weaknesses in incentives to satisfy those wants efficiently. Economists such as Ronald Coase and Harold Demsetz questioned the prevailing logic that the mere existence of market failures warranted government intervention. Coase argued against government intervention without first analyzing whether such intervention would improve economic outcomes (Coase 1960), and Demsetz identified problematic inefficiencies and information asymmetries in government-managed organizations (Demsetz 1969, 1968). The essential defense of privatization which emerged from the literature was that the private sector was capable of providing public goods and services more efficiently than governments due to the disciplining effect of competition and profit motives. Transferring public functions to private, for-profit enterprises instilled incentives to cut costs and produce goods in response to demand. Exposure to competition and supply and demand principles were predicted to improve efficiencies in both production and allocation.

Neoliberal economic doctrines began to gain political traction in developed economies in response to the failure of Keynesian policy to deal with the economic malaise and stagflation of the 1970s. It was first enacted at the level of national policy in the UK under the administration of Prime Minister Margaret Thatcher (1979–1990),

in the USA under the administration of President Ronald Reagan (1980–1988), and in Chile under the administration of President Augusto Pinochet (1974–1990). This was followed by other Western European economies within a decade. In the European Community in the 1990s, the abolition of customs barriers, the liberalization of national markets, common merger controls, and prohibitions on state aid facilitated the entry of private operators to areas that had previously been dominated by governments (Parker 1998). New European rules implied limits on public interventions. Portugal and France, for example, were required to amend their privatization legislation to allow foreign private investment.

In the transitional economies of Central and Eastern Europe, privatization took root as the result of an altogether different process: The collapse of Communism. Following the collapse of the USSR in 1989, privatization became a pillar of state policy in transitional economies. In Russia itself, two major waves of privatization occurred: The first wave, from 1992 to 1994, saw major state enterprises converted to joint-stock companies and put up for public auction, with Russian citizens given vouchers redeemable for stock in the new companies. A second wave of privatization saw the divestment of much of the energy and telecoms sector to private investors in the late 1990s and early 2000s (Leonard and Pitt-Watson 2013). By 1999, only 4% of registered companies in Russia were owned by the state (Munkholt 2000). Similarly, in post-Warsaw pact and other socialist countries, privatization occurred as both an economic and political reaction to the fall of communism. After the collapse, 90% of industrial capital in Eastern Europe remained in public hands (Lipton and Sachs 1990). Privatization thereafter took a variety of forms, often appearing as employee share purchases, public vouchers, or sales to cooperatives and private shareholders. Economic aims included increased efficiency generally, but often also aimed at increasing economic trade with Western Europe. Political aims included returning property to private owners who might have been deprived of it by the nationalization process under socialist governments.

In developing and emerging countries in Asia, Latin American, and Africa, impetus for privatization derived from two significant sources: As a precondition of foreign investment from developed market economies and as a condition for international aid and structural adjustment programs. Conditionality for financial assistance from the IMF, World Bank, and developed-country governments frequently requires states to meet certain objectives of economic liberalization or to pursue debt sustainability by denationalizing assets. Privatization policies have also often been used to counteract a perceived culture of political instability and to signal a commitment to free enterprise. At the dawn of privatization, between 1988 and 1993, the value of privatizations in developing countries reached US \$96bn (World Bank 1995).

China has embraced limited privatization as part of its “Socialist Market System”, while keeping political control and ownership of large, key industries in the economy. Yet it has not been immune to the growth of privatization. From 1978 to 1991, the state’s share in industrial output fell from 77.6% to 52.9% (Wang 1994). In 1995, the State Council endorsed a policy of maintaining state control on large industries and allowing competition in small industries. The role of the state in industry has gradually declined as part of a measured and evolving process of opening up certain sectors of the economy to international investment without compromising political control of key economic sectors.

By the end of the twentieth century, the use of private actors in facilitating the operations of government had become ubiquitous in government policy around the globe. Privatization is no longer the exclusive domain of certain political systems or poles of the political spectrum. The discussion of governmental functions in liberal market economies is now suffuse with market-style language. Public service has been increasingly submitted to what have been called “New Public Management” regimes based on market-based theories (Hughes 2003). Private actors operate prison and immigration systems, determine regulatory standards for key industries, provide health care, and even conduct policing and military operations on behalf of governments.

Within this new global paradigm, the outcomes and theory of privatization have also begun to come under criticism. The results of privatization have often been mixed and context-conditional. Privatization has generated political opposition in many countries, and the theoretical tenets have come under challenge. In theory, policy, and practice, the extent to which privatization should be enacted in each country has become a much more nuanced and particulate debate.

The Economics of Privatization

The rationales for privatization are invariably couched in terms of economic efficiency. A number of rationales emerge from the literature (Parker 2000):

- Increased economic efficiency by introducing private competition and profit motives to what might otherwise be inefficient, monopolistic, and politicized operations under the public sector
- The reduction of government debt by selling financially burdensome assets or operations
- Increased welfare through increased responsiveness to supply and demand and cheaper goods from competition
- The avoidance of capture by politicians overseeing the provision of services by special interest groups such as trade unions or professional bodies
- Cost savings in accomplishing public works by allowing private investment to supplement or displace government financing
- Increased private ownership of capital and the fostering of liberal markets
- The attraction of mobile capital and foreign direct investment in an increasingly globalized world

Broadly, each of the efficiency arguments for privatization is based on two tenets of economic reasoning: (1) Efficiency gains from competition between firms and (2) efficiency gains from information and incentives within private firms (Dunsire et al. 1988; Vickers and Yarrow 1988). Both sources of reasoning stem from the premise that governments do not always pursue maximum

welfare and that imperfections in markets do not automatically justify government intervention.

Economic Efficiencies from Product Market Competition

Economic efficiency is comprised of allocative efficiency and production efficiency. Allocative efficiency occurs when producers produce only those types of goods that are useful and in demand in society, and where the price is equal to marginal cost. In essence, it is concerned that resources are not expended on less-optimal products, and that the most optimal distribution of resources is achieved. Production efficiency occurs when products are created using the least resources possible for the most efficient level of production. Production efficiency in turn has two constituent elements: technical efficiency and price efficiency. Technical efficiency is concerned with the effectiveness of inputs in creating an output. So, for example, a firm which employs too many workers or has underutilized equipment is technically inefficient. Price efficiency is the degree to which a price reflects all available supply and demand information. So, for example, charging a high price for a product that is in low demand would be price inefficient.

Under conditions of competition, firms will seek to optimize allocative and production efficiency, and price will reflect the most efficient marginal costs available. In essence, firms will strive to create the most useful products for the least cost, and charge prices that reflect supply and demand, maximizing economic well-being (Kirzner 1997). Without competition, a firm may restrict output and raise prices. It will not need to pursue technical or price efficiency. Where the monopoly is state-owned, this applies to governments. Governments which own firms have incentives to eliminate competition by regulating the market, subsidizing inefficient industry, or erecting legal barriers to entry (Littlechild 1986).

Efficiencies from Private Firm Ownership

The second source of efficiency from privatization derives from principal-agent theory. Principal-agent theory concerns the information asymmetries

between principals (owners) and agents (managers) of firms. Under this theory, the private sector is considered to have more effective mechanisms of control in place between those who own the firm and those who drive its performance. The first and most obvious reason for this is that incentives for management performance in the private market are often unrestrained by politics and public law. Management may be generously incentivized or quickly demoted more easily than in the public sector, where set pay scales, public unions, and political costs may restrict action. Secondly, a private capital market is considered to have more effective means of control over managers. Managers and boards of directors can be pressured to resign by shareholders. A firm which is losing the value of its shares may see investors sell their shares, further precipitating a drop in the firm's share price. If the share price drops low enough, the firm becomes susceptible to takeovers by more efficient firms, who may then purchase the firm's asset and turn them to more productive uses. In essence, it is argued that inefficient management cannot survive in the private market.

In the public sector, agents are responsible to politicians. The literature on public choice indicates two problems with governments as principal: The first is that the objectives of government are some combination of social welfare, the politician or government's own personal goals, and the aims of special interests groups (Niskanen 1971). Even if weighted only slightly towards the latter two goals, this represents a suboptimal welfare result. Politicians may be subject to political capture and require firms to pursue secondary objectives such as industry employment, high wages, and other aims which may run counter to productive efficiency. These wider social objectives may genuinely contribute to the public welfare; however the literature on public choice suggests that governments are no less susceptible to rent-seeking than private monopolies (Buchanan 1972). For example, a politician elected due to the support of labor interest groups may be insulated from citizen pressure and instead be incentivized to promote high wages in the industry instead of efficient outputs that benefit public welfare as a whole. The second problem is

that, even if the government solely pursues the maximization of efficiency, governments lack information on demand and supply costs. Only under conditions of competition is demand information readily available.

Legal Controls and Methods of Privatization

The divestiture or delegation of basic public services to private actors can raise legal questions both as to legal methods of control on the actions of government and what legal forms privatization may take.

Legal Controls on Privatization

Privatizations are typically subject to three types of legal control: Constitutional law, administrative law, and international or human rights law.

Constitutional law places limits on the existence and exercise of public power. This can include limits on the government's power to privatize. Article 34 of the French Constitution of 1958, for example, states that rules governing the transfer of public assets to the private sector shall be set by law. Privatizations are therefore required to be approved by the legislature. The constitutional principle of equality among citizens and Article 17 of the Declaration of Human Rights (which requires just compensation for expropriation) has been held by the French Constitutional Council to prohibit the sale of public assets at less than the real value. Privatization transactions are therefore justiciable both on the basis of the separation of powers and administrative value. In other countries, such as the UK, the doctrine of parliamentary sovereignty and the discretionary powers of government mean that divestitures by either the legislature (by statute) or the executive (by contract) will not be subject to judicial review without an explicit statutory basis. In the EU, the European Court of Justice has long accepted the existence of an inherent power to delegate where necessary to exercise a power conferred by the Member States,

subject to the limit that privatizations on that basis cannot exceed the limits of the power conferred (*Meroni v. High Authority* [1957–58] ECR 133). In the USA, the Fifth and Fourteenth amendments to the US Constitution contain “due process” clauses which stipulate basic elements of fundamental fairness such as the opportunity to be heard and impartiality in decision-making. The courts have derived a constitutional doctrine of non-delegation from these clauses that requires governments to comply with due process and the separation of powers when delegating.

Administrative law seeks to ensure fairness and compliance with the rule of law in the exercise of state power. In both civil law and common law countries, the legislature may set out detailed codes on administrative procedures and public decision-making. In the USA, federal and state legislation encompass a range of procedural acts. The Federal Procedure Act 1946, for example, offers equal or greater procedural protection than even the due process clause of the constitution. In the EU, conditions and limits on the delegation of state power are regulated in detail by Financial Regulation 1081/2010 [2010] OJ L311/9, which sets out the tasks that can be delegated to private entities or agencies. In the USA and UK, the common law principles of judicial review grant the courts a jurisdiction to review the actions of government decision-makers against standards of procedural unfairness, transparency, bias, and rationality (Taggart 1997).

International and human rights law seeks to protect individuals against the inappropriate use of state power. This becomes particularly relevant to privatization when human rights principles contain procedural protections. For example, Article 6 of the European Convention on Human Rights (ECHR) enshrines the right to a public hearing before an independent and impartial tribunal where civil rights are being determined. It is not difficult to imagine how privatization of a social function such as welfare or public housing could have the potential to deprive the individual of the right to an impartial tribunal against determinations of

their civil rights by the private body. The existence of profit motives of a private operator may provide an incentive to disqualify recipients of welfare or housing benefits, for example. Indeed, in *Feldbrugge v. Netherlands* (1986) 8 EHRR 425, the European Court of Human Rights held that such benefits may give rise to civil rights.

Legal Methods of Privatization

In legal terms, the three main techniques for privatization are legislation, contracts, and legal grants (Donnelly 2007). Each method is subject to a different mixture of legal controls in each state.

Privatization by legislation is, in legal terms, perhaps the most conventional method for privatization. Legislative acts are typically justifiable under general constitutional law principles and are therefore likely to be subject to judicial review principles. The same will generally be true of a privatization engaged in by the executive, subject to rules set out in legislation. In many jurisdictions, however, there are also countervailing limits to the legal control of privatization by legislation. In the UK, the doctrine of parliamentary sovereignty precludes judicial review of legislative acts unless provided explicitly by statute. In practice, statutes granting executive powers are typically drafted broadly and are unlikely to preclude privatization where it may be necessary for the function given.

Privatization by contract transfers the source of the private entity's legal responsibility from an act of the legislature to a private contract. Whereas executive and legislative action is bound by constitutional, administrative, and international law, a private party governed by private contract is typically governed by the terms of the contract. In the UK, decisions by private actors operating under contract are not typically subject to judicial review and administrative law principles, even when exercising a public power (Freedland 1994). In the USA, the definition of "state actor" to which constitutional and administrative procedural rules apply has evolved to exclude most private

actors acting under a contract. Private actors are exempt from disclosure, oversight, bias, and ethical obligations which apply to public bodies under the US Code of Laws, for instance. Similarly, international human rights obligations such as the ECHR bind state parties only, despite the fact that private entities exercising state power may exercise the same capacity to interfere with individual rights. In the UK, for example, despite the fact that the Human Rights Act 1998 applies to "functions of a public nature", the public or private nature of the entity has tended to predominate, rather than its function. Since the legal controls on the private actor are often limited to the scope of the contract, its drafting will have a significant impact on the government's ability to control outcomes. On one hand, government contracts typically fall within the scope of public procurement regulations, and therefore may allow or require governments to include public-service objectives in pecuniary contracts. On the other hand, many government services are highly variable and it may be inappropriate to limit the criteria for performance evaluation. A contract which pays a prison service on a per head basis, for example, may instate a perverse incentive to extend sentences. Yet the more complex the contract, the more difficult the privatization is to monitor, and the less benefit is gained from privatization.

Legal grants consist of the granting of a legal right or funding to build or operate an asset. In contrast with contracts, grants are typically governed by authorizing legislation and often fall outside the scope of legislated procurement regimes. There is often a broader latitude given to the recipient as to how to achieve the aims of the privatization. This can result in an arbitrary difference in legal treatment between contracts and grants (Donnelly 2007).

Criticisms of Privatization

Economic Criticisms

Privatization theory has come under pressure from the emergence of new theories such as New

Institutional Economics theory, which question the predictive power of privatization models and attempts to demonstrate that the optimal allocation of functions cannot be determined a priori in all cases. Privatization is not universally beneficial, and the optimum allocation of ownership often depends on the economic and social attributes of a specific country. Some contemporary economists, such as Joseph Stiglitz, argue that government can generally outperform private enterprise in economic outcomes (Stiglitz 1995). Critics argue that privatization should only take place where the market could perform as well as the benevolent government (Sappington and Stiglitz 1987). This criticism centers around two main heads:

First, markets are often imperfect, and privatization only generates benefits under specific market conditions. The choice is not always between government monopoly and fully competitive capital markets with effective competition. Many state-owned industries maintain large economies of scale (such as rail networks or telecommunication towers) which they then take with them to the private market to create a private monopoly. There is also evidence that private monopolists may be more harmful than public monopolists (Pint 1991). The OECD concludes that private sector monopolists are much more efficient under conditions of competition but also much better at extracting rents under monopolistic conditions (OECD 2000). They may also be less susceptible to regulation: attempts to regulate them may simply result in the private firm passing on the cost of regulation to consumers without bearing any cost itself (De Fraja 1993). Some economists argue that competition is the crucial factor in efficiency, rather than the public or private nature of the owner (Milward and Parker 1983; Vickers and Yarrow 1988).

Second, capital markets may not make for more efficient principal-agent control. Private sector management may still be able to pursue their own objectives to a large degree without shareholder reaction, and a loss of shareholder value may not necessarily lead to takeovers or share sell-offs.

The empirical outcomes of privatization have not always shown efficiency improvements, particularly where privatization has merely resulted in the transfer of a public-sector monopoly to a private-sector monopoly. While there is strong evidence that privatization lowers prices and improves services where markets work effectively, the outcomes can often be the opposite where the privatization did not result in competition (OECD 2000). In other cases, critics have argued that state-owned assets and rights have been sold at an under-value, with no residual government ability to regulate the industry and ensure pareto outcomes for the citizens. The optimum allocation of private and public control is now a matter of theoretical, empirical, and political debate in individual countries and economies. Economists in favor of privatization have continued to argue that the private sector is a more efficient provider of most government functions. For example, Megginson observes that “private ownership must be considered superior to state ownership in all but the most narrowly defined fields or under very special circumstances” (Megginson 2005). Yet other authors argue that privatization increases costs just as often as it reduces them (Brudney et al. 2004). In a survey of the literature, Parker concludes “Empirical evidence, like the economic theory, does not allow us to be confident of predicting accurately the outcome of any privatization on corporate performance” (Parker 2000).

Legal Criticisms

While the scope of public law varies between jurisdictions, government action is often subject to legal constraints under constitutional law, administrative law, and human rights law, which private actors are not. This may allow public actors to divest themselves of their legal obligations. For example, under EU law, the actions of public institutions may be challenged on the basis of the principles of proportionality and non-discrimination, where private institutions may not. In the USA and the UK, the common law principles of judicial review grant the courts a jurisdiction to review the actions of government decision-makers against standards of unfairness,

transparency, bias, and rationality (Taggart 1997), but common law judicial review does not apply to private actors. The principle of bias, for example, forms part of the natural law principles of procedural fairness at common law. It states that where a decision-maker has a pecuniary interest in a decision, the decision may be void for bias. Yet for-profit private actors (for example, particularly those responsible for social welfare or prison systems) will nearly always have a pecuniary interest in the outcome of decisions which it has been delegated.

In civil law countries as well as common law countries, public law sets out detailed codes on administrative procedure and decision-making, but private law remains a matter of contract. This makes courts and legislators slow to recognize and enforce public law duties on private parties in most jurisdictions, even when those private parties are wielding governmental power (Donnelly 2007). Even under the most effective international human rights regimes, such as the European Convention on Human Rights, individual human rights are enforceable only against the state save in exceptional or indirect cases. Such regimes do not typically apply to the actions of private parties, despite the fact that private parties exercising state power may enjoy the same capacity to interfere with individual rights.

Political Criticisms

Political opposition derives primarily from five outcomes: (1) Private restructuring of labor markets and labor retrenchments may result in worker vulnerability and increased unemployment; (2) privatization may result in the discontinuation of less-profitable services; (3) consumer cost increases may result from the withdrawal of government support or private market imperfections; (4) governments may lose control of outcomes in which there is a strong public interest; and (5) privatization may sacrifice the long-term value of assets for short-term gain, because while there are inevitably a finite amount of public assets, there is an infinite capacity for debt growth.

Political opposition therefore derives from both empirical and ideological observation. Underlying each of these criticisms is the

implication that the treatment of citizens as consumers is often inappropriate. First, at an ideological level, citizenship may imply something more than consumerism – it implies a stake in the state itself. For example, as citizen-consumers, the electorate may accept the privatization of postal services provided that efficiency gains are demonstrated, but the electorate as citizens may not accept the privatization of health or military services, regardless of efficiency. The National Health Service in the UK is an example of a public asset that is unlikely to be privatized due to strong public support. Second, at an empirical level, consumers exercise their influence on the state through the exercise of choice. Yet when it comes to public services, citizens will not often have a choice. Welfare recipients or those needing immediate health care, for example, may not have a choice as to the most efficient provider.

References

- Brudney JL, Fernandez S, Ryu JE, Deil W (2004) Exploring and explaining contracting out: patterns among the American States. *J Public Adm Res Theory* 15:393
- Buchanan JM (1972) *Theory of public choice*. University of Michigan Press, Ann Arbor
- Coase R (1960) The problem of social cost. *J Law Econ* 2(1):1–44
- Demsetz H (1968) Why regulate utilities? *J Law Econ* 11(1):55
- Demsetz H (1969) Information and efficiency: another viewpoint. *J Law Econ* 12(1):1
- Donnelly C (2007) *Delegation of governmental power to private parties*. Oxford University Press, New York
- Dunsire A, Hartley K, Parker D, Dimitriou B (1988) Organisational status and performance: a conceptual framework for testing public choice theories. *Public Adm* 66(4):363
- Fraja D (1993) Productive efficiency in public and private firms. *J Law Econ* 50:15
- Freedland M (1994) *Government by contract and public law*. *Public Law* 86:101
- Hughes OW (2003) *Public management and administration: an introduction*. Palgrave Macmillan, Basingstoke
- Kirzner IM (1997) *How markets work: disequilibrium, entrepreneurship and discovery*. Hobart Paper 113, Institute of Economic Affairs, London
- Leonard CS, Pitt-Watson D (2013) *Privatization and transition in Russia in the early 1990s*. Routledge, New York
- Lipton D, Sachs J (1990) *Privatization in Eastern Europe: The case of Poland*. *Brook Pap Econ Act* 2:293

- Littlechild SC (1986) The fallacy of the mixed economy: an 'Austrian' critique of recent Economic thinking and policy. Hobart Paper 90, Institute of Economic Affairs, London
- Meggison W (2005) The financial economics of privatization. Oxford University Press, Oxford, pp 52–96
- Milward R, Parker D (1983) Public and private enterprise: comparative behaviour and relative efficiency. In: Milward D, Parker L, Rosenthal MT, Topham N (eds) Public sector economics. Longman, London
- Munkholt P (2000) Russia seeking a soft landing. *Privat Int* 137:7
- Niskanen WA (1971) Bureaucracy and representative government. Aldine, Chicago
- OECD (2000) Privatisation, competition and regulation. OECD, Paris, pp 10–11
- Parker D (1998) Privatization in the European Union: theory and policy perspectives. Routledge, London
- Parker D (2000) Privatization and corporate performance. Edward Elgar Publishing, Cheltenham, pp xii–xv
- Pint E (1991) Nationalization vs regulation of monopolies: the effects of ownership on efficiency. *J Public Eco* 44(2):131
- Sappington D, Stiglitz J (1987) Privatization, information and incentives. *J Pol Anal Manag* 6(4):110–125
- Stiglitz J (1995) Whither socialism? MIT Press, Cambridge
- Taggart M (1997) The province of administrative law determined? In: Taggart M (ed) The province of administrative law. Hart, Oxford, pp 1–3
- Vickers J, Yarrow G (1988) Privatisation: an economic analysis. MIT Press, Cambridge, MA
- Wang S (1994) The compatibility of public ownership and the market economy: a great debate in China. *World Affairs* 157(1):38
- World Bank (1995) Bureaucrats in business: the economics and politics of government ownership. Oxford University Press, Oxford

Product Promotion Strategy

► Promotional Effort

Productivity and Growth

Jaime Vallés-Giménez and Anabel Zárate-Marco
University of Zaragoza, Zaragoza, Spain

Abstract

Economic growth is a long-run process that occurs when an economy's potential output increases, and it can be measured by the

product method, the income approach or the expenditure method. Actual growth is the percentage annual increase in national output. Potential growth is the speed at which the economy could grow, i.e. the percentage annual increase in the economy's capacity to produce and it can be shown by an outward shift in the economy's production possibility frontier. The theory and empirical studies suggest that potential economic growth is associated with the increase in the use of factors of production (capital, labor, energy, etc), but primarily to increases in productivity or efficiency with which these factors are used, through advances in labor skills and organization of production or improves in technology. Productivity and economic growth are then closely linked because economic growth occurs when productivity increases to allow for such growth. Productivity is therefore the cornerstone of economic growth.

Productivity is an indicator of the efficiency of production and can be defined as the ratio of output to inputs in production. Higher productivity means that the economy can produce more goods and services at a lower cost per unit. This will help to reduce prices and increase consumer welfare and living standards, because more real income improves people's ability to purchase goods and services, enjoy leisure, improve housing and education and contribute to social and environmental programs. Higher productivity increase total output from the scarce factor resources, causing an outward shift of the production possibility frontier. Productivity also affects our competitive position: the more productive we are the better we are able to compete on world markets. Productivity growth also helps businesses to be more profitable. There are broadly two ways of measuring productivity. On one side are the partial productivity measurements that relate to an input (labor, capital, etc), and on the other side it is a measure of total factor productivity (TFP) or multifactor productivity (MFP), which measures the effects in total output not caused by measured inputs of labor, capital and intermediate outputs.

Definitions

Economic growth is a long-run process that occurs when an economy's potential output increases, i.e., as economy's ability to produce goods and services rises.

Productivity is an indicator of the efficiency of production.

Concepts and Connections Between Growth and Productivity

The fruit of economic activity is the amount of goods and services produced by labor, capital, and other inputs. So arguably, **economic growth** is the increase in the market value of the goods and services produced by an economy over time. It is conventionally measured as the percent rate of increase in real gross domestic product, or real GDP, i.e., $(\text{real GDP}_t - \text{real GDP}_{t-1}) / \text{real GDP}_{t-1}$.

Since GDP is a macroeconomic variable, i.e., it is the result of multiplying the quantities of goods and services produced in one country by their prices, we will only have a proper idea of the growth of an economy's production if we eliminate the distorting effect of inflation on the price of goods produced, and the evolution of real output is analyzed.

Another element to consider in economic growth is the increase in population. Only if the population increase is known can it be determined whether the per capita product increases or not. For this reason, of more importance is the growth of the ratio of real GDP to population (real GDP per capita).

On the other hand, as the GDP figures of each country are measured in its local currency, they have to be converted into a common currency (e.g., dollars or euros) at the current exchange rate, to be able to be compared. But the exchange rate may be a poor indicator of the purchasing power of the currency at home. To compensate for this, GDP can be converted into a common currency at a *purchasing-power parity rate*. This is a rate of exchange that would allow a given amount of money in one country to buy the same amount of goods in another country after

exchanging it into the currency of the other country.

Besides this method of measuring GDP, known as the *product method*, GDP can be calculated in other two different ways. The production of goods and services generates incomes for households in the form of wages and salaries, profits, rent, and interest. Therefore, GDP can also be calculated by adding up all of the income received by labor and other inputs in the economy. This is known as the *income approach*. The third method, expenditure method, focuses on the expenditures necessary to purchase the nation's production by different groups in the economy. The four main components are consumption expenditures by households, gross private investment spending principally by firms, government purchases of goods and services, and net exports (exports minus imports). Because of the way the calculations are made, the three methods of calculating GDP must yield the same result.

It is essential to distinguish between actual and potential economic growth. Actual growth is the percentage annual increase in national output: the rate of growth in actual output (GDP). Potential growth is the speed at which the economy could grow. It is the percentage annual increase in the economy's capacity to produce: the rate of growth in potential output. An increase in an economy's productive potential can be shown by an outward shift in the economy's production possibility frontier.

If the potential growth rate exceeds the actual growth rate, there will be an increase in spare capacity and probably an increase in unemployment: there will be a growing gap between potential and actual output. To close this gap, the actual growth rate would temporarily have to exceed the potential growth rate. In the long run, however, the actual growth rate will be limited to the potential growth rate. Although growth in potential output varies to some extent over the years – depending on the rate of advance of technology, the level of investment, and the discovery of new raw materials – it nevertheless tends to be much more steady than the growth in actual output. Actual growth tends to fluctuate. In some years, countries will experience high rates of

economic growth. In other years, economic growth is low or even negative. This cycle of booms and recessions is known as the business cycle or trade cycle. The business cycle moves up and down, creating fluctuations around the long-run trend in economic growth.

Anyway, these measures of growth do not determine economic development as this is a more complex concept as it has social connotations also related to the improvements of living standards in the country. They ignore the distribution of income. Although per capita real income may be increasing in a country, this does not necessarily mean that all the inhabitants of that country are benefiting from this improvement. Maybe along with the growth of real income it takes place a change of income that impoverish certain people while others enjoy her growth above average. While economic growth is necessary, it is not sufficient for progress on reducing poverty. For this reason, the economic growth may be a poor indicator of society's well-being.

Moreover, it does not compute the human costs of production. If production increases, this may be due to technological advance. If, however, it increases as a result of people having to work harder or longer hours, its net benefit will be less. Leisure is a desirable good and so too are pleasant working conditions, but these items are not included in the economic growth figures.

Another problem in the computation of economic growth is some external benefits or costs that are not included in GDP statistics. For example, growth has the disadvantage that can both create negative externalities, e.g., higher levels of noise pollution and lower air quality arising from air pollution and road congestion and causes depletion of resources. And finally, GDP only measures the market economy, thereby excluding "do-it-yourself" and other home-based activities as well as the underground economy, so the GDP statistics understate the true level of production in the economy.

Since the pioneering work of Solow (1956), the theory and empirical studies suggest that potential economic growth is associated with two factors: the increase in the use of factors of production (capital, labor, energy, etc.), but

primarily the increase in productivity or efficiency with which these factors are used, through advances in labor skills and organization of production or improvements in technology. Productivity and economic growth are then closely linked because economic growth occurs when productivity increases to allow for such growth. Productivity is therefore the cornerstone of economic growth.

Productivity can be defined as the ratio of output to inputs in production. It is an average measure of how efficiently goods and services are produced. Higher productivity means that the economy can produce more goods and services at a lower cost per unit. This will help to reduce prices and increase consumer welfare and living standards, because more real income improves people's ability to purchase goods and services, enjoy leisure, improve housing and education, and contribute to social and environmental programs. Higher productivity increase total output from the scarce factor resources, causing an outward shift of the production possibility frontier. Productivity also affects our competitive position: the more productive we are the better we are able to compete on world markets. Productivity growth also helps businesses to be more profitable.

There are broadly two ways of measuring productivity. On one side are the partial productivity measurements that relate to an input (labor, capital, etc.), so we can say that there are so many measurements of productivity as resources used in production. This partial productivity has usually been measured in terms of labor, by the availability of data. Labor productivity is then the value of goods and services produced in a period of time, divided by the hours of labor used to produce them. In other words, the labor productivity measure the output produced per unit of labor, usually reported as output per hour worked or output per employed person.

However, partial productivities do not show overall efficiency of the use of all the resources, so it is important to have a simultaneous measurement of the efficiency of the totality of resources, i.e., a measure of total factor productivity (TFP) or multifactor productivity (MFP). TFP measures

the effects in total output not caused by measured inputs of labor, capital, and intermediate outputs. If all inputs are accounted for, total factor productivity (TFP) represents improvements in ways of doing things, that is to say, it can be taken as a measure of an economy's long-term technological change or technological dynamism, which is the primary source of real economic growth. In the short term, however, also reflect unexplained factors such as cyclical variations in labor and capital utilization, economies of scale, and measurement error. Total factor productivity is the most commonly known and widely used method of productivity measurement. However, TFP cannot be measured directly but is a residual. It accounts the residual growth that cannot be explained by the rate of change in the inputs.

Sources of Productivity and Economic Growth

Despite the lack of a unifying theory, there are several partial theories that discuss the role of various factors in determining long-term economic growth. Two main strands can be distinguished. The neoclassical model based on the growth model of Solow (1956) has emphasized the importance of investment. And the more recent theory of endogenous growth developed by Romer (1986, 1990) and Lucas (1988) has drawn attention to human capital and innovation capacity. Furthermore, important contributions on economic growth have been provided by the cumulative causation theory of Myrdal (1957) and by the New Economic Geography school of Krugman (1991). These two schools assert that economic growth tends to be an unbalance process favoring the initially advantaged economies. In addition, other explanations have highlighted the significant role that non-economic (in the conventional sense) factors play on economic performance.

The theoretical and empirical studies about the growth theories have been plentiful and have used differing conceptual and methodological viewpoints (especially interesting is the review of Helpman 2004). These studies have placed

emphasis on different explanatory parameters and have offered various insights to the sources of economic growth. The review of this literature suggests that no single policy or factor leads to productivity growth. Rather, it is a matter of getting a lot of interconnecting things right, and by ensuring incentives are aligned, creating an environment where firms can create and take advantage of opportunities. The main factors that may influence the long-term economic growth are as follows:

1. *Physical capital.* The rate of accumulation of physical capital is one of the main factors determining the level of real output per capita although its effects could be more or less permanent depending on the extent to which technological innovation is embodied in new capital. Investment in physical capital is a key determinant of economic growth identified by both neoclassical and endogenous growth models. However, in the neoclassical model, physical capital has impact on the transitional period, while the endogenous growth models argue for more permanent effects. Whatever the transition mechanism from capital accumulation to growth, the significant differences in the investment rate across countries, and over time, point to it as a possible source of cross-country differences in output per capita.

Physical capital includes factories, tools, computers, machinery, production equipment, and structures such as infrastructure or fixed social capital that are often the result of investments made by the state. These infrastructures range from transport infrastructure (roads, airports, ports, railways), energy, telecommunications, to universities, hospitals, water projects, and other public health measures, e.g., diseases control, etc.

The more capital workers have at their disposal, generally the better they are able to do their jobs, producing more and better quality output. The role of infrastructures is to expand production, to increase resources, and to enhance the productivity of private capital. A good transport system enhances

cohesion and improves access to outlying regions through the reduction of transport costs for both goods and people traveling for leisure or work. Telecommunications are the modern substitute for the connections made through transport and a prerequisite for the development of industries and modern services that rely on the phone, fax, and data transmission systems. An undersized or inadequate infrastructure, such as an electrical network with frequent failures, cuts electricity to homes and businesses is a major obstacle to economic growth in some countries. Without necessary infrastructure, it can be difficult for firms to be competitive in the international markets. The lack of infrastructure is often a factor holding back some developing economies

2. It is not enough that a worker has good equipment; he must also know what to do with it and how to use it in the safest, most effective, and efficient manner. For this reason, *human capital* is the main source of growth in several endogenous growth models as well as one of the key extensions of the neoclassical growth model, because of its role as a facilitator of both technology adoption from abroad (absorption capacity) and the creation of appropriate domestic technology (innovation). It would be impossible to operate the current economy with a population with the literacy levels and formation of a century ago.

The term human capital refers both to the improvement and training of manpower produced by the education and knowledge that is incorporated into the work force and by the learning by doing, as well as to the improvements in their health. A population that is well educated and well trained helps a society to increase its ability and acquire as well as use relevant knowledge (absorption capacity). As knowledge is created by a small number of leader countries in technological terms and most countries do not produce state-of-the-art technology themselves, these latter countries must acquire the technology from elsewhere via trade or foreign direct investment.

Focusing on education, basic education would be important for learning-capacity and utilizing information, while the higher education would be necessary for technological innovation.

In turn, we must bear in mind that in a population with good health, workers can be more productive and learning ability of children to be greater. A longer life expectancy makes it more attractive to invest in human capital and even foreign direct investment, and savings incentive and productivity can be increased. The governments can also play an important role in the accumulation of human capital from the time they can invest in education.

3. Probably the most important factor for productivity growth and therefore economic growth is the *innovation and technological progress*. In practice, most technological improvements are due to deliberate actions, such as research and development (R&D) carried out in research institute or firms.

R&D evolves new ideas and designs and is used by firms in search for blueprints of new varieties of products or higher-quality products. New ideas can also take the form of new technologies, new products, or new corporate structures and ways of working. Expenditure on R&D can be considered as an investment in knowledge that translates into new technologies as well as more efficient ways of using existing resources. Such innovations contribute to the expansion of the so-called frontiers of knowledge, and the accumulation of knowledge will generate growth. Hence, technological change emerges from technical innovations generated by research and development, patenting and software, and productivity enhancing developments in the fields of education management and marketing. And workers today are capable of producing more than in the past, even with the same amount of physical and human capital, because the technology has advanced over time.

The amount of resources that are devoted to R&D can be influenced by government

intervention. In particular, the potential benefits from new ideas may not be fully appropriated by the innovators themselves due to spillover effects, which imply that without policy intervention the private sector would likely engage in less R&D than what could be socially optimal. This can justify some government involvement in R&D, both through direct provision and funding, but also through indirect measures such as tax incentives and protection of intellectual property rights to encourage private-sector R&D.

4. The degree of *openness* of the economy also affects productivity and economic growth. It does through several channels such as exploitation of comparative advantage, technology transfer and diffusion of knowledge, increasing scale economies, and exposure to competition. Trade liberalization promotes competitiveness, efficiency of input, creates incentives to innovate, and ensures that resources are allocated to the most efficient firms. It also forces existing firms to organize work more effectively through imitations of organizational structures and allow to introduce foreign (relatively advanced) technology into domestic production, which in turn has a positive effect on productivity and economic growth. To have a good absorption capacity is a key factor for a good use of the technology transferred. In particular, certain kinds of imports, namely, machinery and equipment relating to foreign R&D, are expected to generate a lot of technology transfer because the technology is often embodied in goods. Moreover, trade liberalization increases investment opportunities and international contacts.
5. *Foreign Direct Investment* can play a crucial role of internationalizing economic activity and can stimulate economic growth by improving technology and productivity. Host economies are expected to benefit from the positive externalities driven by foreign direct investment. Those include knowledge spillovers generated by technology transfers, introduction of new processes, and

managerial skills and know-how diffusion to the domestic market.

6. *Macroeconomic stability-oriented policies* can also have a significant impact on economic growth. A stable macroeconomic environment characterized by low and predictable inflation, sustainable budget deficits, and limited departure of the real exchange rate from its equilibrium level, sends important signals to the private sector about the commitment and credibility of a country's authorities to efficiently manage their economy and increase the opportunity set of profitable investments.

The usual arguments for lower and more stable inflation rates include reduced uncertainty in the economy and enhanced efficiency of the price mechanism. Uncertainty related to higher volatility in inflation could discourage firms from investing in projects that have high returns but also a higher inherent degree of risk. Moreover, large fiscal deficits and high net international debt position make a country vulnerable to global financial shocks and terms of trade shocks (e.g., oil price spikes)

7. *Institutions* also matter for economic growth because they establish the rules of the game in a society or, more formally, they set the humanly devised constraints that shape human interaction (North 1990). Institutions decide how to organize the societies and so determine whether or not the economy prospers. Certain forms societies encourage people to innovate, to take risks, to save for the future, to find better ways of doing things, to learn and educate themselves, solve problems of collective action, and provide public goods, while others do not. Institutions are therefore a key factor for potential economic growth.

Institutions encompass both informal constraints as customs, traditions, codes, or taboos and formal constraints as property rights, legal rules, constitutions, contract enforcement, the political system, and electoral rules. Also encompass public sector imperfections, the degree to which laws and

regulations is fairly applied, the extent of corruption, etc.

Economic institutions are the set of norms relating to production, allocation, and distribution process of goods and services. They can guarantee the rights of intellectual and industrial property so necessary for there to be greater efficiency in the use of resources and also greater technological progress and innovation leading to economic growth. They must also guarantee some degree of equality of opportunity in society, including such things as equality before the law, so that those with good investment opportunities can take advantage of them. Economic institutions also help to stabilize the economy and ensure the proper functioning of the financial system, so it is necessary to have high rates of savings and investment, a good allocation of resources, specialization of the economy, and creating incentives. A well-developed financial system provides funding for capital accumulation, helps the diffusion of new technologies, mobilizes savings by channeling small savings of individuals into profitable large-scale investments, while offering savers a high degree of liquidity.

Economic institutions can also remove or reduce the tariff barriers and the impediments to foreign investment so they may favor a greater exposure to international competition. They may enforce contracts, discourage unfair or abusive business practices, limit the power of rulers, protect individuals both from one another and from the state, increase safety at work, contribute to public health and safety and the development of a more productive and fairer society, set limits on pollution, help protect consumers from potentially hazardous products, ensure that they are able to make informed choices, and influence investments in physical and human capital and technology and the organization of production.

Economic institutions are linked to political institutions because the latter are necessary to the former work. Political institutions

are those that determine the structure of the state and the procedures of the political decision-making process. Political institutions shape the political process that produces legislation and regulation. They also determine the legal system and coordinate the processes that create and enforce the law. Political institutions therefore produce economic institutions and determine their quality. Institutions like democracy and social protection legitimize market outcomes and ensure their endurance. Political institutions can support a market economy by shaping and safeguarding property rights and making the market compatible with social stability and social cohesion. A stable and corruption-free government, a strong independent judiciary, efficient bureaucracy, and political constraint on executive are keys to generate certainty and thus economic growth, since political instability would increase uncertainty, discouraging investment, and eventually hindering economic growth.

8. Certain structural characteristics such as *geographical conditions* are also a powerful driver of economic growth. The country's location, its topography, and access to the sea affect the transport costs and the efficient allocation of resources because where geography is not propitious, the diffusion of technology is more complicated and more costly trade. Thus, it also affects the productivity and competitiveness.

Moreover, the existence of natural resources in abundance is essential for economic growth. The natural resources of a country such as minerals and oil-resources, soil quality, forest wealth, and good climate river system affect the returns to agriculture and its economic structure. A country deficient in natural resources may not be in a position to develop rapidly, although natural resources are a condition for economic growth necessary but not sufficient to one.

9. *Demographic factors* like population growth, population density, migration, and age distribution can play a major role in economic growth.

The population growth can undermine per capita economic growth, and moreover, composition of the population has important implications for growth. A large working-age population is deemed to be conducive to growth, whereas population with many young and elderly dependents is seen as impediment because influences the dependency ratio, investment and saving behavior, and quality of human capital. Population density, in turn, may be positively linked with economic growth as a result of increased specialization, knowledge diffusion, and so on. Migration would affect growth potential of both the sending and receiving countries.

10. Other factors of *sociocultural nature*, e.g., ethnic diversity or cultural diversity may affect growth although in a much more residual, indirect, and unclear manner. For instance, it may have a negative impact on growth due to emergence of social uncertainty or even of social conflicts or a positive effect since it may give rise to a pluralistic environment where cooperation can flourish.

References

- Helpman E (2004) The mystery of economic growth. The Belknap Press of Harvard University Press, Cambridge, MA
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99:183–199
- Lucas R (1988) On the mechanics of economic development. *J Monet Econ* 22:3–42
- Myrdal G (1957) Economic theory and underdeveloped regions. Hutchinson Publications, London
- North DC (1990) Institutions, institutional change, and economic performance. Cambridge University Press, New York
- Romer P (1986) Increasing returns and long run growth. *J Polit Econ* 94(2):1002–1037
- Romer P (1990) Endogenous technological change. *J Polit Econ* 98(5):S71–S102
- Solow RM (1956) A contribution to the theory of economic growth. *Q J Econ* 70(1):65–94

Professional Guides

- [Codes of Conduct](#)

Prohibition

Nicholas A. Snow

Department of Economics, Indiana University, Bloomington, IN, USA

Abstract

A prohibition is a government decree against the production and exchange of a good or service. Recent studies on prohibitions, for a variety of goods and services, such as drugs, alcohol, and prostitution, suggest that prohibitions impose heavy costs and are extremely difficult to enforce.

Prohibition

Governments throughout history have utilized prohibitions on a multitude of goods and services throughout history for a variety of reasons, such as attempts to protect domestic producers from foreign trade, protect consumers, regulate morality, etc. Thus, a prohibition is a means to achieving a broader social end.

In theory, the economics of prohibition is simple. Ultimately, prohibitions are a supply reduction legislation designed to curtail the production, exchange, and consumption of a good through the use of penalties, such as fines, confiscation of assets, and even jail sentences. Prohibitions attempt to reduce supply by making it more difficult for suppliers to produce and distribute the good in question. Within the supply and demand model, this causes a shift of the supply curve up and to the left resulting in a higher price and decrease in quantity demanded.

If enforcement is successful, this results in lower consumption, which has several welfare implications. First, it leads to a loss for consumers resulting from the higher price and the substitution of consumption to lower-valued goods. Second, producers experience a loss from the higher production costs and risks or in a loss of income and utility in needing to shift to other occupations, which differ from their comparative advantage.

Finally, a further utility loss occurs for both consumers and producers because the decrease consumption creates an allocative inefficiency due to the lost gains from trade or what economists call a deadweight loss.

The level of prohibition enforcement is also determined by economic factors. In enforcing prohibitions the government faces scarcity. Thus, enforcement is unlikely to be increased until the goal of eliminating the supply is reached. Just as individuals must equate marginal costs with marginal benefits in decision-making, so too must policy makers. Enforcement is not costless and requires the use of resources. Every dollar spent on prohibition enforcement is a dollar not spent on some other policy or a dollar not spent on something else by the taxpayer. Thus, the law of diminishing marginal utility exists in the policy world as well as for the individual. Policy makers must find the optimal level of enforcement, which is unlikely to lead to zero consumption. Enforcement will lower consumption but whether or not this leads to a level of consumption that achieves the policy goal is determined by the opportunity costs perceived by the individuals who make up society.

From a policy perspective, the success of the prohibition should not stem from the evidence of reduced consumption but rather whether the reduced consumption leads to the desired outcomes or not. A further difficulty in regard to the enforcement of a prohibition is the dynamic effects, or the unintended consequences, of the law. First, in pushing the trade underground information is greatly restricted due to the necessary premium on secrecy, which in turn greatly reduces price competition. This leads many in the market to charge monopoly prices, which widens profit margins. These large profits create an incentive for many, normally quite law abiding, individuals to break the law. This will include not only those directly participating in the trade but also government officials who succumb to the temptation of corruption. This has an effect of increasing both supply and demand, which attracts the attention of government officials, who then attempts to crack down. As a result the amateur is pushed out of the market but with the continued profit opportunities

for individuals who are skilled at evading the law remain.

Second, enforcement can also lead to consequences concerning the good itself, such as a lower quality. With fewer competitors, suppliers do not need to rely on quality as much as they would in a legal market. The principle consequence, however, is what Richard Cowan, in a 1986 article, called the “iron law of prohibition,” which is “the more intense the law enforcement, the more potent the drugs become.” Cowan’s point was related to drug prohibition but the principle stands for other goods as well. There are two reasons for this. First, the prohibition makes smuggling a necessary activity in order to get the goods to the consumers, and smugglers will prefer to minimize the bulk of their goods. For example, typically a shot of liquor, glass of beer, and a glass of wine have roughly the same amount of alcohol. Thus, due to its smaller size, liquor becomes more profitable to smuggle, driving beer and wine out of the market. Second, from the demand side, the same considerations apply. Liquor is more potent than beer and wine; it takes less of the good to do the job, of getting drunk in this example, for the consumer. When taking a hip flask with you for a night on the town filling it with whiskey is more effective than filling it with your favorite Pinot Noir.

These unintended consequences of prohibition are a causal result of the relationship between the market process and the intervention into that process. While these consequences are not completely predictable in detail, they are not completely surprising either. Interventions into the market’s discovery process alter that discovery process; it does not eliminate it. Economist Israel Kirzner categorizes the four different ways interventions can broadly alter the market’s discovery process, namely, by creating an undiscovered discovery process, an unsimulated discovery process, a stifled discovery process, and a wholly superfluous discovery process. All four of these are important for an intervention like a prohibition.

The undiscovered discovery process refers to the demand for government intervention due to either, or both, an ignorance or impatience with the market’s ability to achieve the desired end or

state of affairs. The market is a process not an instantaneous state of affairs. Further markets operate under imperfect knowledge. Thus, corrections take time and occur under a decentralized spontaneous process. Prohibitions are often implemented due to perceived market failures, but often these failures are a result of other government interventions that impaired the market process in the first place. In other words, the undiscovered discovery process matters because government interventions are further demanded because imperfections of interventions are not seen or are blamed on the market and/or individuals are ignorant and/or impatient with the markets progress toward a solution.

The unsimulated discovery process refers to the inability of the bureaucracies in charge of a regulation to simulate the market's discovery process. Prohibition enforcement is made difficult because the bureaucracies in charge of enforcement are unable to simulate the discovery, or success, of the market. Bureaucrats have little incentive to innovate in order to improve efficiency within their tasks because they lack the incentive structure created by a competitive environment. As a result enforcement tends to be inefficient relative to what the market could or would produce. And this obviously creates a difficult hurdle for enforcement.

Interventions into the market also help to stifle the market's discovery process. Innovations in new techniques, safety, product characteristics, sources of supply, etc. created through this process are no longer discovered. In the case of prohibitions, the discovery process can be completely destroyed and also curtail and distort the process for other goods, as innovations in one area may have implications in others. And understanding the magnitude of this is simply not possible.

Finally, the wholly superfluous discovery process is the creation of new economic profit opportunities in response to the government's intervention. These new discoveries in search of profits are not always desirable. The new discovery process often creates wholly unexpected and undesired outcomes that create tremendous change in terms of both the product and the supply chain. Essentially, the government's attempts at

enforcement are undercut as entrepreneurs figure ways to supply the good to consumers, and the means they use may even be worse than what the regulation is trying to fix. As economist Milton Friedman, in his 1975 book *An Economist's Protest*, put it, "There is ample evidence that imagination and innovation are not stilled by restrictive legislation – only diverted to figuring ways around it."

The altered profit opportunities not only affect market participants but also government officials responsible for enforcing the prohibition. Since the bureaucrats are not the residual claimant of the bureaucracy they are a part of, black markets provide chances to gain by providing market participants either protection or selective enforcement. In other words, a problem of graft and corruption within their own ranks will further impede the government's efforts to enforce the law, and this is a direct result of the wholly superfluous discovery process.

While not all prohibitions will be the same, the economic theory of prohibition does tend to illustrate difficulties with the enforcement of restricting market activities for many goods and services.

Cross-References

► [Alcohol Prohibition](#)

References

- Cowan R (1986) How the narcs created crack: a war against ourselves. *Natl Rev* 38(23):26–34
- Friedman M (1975) *An economist's protest*. Thomas Horton and Daughters, Glen Ridge

Further Reading

- Becker GS, Murphy KM, Grossman M (2006) The market for illegal goods: the case of drugs. *J Polit Econ* 114(1):38–60
- Boettke PJ, Coyne CJ, Hall AR (2013) Keep off the grass: the economics of prohibition and U.S. Drug Policy. *Oregon Law Rev* 17(4):1069–1095
- Miron JA (2004) *Drug war crimes: the consequences of prohibition*. Independent Institute, Oakland
- Thornton M (1991) *The economics of prohibition*. University of Utah Press, Salt Lake City

Promotional Activities

► Promotional Effort

Promotional Effort

Nikolaos Georgantzis^{1,3} and Christian Boris Brunner²

¹Economic and Social Sciences, School of Agriculture Policy and Development, University of Reading, Reading, Berkshire, UK

²School of Agriculture, Policy and Development, University of Reading, Reading, Berkshire, UK

³LEE and Economics Department, Universitat Jaume I, Castellon, Spain

Abstract

The term promotional effort refers to all strategies aimed at broadening a firm's market scope through the establishment of a larger and more loyal consumer basis. Advertising, public relations, sales promotion, personal selling as well as price-related strategies affecting a firm's sales potential are addressed. Both positive and normative approaches are briefly reviewed, discussing the theoretical and empirical issues studied in the existing literature.

Synonyms

Product promotion strategy; Promotional activities

Definition

The term is used to refer to the qualitative and quantitative aspects of a firm's strategies aimed at broadening its market scope through the establishment of a larger and more loyal consumer basis. According to Kotler et al. (2013), such activities can generally be classified into product management, pricing, promotion, and distribution. Promotional activities include advertising, public relations, sales promotion, personal selling as

well as database marketing, direct response marketing, sponsoring, social media, and other alternative marketing activities (Clow and Baack 2014). In formal economic models, promotional effort is treated separately from pricing, in which case it refers to investments enhancing a firm's potential market before pricing is taken into account. However, price-related strategies like price announcements, bundle pricing, or low price guarantees could be considered as part of a firm's promotional effort, rather than merely a pricing decision.

Impact Measurement and Responses to Promotional Effort

The measurement of a firm's promotional effort is a challenge for marketers. First, a problem arises due to the difficulty in identifying the costs specific to different activities (e.g., Chapman 1986). Second, a problem arises with the measurement of a direct causal relationship between a given promotional strategy and its outcome (Berger et al. 1964), like, for example, consumers' reactions such as their attitude toward the brand or the sales rate of products. Third, this relationship is moderated and/or mediated by unknown and uncontrolled factors (Kuehn 1964; Berger et al. 1964). Fourth, the outcome of promotional effort is delayed, which causes the problem of accountability over time between promotional effort and outcomes (e.g., Mills 1959; Miller and Strain 1970). Finally, multiple promotional activities of a firm may support or weaken each other.

Similar problems arise with respect to various products of a multiproduct firm. To overcome the complexity of the resulting setup, several experimental studies have been performed in order to isolate promotional effort effects within each separate domain (e.g., Berger et al. 1964; Miller and Strain 1970). To find the optimum promotional effort of a firm, economic models have taken into account the multiperiod, multicompetitor, and multiproduct nature of the problem (e.g., Gupta and Krishnan 1967a, b). Furthermore, the price of the firm's product(s) (e.g., Carpenter 1987; Krishnan and Gupta 1967) as well as a firm's price changes (price reductions or price

promotions) have been analyzed more in detail, including consumer switching or variety-seeking behavior (Kahn and Raju 1991). In order to be effective and efficient with respect to its promotional activities, a firm needs to: (1) *choose* the *right* promotional activities, to which its target group pays attention and responds positively. In addition to this, the selected promotional activities must fit to the firm's positioning and image and differentiate the firm from competitor; (2) *execute* the selected promotional activities in an efficient and effective way; and (3) *integrate* all promotional activities to one "main picture" in order to create a unique brand image in the consumer's mind.

Qualitative and Quantitative Aspects of Promotional Effort

From a managerial point of view, qualitative and quantitative considerations define the two major domains along which promotional effort enters into a firm's decision-making process. Qualitative considerations regard the choice of the actual strategy mix and the nature of explicit or implicit messages and signals transmitted to the consumer, whereas quantitative considerations regard the firm's investment in the different promotional activities. Along the qualitative domain, a firm's promotional effort could aim at creating or reinforcing the consumer's or stakeholder's brand awareness, stimulating their interest in the firm's products, encouraging first or more frequent purchases of the firm's products and building a long-term relationship (Bester and Petrakis 1996; Moraga-Gonzalez and Petrakis 1999) between the firm and its customers (Clow and Baack 2014; Keller 2013). The marketing literature emphasizes the need for an integrated strategy in order for promotional effort to lead to a unified image of the firm and its brands in the consumer's mind (Keller 1993, 2013), so that all promotional activities should support each other, transferring the same meaning to the receiver (Clow and Baack 2014). The importance of these qualitative aspects of promotional effort in a firm's decision-making problem has contributed to the fact that the issue has attracted researchers from many different

disciplines, like economics, management, psychology, sociology, neuroscience, media and communication science, and even sociolinguistics.

Thus, the use of promotion techniques is informed by all the aforementioned approaches, pointing clearly to a possibility and desire of firms to intervene and affect the consumer's decision-making context and overall attitude, beyond pure product-related information, so that along all touch-points between customer and brand, the firm needs to communicate the same information to create a strong brand image in the customer's mind and to reinforce customer's brand knowledge (Keller 2013; Rossiter and Bellman 2005). Despite that, mainstream approaches within both the economics and marketing disciplines have strongly insisted on the information-enhancing role of advertising and promotional effort in general, adopting a preference invariance approach, similar to what could be motivated by Stigler and Becker's (1977) *De gustibus non est disputandum* (see also Becker 1996; Becker and Murphy 1993). As a consequence, all relevant legislations are almost exclusively concerned with the truthfulness of messages contained in informational campaigns, rather than with the persuasive effects of advertising, aimed at the subconscious processes underlying the consumer's choices (except for a clear ban of advertising aimed at trapping children's preferences).

Regarding the channel through which promotional effort broadens a firm's market potential, promotional effort may target retailers and wholesalers or end consumers. In the former case, the firm uses a *push* strategy, offering incentives directly to retailers and wholesalers (e.g., through personal selling or promotion). The aim is that retailers and wholesalers are the ones who engage in specific marketing activities such as personal selling, sales promotion, or advertising directly to the end consumers (Kotler et al. 2013; Barreda and Georgantzis 2002). In the latter case, the firm uses a *pull* strategy, promoting its brands and products directly to end consumers, for example, through TV spots, print advertisements, mobile marketing, and mailings.

Finally, the quantitative approach to promotional effort focuses on the issue of advertising expenditure and its efficiency as a market-

enhancing mechanism. Both types of advertising, informative (Grossman and Shapiro 1984) and persuasive (Bloch and Manceau 1999; von der Fehr and Stevik 1998) have been studied in oligopolistic contexts to reach a rather robust conclusion that advertising expenditures may be excessive from a social point of view and may lead firms to a prisoner dilemma-type of situation in which firms are trapped into “run-to-stay-still” competition. In such a case, firms may also end up earning lower profits than they would in a world without advertising. In the case of persuasive advertising, enhancing consumer heterogeneity seems to be the desired effect of promotional effort due to its competition-reducing ability. Generally speaking, this can be in detriment of social welfare even in the extreme case in which heterogeneity per se is in favor of a prosocial or environmental-friendly consumer behavior (Garcia-Gallego and Georgantzis 2009). Most of this wisdom has not been translated sufficiently into specific measures of economic policy and is largely neglected by the existing laws, except for some regulation regarding limitations of the overall time and space that advertising should be allowed in certain media.

Within the scope of informative advertising, special attention has been paid recently to best-price guarantees (e.g., *if you could find it cheaper, we refund the difference*) advertised by many large retailers all over the world. While, at a first glance, such a strategy appears to be a signal of the firm’s commitment to prices which are lower than any of its rivals, suspicion has been raised by some authors (Arbatskaya et al. 2004, 2006), arguing that price guarantees may be used by firms wishing to facilitate cartel sustainability and discourage price cuts, because the transparency achieved regarding rival prices makes deviations from collusive agreed prices easier to detect. Both the promotional and the cartel facilitating explanations of price guarantees have received some support. The controversy persists, calling for a case-by-case treatment (Fatas et al. 2013) when concerns arise regarding the true motivation and effects of such guarantees. Other subtle forms of informative advertising relate to signaling of the firm’s attitude toward certain quality aspects of their products, like

safety (ISO standards), chemical composition, fair trade rule compliance, and environmental performance achieved by certificates and labeling (Loureiro and Lotade 2005). In those cases, legislation is assisted by the certifying or label-awarding authority so that no issues arise regarding imperfect and asymmetric information.

Cross-References

- ▶ [Cartels and Collusion](#)
- ▶ [Distance Selling and Doorstep Contracts](#)
- ▶ [Labeling](#)
- ▶ [Liability and Information](#)
- ▶ [Internet Governance](#)
- ▶ [Labeling](#)
- ▶ [Signalling](#)

References

- Arbatskaya M, Hviid M, Shaffer G (2004) On the incidence and variety of low price guarantees. *J Law Econ* 47:307–332
- Arbatskaya M, Hviid M, Shaffer G (2006) On the use of low-price guarantees to discourage price cutting. *Int J Ind Organ* 24:1139–1156
- Barreda-Tarrazona I, Georgantzis N (2002) Regulating vertical relations in the presence of retailer differentiation costs. *Int Rev Law Econ* 22:227–256
- Becker G (1996) *Accounting for tastes*. Harvard University Press, Cambridge, MA
- Becker G, Murphy M (1993) A simple theory of advertising as a good or bad. *Q J Econ* 108:941–964
- Berger PK, Fraley GW, Tarpey LX (1964) The effects of advertising on the use of a train’s diner. *J Advert* 9:25–29
- Bester H, Petrakis E (1996) Coupons and oligopolistic price discrimination. *Int J Ind Org* 14:227–242
- Bloch F, Manceau D (1999) Persuasive advertising in hotelling’s model of product differentiation. *Int J Ind Organ* 17:557–574
- Carpenter GS (1987) Modeling competitive marketing strategies: the impact of marketing-mix relationships and industry structure. *Market Sci* 6:208–221
- Chapman RG (1986) Assessing the profitability of retailer couponing with a low-cost field experiment. *J Retail* 62:19–49
- Clow KE, Baack D (2014) *Integrated advertising, promotion, and marketing communications*. New York: Pearson Education Ltd
- Fatás E, Georgantzis N, Sabater-Grande G, Máñez J (2013) Experimental duopolies under price matching and price beating guarantees. *Appl Econ* 45:15–35
- García-Gallego A, Georgantzis N (2009) Market effects of changes in consumers’ social responsibility. *J Econ Manage Strat* 18:235–262

- Grossman GM, Shapiro C (1984) Informative advertising with differentiated products. *Rev Econ Stud* 51(1):63–81
- Gupta SK, Krishnan KS (1967a) Differential equation approach to marketing. *Oper Res* 15:1030–1039
- Gupta SK, Krishnan KS (1967b) Mathematical models in marketing. *Oper Res* 15:1040–1050
- Kahn BE, Raju JS (1991) Effects of price promotions on variety-seeking and reinforcement behavior. *Market Sci* 10:316–337
- Keller KL (1993) Conceptualizing, measuring, and managing customer-based brand equity. *J Market* 57:1–22
- Keller KL (2013) *Strategic brand management: building, measuring, and managing brand equity*. Prentice Hall, Upper Saddle River
- Kotler P, Armstrong G, Harris L, Piercy NF (2013) *Principles of marketing*. Essex: Pearson Education Ltd
- Krishnan KS, Gupta SK (1967) Mathematical model for a duopolistic market. *Manage Sci* 13:568–583
- Kuehn AA (1964) *The marketing concept in action*. American Marketing Association, Chicago
- Loureiro ML, Lotade J (2005) Do fair trade and eco-labels in coffee wake up the consumer conscience? *Ecol Econ* 53:129–138
- Miller BR, Strain CE (1970) Determining promotional effects by experimental design. *J Market Res* 7:513–516
- Mills HD (1959) A study in promotional competition. The Harlan D. Mills Collection. http://trace.tennessee.edu/utk_harlan/48
- Moraga-González JL, Petrakis E (1999) Coupon advertising under imperfect price information. *J Econ Manage Strat* 8:523–544
- Rossiter JR, Bellman S (2005) *Marketing communications: theory and applications*. New York: Pearson.
- Stigler G, Becker G (1977) De Gustibus Non Est Disputandum. *Am Econ Rev* 67:76–90
- von der Fehr NH, Stevik K (1998) Persuasive advertising and product differentiation. *South Econ J* 65:113–126

Property Rights: Limits and Enhancements

Geoffrey M. Hodgson
Hertfordshire Business School, University of
Hertfordshire, Hatfield, Hertfordshire, UK

Definition

This essay considers the concept of “property right” as typically employed in “the economics of property rights.” Some early uses of the term

“property right” in economics are quoted before moving to the predominant usages in the standard “economics of property rights” today. This prevailing usage is compared with the more nuanced notion of “property right” in legal theory. It is argued that, for the purposes of understanding the role and impact of property rights on economic performance, it would be better for economists to return to the legal meaning of property rights. This legal meaning can cope better with important developmental phenomena such as the use of property as collateral to finance loans. Also the recognition of the legal character of property opens doors for a richer understanding of motivations to respect property rights, which is consistent with empirical data indicating that people are not simply motivated to respect laws out of instrumental calculations of costs and benefits.

Introduction

Nowadays, many economists concur that the development of secure and enforceable property rights is a vital factor in the process of modern economic development. But what are property rights? Different answers to this question focus on different kinds of social relationship or institution, which, in turn, can lead to different policy recommendations. On closer inspection, there are major differences of opinion on this issue.

These differences can be divulged by posing two connected queries. First, are property rights conceived as an eternal feature of the human condition, or are they historically specific, emerging (say) with states and legal systems? Many researchers in this field treat property as an evolved – spontaneously or endogenously emerging – response to human interactions over territory or other wealth, and do not assume that states or legal systems are necessary for its emergence.

The division of opinion on the first query often related to the divide in the history of economics between those that emphasise universal principles or laws, and those that claim that concepts such as exchange, supply, demand, and even calculative rationality are historically specific, depending on particular economic institutions. Much of modern

This chapter distils and extends some material from Hodgson (2015a, b).

economic theory emphasizes universal principles, whereas others, including the German historical school and the original American institutionalists, emphasized historical specificity.

The second, related, query is whether “property” is treated as synonymous with mere possession – i.e., *de facto* control – over a resource, or does it involve the acquisition of declared rights, backed up by institutions of legislation, adjudication, and enforcement? Again the rival choice of emphasis on universality or historical specificity is relevant here. Those that seek a universal concept of property must find it also in worlds where there is no state, where custom rules without institutionalized law. By contrast, those that make property more than possession and relate rights to legal institutions are obliged to abandon the notion that property is a universal or ahistorical concept.

This entry argues that this fundamental difference of outlook over the nature of property has major theoretical and practical consequences. It has divided opinion since the disciplines of law and economics began to interact in the nineteenth century. It has yet to be played out to an adequate solution.

Consider some examples. John R. Commons was an early institutional economist and a leading explorer of the overlapping territories of economics and law. He was clear: “in the end, the actual title to property rests on the sovereign power of the state to enforce its decrees. . . . There is, strictly speaking, no such thing as absolute, unlimited right of property, which law steps in as an afterthought to restrict” (Commons 1893, p. 110).

The institutional economist Thorstein Veblen also pointed out that property is historically specific: “no concept of ownership, either communal or individual, applies in the primitive community. The idea of communal ownership is of a relatively later growth” (Veblen 1898, p. 358).

A historically specific, state-related concept of property also pervades much analysis by legal scholars. For example, the leading legal theorist Antony Honoré (1961, pp. 107, 115) insisted that property is much more than possession or control:

A people to whom ownership was unknown, or who accorded it a minor place in their arrangements, who meant by *meum* and *tuum* no more than

“what I (or you) presently hold” would live in a world that is not our world. . . . To have worked out the notion of ‘having a right to’ as distinct from merely “having” . . . was a major intellectual achievement.

Legal scholars Daniel Cole and Peter Grossman (2002) have documented how the notion of “property rights” in prevailing contemporary approaches in economics departs radically from that notion in law. For other insights and dissenting voices on the treatment of property, see Hernando (2000), Steiger (2008), Fukuyama (2011, pp. 66–71), Arruñada (2012, 2017), and Heinsohn and Steiger (2013).

This entry explores this crucial divergence of usage and interpretation of a key concept. The following section lays out the concept of property in much of “the economics of property rights.” Sections “[Why Do People Obey the Law?](#)” and “[Property Rights and Economic Development](#)” explore some implications of the discussion in terms of practical analyses of the roles of property. Section “[Concluding Remarks](#)” concludes the chapter.

The Concept of Property in the Contemporary “Economics of Property Rights”

At least in its early stages of its development, the now-conventional “economics of property rights” was influenced by the treatment of institutions in the Austrian school of economics, particular by Carl Menger, Ludwig Mises, and Friedrich Hayek.

Menger famously located the essence of some institutions in the spontaneous arrangements that engender or sustain them, rather than in acts of decree by a state or other public authority. Similarly, the standard economics of property rights strives to understand property as a spontaneous institution, which does not necessarily involve the state.

Mises made a sustained attempt to make core economic concepts as universal as possible. So when he considered the nature of ownership, he treated the legal aspect as merely a normative

(“ought to have”) justification of de facto “having” something:

From the sociological and economic point of view, ownership is the *having* of the goods . . . This *having* may be called the natural or original ownership, as it is purely a physical relationship of man to the goods, independent of social relations between men or of a legal order. . . . Economically . . . the natural *having* alone is relevant, and the economic significance of the legal *should have* lies only in the support it lends to the acquisition, the maintenance, and the regaining of the natural *having*. (Mises [1932] 1981, p. 27)

Hence, for Mises, ownership was natural and ahistorical rather than legal or institutional. An individual-physical rather than a social relationship, it was deemed independent of law or any other social institution. He downgraded the institutions required for the legitimation, protection, and enforcement of the capacity to have, and neglected social aspects of ownership that may signal power or status.

Hayek’s position was in some respects different from Mises, but by arguing that all law was basically custom, Hayek (1973, pp. 72–5) echoed Mises’s argument by removing any historical specificity to the notion of law. In part, this is a matter of the definition of law. But there are good reasons for regarding law as a characteristic of systems with a fully institutionalized judiciary and legislature (Hodgson 2015a, b). Hayek not only reduced law to custom, but he treated common law as essentially spontaneous and customary, whereas in fact it was by large part codified and made consistent by the state (Fukuyama 2011, pp. 254–60).

It is possible to reject Hayek’s extremely wide characterization of law without adopting the view that law emanates simply from the state, and without underestimating the crucial role of custom in legal development. Indeed, all law depends to a large degree on custom, for its origin or for its implementation. While law, in the narrower and historically specific definition, implies the state, it is not constituted simply by decree and it depends upon custom.

Notably, in contrast to Hayek, Ronald Coase never suggested that property in a modern market economy could exist or be understood without a state legal system. Coase (1988, p. 10) wrote:

When the physical facilities are scattered and owned by a vast number of people with very different interests . . . the establishment and administration of a private legal system would be very difficult. Those operating in these markets have to depend, therefore, on the legal system of the State.

Subsequent writers on “the economics of property rights” departed from this maxim, enabling them to tackle the definition of “property right” in a very general way, ignoring its dependence on “the legal system of the state.”

Armen Alchian was a key figure in the development of the contemporary “economics of property rights.” Alchian (1965) defined private property rights in terms of assignments of the ability to choose the use of goods (without affecting the property of other persons). Following Mises, his definition was largely in terms of de facto powers of control rather than legal or moral rights.

Alchian removed morality from the notion of a “right.” He downgraded the role of law, from being the source of legitimate authority concerning rights to one means among others for enforcing possession and control. But typically, we refer to moral or legal rights, we do not simply mean de facto ability to control resources. The *Oxford English Dictionary* defines “right” (as a noun) as “that which is morally correct, just, or honourable” or “a moral or legal entitlement to have or do something.” It means much more than the ability to do something.

Later Alchian (1977, p. 238) defined the “property rights” of a person in universal and institution-free terms including “the probability that his decision about demarcated uses of the resource will determine the use.” Alchian’s definitions of property neglect the essential concept of legitimated, rightful ownership. This concept is important, even if what is rightful is contestable or difficult to identify. His definitions denote possession rather than property.

Similarly, the highly influential property-rights economist Yoram Barzel (1994, p. 394) defined property as:

an individual’s net valuation, in expected terms, of the ability to directly consume the services of the asset, or to consume it indirectly through exchange.

A key word is *ability*: the definition is concerned not with what people are legally entitled to do but with what they believe they can do.

This explicitly removed the question of legal title from the definition of property. The upshot of this is that if a thief manages to keep stolen goods then he acquires a substantial property right in them, even if, on the contrary, legal or moral considerations would suggest that they remain the rightful property of their original owner. Elsewhere Barzel (1997, p. 3) argued:

The term 'property rights' carries two distinct meanings in the economic literature. One ... is essentially the ability to enjoy a piece of property. The other, much more prevalent and much older, is essentially what the state assigns to a person. I designate the first 'economic property rights' and the second 'legal (property) rights.' Economic rights are the end (that is, what people ultimately seek), whereas legal rights are the means to achieve the end. Legal rights play a primarily supporting role...

Barzel's distinction between "economic property rights" and "legal property rights" has proved influential in this genre of literature. Barzel made it clear that his version of "the economics of property rights" is not about legalities. But it is questionable to regard "the ability to enjoy" something as a "right." Enjoyment can exist without rights, and rights without enjoyment. Rights result from institutionalized rules involving assignments of potential benefit. They always involve relations between people as well as relations with things. The "ability to enjoy" may not involve more than an individual's relationship with an object.

One possible response is that all this is essentially about the choice of definitions and the uses of words. In particular, property rights economists might simply say that what some call "possession" they call "economic property rights" and what some call "property rights" they call "legal property rights." Consequently, we can go through and edit the relevant economics of property rights literature, making the appropriate terminological substitutions, and everything else would remain valid.

This is contestable. Words carry meanings that we cannot simply alter at will. While the foundational economics of property rights

literature brings important and valid insights, it would still be impaired after these terminological substitutions. First, using the word "right" to describe something that is not a right but a matter of de facto control is misleading: it obscures the adopted legal meaning of rights in modern legal and economic systems (Cole and Grossman 2002).

Second, attention is diverted from the roles of moral sentiments and dispositions to obey authority: they are important parts of human motivation, alongside greed and self-interest. Third, even with the terminological substitutions, the foundational property rights literature would have an inadequate treatment of (legal) property rights as collateral to obtain loans, and of the historically specific legal institutions that make collateralization possible. These issues are explored in the following two sections.

Why Do People Obey the Law?

Consider the intrinsic role of law in human motivation. Intrinsic motivation stems from the perceived nature of law itself, rather than from the sanctions or rewards of the legal system. The "economics of property rights" often neglects this intrinsic motivation: it is often assumed that law impinges on behaviour principally as a cost or constraint. Obeying the law is simply a matter of expected costs and expected benefits, where no benefit is assumed from legal compliance itself.

Hence, having made his distinction between two kinds of "rights," Barzel (2002, pp. 16, 157) claimed that: "What individuals maximize (subject to their personal safety) is the value of their *economic rights*." These exclude "legal rights," which are defined as "*claims over assets delineated by the state*." In other words, individuals are *ceteris paribus* indifferent to "legal rights" and act solely to maximize their enjoyment of assets under their control, whether these assets are obtained legally or illegally. This assumes a particular form of maximizing behaviour where law itself has no direct input as an argument in the preference function.

Barzel accepted that the law may matter. It may be a major factor in determining outcomes in terms of control and enjoyment of resources. But Barzel (1997, p. 3) argued that law matters only insofar as it leads to “what people ultimately seek,” namely the enjoyment of resources. Legal obedience has no intrinsic value, and it is simply treated as an instrumental “means” to that “end.” Legal obedience would not appear as an argument in any presumed preference function. Law is regarded as simply an instrumental means to an end. According to Barzel, obeying the law is not an end in itself.

By contrast, the subfield of psychological jurisprudence has gathered considerable survey data to establish that people place some normative value on obeying the law, in addition to any instrumental consideration of expected personal costs or benefits. Tom R. Tyler is the leading authority in this area. Tyler (1990, p. 3) contrasted the “instrumental perspective” where “people are viewed as shaping their behaviour to respond to changes in the tangible, immediate incentives and penalties associated with following the law” with the “normative perspective” concerned with “what people regard as just and moral as opposed to what is in their self-interest.”

Supported by evidence gained from a survey of several hundred citizens in Chicago, Tyler (1990) argued that citizen may be inclined to obey the law for noninstrumental as well as instrumental reasons. For some actors, the estimated costs (involving estimates of the chances of being caught and likely punishments) and benefits of breaking the law are weighed against the costs and benefits of compliance. But the evidence of Tyler and others suggests that such instrumental calculations have relatively little effect on compliance. More important is internalized obligation, stemming from other considerations.

First, citizens often comply with the law because they regard the legal authority as having a legitimate right to lay down rules that people must obey. The notion of legitimate authority can have several possible bases, but in modern democracies, it mainly derives from the belief that the popular election of a government makes such state authority legitimate. Max Weber ([1922]

1968, p. 215) saw legal authority and legitimation as “resting on a belief in the legality of enacted rules and the right of those elevated to authority under such rules to issue commands.”

Second, citizens sometimes comply with a particular law because it is believed to be moral. This is particularly the case with laws against murder or rape, but may not be so with other laws, such as those governing traffic on roads. Tyler argued that moral beliefs sometimes help to sustain laws. But this does not work when people believe that a particular law lacks sufficient moral force.

Instrumentalists in general, and proponents of utility-maximization in particular, argue that the issues of legitimacy and morality can be incorporated in the calculus of costs, benefits or utility. Hence advocates of utility-maximization claim that when someone acts according to their perceptions of legitimacy or morality, they gain extra utility from a “warm glow” of self-satisfaction. A problem with this argument is that it undermines the very meaning of legitimacy and morality. The whole point of actions in accord with perceived legitimacy or morality is that they do not necessarily serve self-interest. Acting morally means “doing the right thing,” even if it is costly to the actor. People obey legitimate legal authority saying “because it’s the law.” These deontic motivations cannot be reduced to convenience, convention, or cost-benefit calculation, in accord with a unidimensional calculus of utility or desire.

Both adults and children feel strong obligations to obey the law. Austin Sarat (1975) found that 70% of adults in his US sample agreed that a law “must always be obeyed.” A survey of US high school students found that 77% of whites and 72% of blacks agreed that “people should always obey the law” (Rodgers and Lewis 1974). Tyler (1990, p. 178) summarized a major implication of his survey of Chicago citizens:

People obey the law because they believe that it is proper to do so, they react to their experiences by evaluating their justice or injustice, and in evaluating the justice of their experiences they consider factors unrelated to outcome, such as whether they have had a chance to state their case and been treated with dignity and respect. . . . The image of the person resulting from these findings is one of a person whose attitudes and behavior are influenced

to an important degree by social values about what is right and proper. This image differs strikingly from that of the self-interest models which dominate current thinking. . .

Our ingrained inclinations to respect those in authority were dramatized by the experiments of Stanley Milgram (1974), who invited members of the public to help in a laboratory study ostensibly about learning. A “scientist” asked these recruits to administer electric shocks to a subject, to punish wrong answers to questions. Milgram found that a majority of adults would administer shocks that were apparently painful, dangerous, or even fatal, if ordered to do so by the person in authority. In fact, there were no shocks and the subject was an actor, feigning agony or even death. This experiment shows that people can willingly accept the orders of perceived authority figures, even when their own moral feelings are violated.

Milgram (1974, pp. 124–5, 131) argued that our dispositions to respect authority emanate from the evolutionary survival advantages of cohesive social groups. While socialization and learning are clearly important in developing propensities to obey authority and the law (Engel 2008), Milgram also proposed that the human species has evolved an inherited, instinctive, propensity for obedience that is triggered by specific social circumstances. He suggested that dispositions to respect authority have both genetic and cultural foundations. This is in accord with Charles Darwin in his *Descent of Man*, who proposed that human tribes that developed systems of social obedience and cooperation, and a moral code to buttress these attributes, would survive in competition with other human tribes and in dealing with their environment.

These arguments undermine the view of Barzel and others that respect for the law – based on its perceived legitimacy or moral concordance – plays little or no part in attitudes toward property or other legal rights. The insistence that property is a legal right does not imply that people never break the law, or that law alone somehow predicts behavior. But the establishment of legal rights, through perceptions of moral legitimacy and the use of state power, can affect intentions or behavior. As Honoré (1961, pp. 107, 115) insisted, an

economy involving mere possession is very different in nature and outcomes from one that has institutionalized rights of property.

The mistaken removal of legal rights from the definition of property cannot be justified on the ground that they are unnecessary to explain or predict behaviour. Any explanation of dispositions, choices, or preferences must take such factors into account. If economists are interested in predicting behavior on the basis of some scientific understanding of what causes it, then they must take matters of motivation, including the instrumental and the normative into account (Merrill and Smith 2007).

Property Rights and Economic Development

Arguments emphasizing the perceived legitimacy of the legal system have implications for establishing the rule of (state) law, and particularly installing just and secure property rights to help promote economic development.

China is an important test case for these arguments. China began its market reforms in 1978 and its systems of property, commercial, and corporate law are still relatively underdeveloped compared to Europe or North America. This fact, alongside its highly impressive economic growth since 1978, has led some economists to conclude that legally-enforced property rights are of lesser significance.

But, despite superficial appearances to the contrary, there is evidence that legal systems and legal property rights matter. China’s explosive growth started when land-use (*usus fructus*) rights were widely conceded to the peasants after 1978 (Coase and Wang 2012). Relevant legislation concerning land leasing followed rather than preceded this concession. But this does not mean that legal land-use rights were unimportant. Local power from below tentatively established *de facto* powers, which spread widely and became *de jure* when it was legally ratified by the state. This endorsement, along with the institutional arrangements established from below, was vital to safeguard these rights.

Legalities matter and evidence suggests that they matter still more as capitalism develops. Further economic development in East Asia may depend in part on the installation of superior state legal and political systems governing and protecting property and contracts. Private ordering is vital but insufficient. The cross-country evidence of Robert J. Barro (1997) and others suggests that economic growth is correlated with the rule of law, among other factors.

By contrast, the Alchian-Barzel approach might see it as sufficient that Chinese entrepreneurs can control resources sufficiently to enable prosperity and rapid growth. Matters of legal infrastructure and enforcement would be secondary. Instead, the alternative perspective outlined here would point to the priority of developing legal institutions, their separation from political authority, and the related reform of the political system.

As Hernando (2000) and others have pointed out, the registration and enforcement of property rights (particularly in land and buildings) is a necessary condition for economic development. But Arruñada (2017) argued persuasively that property registries, while necessary, are insufficient. Among other things, they need to be backed up by well-functioning systems of law enforcement.

The exclusive focus on control in “the economics of property rights” overlooks the use of property as collateral for loans. The possibility of collateralization – which relies on legal and financial institutions – cannot be based on possession alone. It involves institutions: relations between individuals as well as relations between individuals and things. While emphasizing the importance of “property rights,” much of this discourse side-lines the vital institutions that are required to sustain them and make them fully operational in a developed economy.

Another relevant historical case study is the English Glorious Revolution of 1688. Douglass North and Barry Weingast (1989) famously claimed that this event made property rights more secure by constraining the power of the monarch. But on closer inspection, it becomes clear that property rights in England were

relatively secure centuries before 1688 (Fukuyama 2011, pp. 418–20). The problem instead was that these rights involved *entails*, which kept land and buildings in the family and prevented their use as collateral to secure loans.

Instead of making property rights secure, 1688 led to new alliances and a series of wars, requiring the state to reform its administration, gather more taxes, and help create a new financial system. This created conditions in the eighteenth century that eventually incentivized and facilitated the use of land as collateral for loans for agricultural, infrastructural, and industrial development (Hodgson 2017). The overly simplistic treatment of property rights as mere possession obscures the importance of different aspects of property (Honoré 1961) and the crucial role of potential collateralization in creating the conditions for the Industrial Revolution.

Attempts to make “property rights” analysis universal, so that it applies to societies without effective (state) legal authority draw our attention away from the importance of property rights, properly defined (Arruñada 2017). Of course it is important to understand extra-legal enforcement mechanisms, such as with pirates and mafias, but we should not pretend that might and right are the same. Of course, studies of worlds without (state) legal enforcement can help us devise policies that apply to developing countries where the state is weak or dysfunctional. But – following Coase (1988) among others – we should not assume that such spontaneous mechanisms are sufficient, or can apply to large-scale, complex economies. The reinstatement of a more genuine concept of “property rights” in development economics would lead to greater emphasis on the importance of an effective legal system that enjoys some autonomy from political or sectional power with some perceived justice in its proceedings.

Concluding Remarks

The success of capitalism has depended on systems of law enforcement. But these took a long time to establish. Even today, in much of the world, systems of law enforcement are weak,

expensive, corrupt, or inaccessible. In their absence, people fall back on other means of establishing obligations and ensuring compliance. Commerce then works through clan or family ties, shared religion, or ethnicity, bureaucratic cooption and corruption, or threats of violence to person or property. Systems of spontaneous enforcement show how commercial agreements can be maintained in the absence of adequate state systems of law. Such systems existed in history and persist today in some contexts. Hence they are important objects of analysis. But this should not mislead us into believing that fully-developed modern capitalist systems can rest on purely spontaneous or customary foundations.

A crucial argument here is that legal institutions have to be taken into account in “economic” analysis. Law is much more than a constraint: it matters too for an adequate understanding of human motivation and the financial dynamics of capitalism. An alternative approach – which takes the impact of legal institutions more seriously – is described as *legal institutionalism* (Hodgson 2015a).

An appreciation of noninstrumental motivations for legal compliance, which are part and parcel of the arguments here concerning the nature of property rights, challenges the use of utility-maximization as sufficient model of human behaviour. If “economics” is confined to utility-maximizing agents, then “economics” cannot adequately deal with the reality of property rights. But if economics takes on board the insights of Adam Smith, Ronald Coase, Amartya Sen, and several others, who all adopted more complex views of human motivation involving moral sentiments, then the “economics of property rights” can be greatly enriched.

Cross-References

- [Austrian School of Economics](#)
- [Capitalism](#)
- [Constructivism, Cultural Evolution, and Spontaneous Order](#)
- [Economic Analysis of Law](#)

- [Economic Development](#)
- [Economics Imperialism in Law and Economics](#)
- [Hayek, Friedrich August von](#)
- [Institutional Economics](#)
- [Mises, Ludwig von](#)

References

- Alchian AA (1965) Some economics of property rights. *Il Politico* 30:816–829
- Alchian AA (1977) Some implications of recognition of property right transaction costs. In: Brunner K (ed) *Economics and Social Institutions: Insights from the Conferences on Analysis and Ideology*. Martinus Nijhoff, Boston, pp 234–255
- Arruñada B (2012) Institutional foundations of impersonal exchange: theory and policy of contractual registries. University of Chicago Press, Chicago
- Arruñada B (2017) Property as sequential exchange: the forgotten limits of private contract. *J Inst Econ* 13(4):753–783
- Barro RJ (1997) Determinants of economic growth: a cross-country empirical study. MIT Press, Cambridge, MA
- Barzel Y (1994) The capture of wealth by monopolists and the protection of property rights. *Int Rev Law Econ* 14(4):393–409
- Barzel Y (1997) *Economic Analysis of Property Rights*, 2nd edn. Cambridge University Press, Cambridge
- Barzel Y (2002) A theory of the state: economic rights, legal rights, and the scope of the state. Cambridge University Press, Cambridge
- Coase RH (1988) *The firm, the market, and the law*. University of Chicago Press, Chicago
- Coase RH, Wang N (2012) *How China became capitalist*. Palgrave Macmillan, London/New York
- Cole DH, Grossman PZ (2002) The meaning of property rights: law versus economics? *Land Econ* 78(3):317–330
- Commons JR (1893) *The Distribution of Wealth*. Macmillan, London/New York. Reprinted 1963 (New York: Augustus Kelley)
- Engel C (2008) Learning the Law. *J Inst Econ* 4(3): 275–297
- Fukuyama F (2011) *The Origins of Political Order: From Prehuman Times to the French Revolution*. Profile Books/Farrar, Straus and Giroux, London/New York
- Hayek FA (1973) *Law, legislation and liberty; volume 1: rules and order*. Routledge/Kegan Paul, London
- Heinsohn G, Steiger O (2013) *Ownership economics: on the foundations of interest, money, markets, business cycles and economic development*. Routledge, London/New York. Translated and edited by Frank Decker
- Hernando DS (2000) *The mystery of capital: why capitalism triumphs in the west and fails everywhere else*. Basic Books, New York

- Hodgson GM (2015a) Conceptualizing capitalism: institutions, evolution, future. University of Chicago Press, Chicago
- Hodgson GM (2015b) Much of the “economics of property rights” devalues property and legal rights. *J Inst Econ* 11(4):683–709
- Hodgson GM (2017) 1688 and all that: property rights, the glorious revolution and the rise of British capitalism. *J Inst Econ* 13(1):79–107
- Honoré AM (1961) Ownership. In: Guest AG (ed) Oxford Essays in Jurisprudence. Oxford University Press, Oxford, pp 107–147. Reprinted in *J Inst Econ*, 9(2), 2013, pp. 227–55
- Merrill TW, Smith HE (2007) The morality of property. *William Mary Law Rev* 48(5):1849–1895
- Milgram S (1974) Obedience to authority: an experimental view. Harper and Row/Tavistock, New York/London
- Mises L (1981) Socialism: An Economic and Sociological Analysis. Liberty Classics, Indianapolis. Translated from the second (1932) German edition
- North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. *J Econ Hist* 49(4):803–832
- Rodgers HR, Lewis E (1974) Political support and compliance attitudes: a study of adolescents. *Am Polit Q* 2:61–77
- Sarat A (1975) Support for the legal system. *Am Polit Q* 3(1):3–24
- Steiger O (ed) (2008) Property economics: property rights, Creditor’s money and the foundations of the economy. Metropolis, Marburg
- Tyler TR (1990) Why people obey the law. Yale University Press, New Haven
- Veblen TB (1898) The beginnings of ownership. *Am J Sociol* 4(3):352–365
- Weber M (1968) Economy and Society: An Outline of Interpretative Sociology. University of California Press, Berkeley. 2 vols, translated from the German edition of 1922

Proportionality Test

Tomáš Sobek¹ and Josef Montag²

¹Faculty of Law, Masaryk University, Brno, Czech Republic

²International School of Economics, Kazakh-British Technical University, Almaty, Kazakhstan

Abstract

Proportionality test is a legal method used by courts, typically constitutional courts, to decide hard cases, which are cases where two

or more legitimate rights collide. In such cases a decision necessarily leads to one right prevailing at the expense of another. In order to decide such cases correctly, the court must balance the satisfaction of some rights and the damage to other rights resulting from a judgment. This entry overviews the proportionality test and the four steps of implementing the test. We also discuss the incommensurability problem, which is the main criticism of the balancing approach.

Proportionality Approach to Hard Cases

Courts, constitutional courts, in particular, sometimes face cases that cannot be decided by referring to the letter of the law or an interpretation thereof. Such cases, henceforth “hard cases,” are characterized by a collision of legitimate rights, or public interests, that are protected by the law, typically by the constitution. Adjudicating such hard cases inevitably results in a decision as to which right will prevail at the expense of limiting the other right. That is to say, in such cases courts face a trade-off situation.

In order to decide such cases in a rational manner, the adjudicators need a criterion to distinguish good and bad trade-offs. After the Second World War, the German constitutional theory developed the proportionality test as its main methodological tool for reconciling conflicting rights and values. Proportionality test is a method that allows courts to structure the decision dilemma at hand in a way that facilitates the decision to be carried out in a rational and transparent manner by balancing the benefits and costs – broadly understood – of alternative courses of action. And this approach proved to be viable. The proportionality test, indeed, became a dominant technique of adjudication of hard cases around the world (Aleinikoff 1987).

Proportionality test can be viewed as a set of rules for determining the necessary and sufficient conditions for a limitation of a constitutionally protected right and ascertaining whether these conditions are satisfied. The proportionality test consists of four steps or sub-tests (Alexy 2014).

A limitation of a constitutional right by a legal act is constitutionally permissible if and only if (i) the act pursues a legitimate aim (legitimacy test), (ii) the act is capable of achieving this aim (suitability test), (iii) the act impairs the affected right as little as possible (necessity test), and (iv) the importance of achieving the aim outweighs the importance of preventing the limitation on the affected right (balancing test or proportionality *stricto sensu*).

If these conditions are satisfied, the test concludes that the right under consideration has greater weight and should prevail over the conflicting right. The resulting preference relation, however, is specific to the case at hand only and its context. The outweighed right is not to be considered invalid; it may itself outweigh some other right in different circumstances.

From the economic point of view, the proportionality framework is interesting because it embodies the requirement for efficiency. The rules of proportionality require that the tradeoffs that courts make are efficient in the sense that the prevailing right is of higher social value and that its protection is achieved with minimal costs for the “losing” right. The test postulates that there should be a reasonable relationship between a particular purpose to be achieved by law and the legal means used to achieve that purpose. Judges must pay attention to the social importance of achieving the law’s purpose as well as the social importance of preventing the harm to other constitutionally protected rights. Put simply, they must consider not only the benefits associated with the proposed decision but also the damage to other rights caused by it.

The Steps

Legitimacy Test

The legitimacy test is simply a requirement that the legal principle, right, or public interest which the decision would uphold is lawful and reasonable, and so it should not be dismissed right away. This step simply aims to filter out cases, which are to be decided on different grounds and for which the proportionality test would be therefore

obsolete. The second, third, and fourth sub-tests express the idea of optimization relative to the factual and legal possibilities, and so we will discuss them in more detail.

Suitability Test

The stage of suitability review examines whether there is a reasonable connection between the act that interferes with a right and its legitimate aim. The interference must be capable of contributing to the achievement of the legitimate aim at least to a small extent. The compliance with this requirement excludes the adoption of means which obstruct some right without promoting any other right or interest for which it has been adopted. The point of this stage is to sort out those cases where there does not actually exist a conflict of rights (or a conflict of a right and public interest). Naturally, a conflict of two rights means that one cannot realize one right without detriment to the other. If the interference with one right does not contribute to the realization of any other right then this interference would be merely an obstruction without any conflict of rights justifying it.

Necessity Test

The stage of necessity reviews examines whether the proposed legal act impairs an affected right as little as possible. The act is necessary if there are no alternative measures that may achieve the same purpose with a lesser degree of impairment. If there is an equally suitable means that interferes with the affected right less intensively, then things can be improved without any costs. The necessity requirement postulates that no such free lunch is left on the table. So, this sub-test requires that of two means promoting the same aim (a certain right or public interest) to the same degree, the one that interferes less intensively with a conflicting right, i.e., generates lower costs, has to be chosen. This shows that the necessity stage is an expression of the idea of efficiency.

Balancing Test

While the stages of suitability and necessity essentially involve optimization relative to the factual possibilities, the stages of legitimacy and balancing (proportionality in the narrower sense)

refer to what is legally possible. Balancing, the last stage of the proportionality test involves optimization relative to the competing rights. The greater the degree of non-satisfaction of, or detriment to, one right, the greater is the required importance of satisfying the other right for the balancing test to tilt in its favor (Alexy 2002, p. 102). This requirement excludes an intensive interference with one right that is justified only by a low importance assigned to the satisfaction of the colliding right.

The process of balancing can be broken down into three steps. (i) In the first step, the degree of non-satisfaction of or detriment to a right is established. (ii) The second step consists of establishing the importance of satisfying the competing right. (iii) And in the last step, it is examined whether the importance of satisfying the latter right justifies the detriment to or non-satisfaction of the former.

These three steps of balancing are more explicitly captured by the weight formula:

$$W_{i,j} = \frac{I_i \cdot W_i \cdot R_i}{I_j \cdot W_j \cdot R_j},$$

where i and j index the colliding principles, I_i and I_j capture the intensity of interference in the respective rights in case of a decision in favor of the other right, W_i and W_j represent the abstract weights of the two colliding principles, and R_i and R_j capture the reliability of empirical assumptions on both sides of the argumentation (Alexy 2014, pp. 55–56). If the weight is greater than one, principle i should prevail; if it is less than 1, the court should rule in favor of principle j .

The data entering the formula is necessarily imprecise and based on court's judgment. This, however, stems from the very nature of the decision problems that courts face and does not invalidate the balancing approach itself. To illustrate how such balancing can be approached, Alexy (2014) suggests to use the triadic scale often used in the constitutional discourse: light (l), moderate (m), and serious (s). The levels of this scale can then be mapped to a numeric scale with more than proportional increments, such as the geometric sequence such $2^0, 2^1, 2^2$. This captures the idea

that the power of principles increases with the intensity of interference more than proportionally (the “marginal cost” of interference is increasing).

Incommensurability Criticism

At a first glance, it looks like the proportionality test requires the judges to reconcile irreconcilable interests: the freedom of expression versus the right to good reputation, the protection of public safety versus the right to personal privacy, the protection of natural environment versus the freedom of business, the protection of public order versus the freedom of religion, the right to a fair trial versus the economic efficiency of judicial process, and so on.

An important objection against the proportionality test is that the idea of balancing presupposes comparing rights as value bearers without a common metric. The main point of this line of criticism is that the balancing of rights (or the public interest) compares items which are “incommensurable” (see Chang 1997, 2013; Frank 2008; Raz 1986, Chap. 13). The late US Supreme Court Justice Antonin Scalia, who may count as one of the pronounced proponents of the incommensurability objection, put it bluntly: “[T]he scale analogy is not really appropriate since the interests on both sides are incommensurate. It is more like judging whether a particular line is longer than a particular rock is heavy” (Bendix Autolite Corp. v. Midwesco Enterprises 1988).

This variant of the incommensurability objection is based on intuition that we can only compare two things of the same kind so that it is impossible to perform a rational cardinal comparison between gains and losses for colliding rights or the public interest that are incommensurable. How convincing is this objection?

We suggest that the famous idiom about the comparison of apples with oranges should not be used in a mechanical way because comparing things of different kinds can, indeed, be meaningful. Incommensurability need not imply incomparability (Da Silva 2011, see also Chang 2002). Balancing can work in a satisfactory manner as long as comparability among the colliding rights

is rationally justified, no matter whether they are commensurable or not. And for this we would need to know what exactly we are actually comparing during the balancing exercise.

The answer to that may be relatively straightforward. We compare the marginal social importance in fulfilling one right and the marginal social importance in preventing the harm to another right (Barak 2012, p. 484). However, the problem here is that “[i]dentifying a single criterion does not eliminate incommensurability if the application of the criterion depends on considerations that are themselves incommensurable” (Endicott 2014, p. 311). And so one can still be legitimately worried that if the judges are required to balance things that cannot be compared, then the conclusion of such a balancing is nothing more than the outcome of their arbitrary preferences or subjective feelings (Alder 2001, p. 717).

The severity issues raised by the incommensurability criticism notwithstanding the problem with this argument as, we see it, is twofold: First, if one needs to make a decision, it is not of much help to point out that the decision problem is difficult or impossible to crack. One can contemplate indefinitely what is the proper course of action in the trolley dilemma (Edmonds 2014) as long as this happens in a thought experiment. When, however, the trolley is actually hurtling down a track toward the five people tied to it, a decision has to be made whether to push the fat man over the bridge, saving the five, or let them die, saving the fat man. Courts are not in the business of thought experimentation; they need to make the actual decisions.

The second problem with the incommensurability criticism can be seen when we bring it *ad absurdum*. One may argue that individuals make decisions based on comparisons of incommensurable courses of actions on a daily basis. Should I order a steak or a salad? Should I snooze the alarm or get up right now? Should I stay in the office or give my dog a proper walk? It is hard to see how these decisions are made using some common objective scale. Indeed, all of these choices are different from one another along different dimensions, some of which they may not share, making it hard to argue that there is a common “objective”

denominator in any of these dilemmas. If the incommensurability argument was valid, all peoples’ choices could be regarded as purely incidental. Such an argument would be problematic not least from the legal point of view as the bulk of law presupposes individual rationality and the ability of people to make sensible choices for themselves (see Chang 2012).

Conclusion

When individuals choose the desired course of action, they do so based on their preferences or feelings. The courts are not in a fundamentally different position in this respect, save for two important aspects: (i) unlike individuals who decide for themselves, judges sometimes must make decisions on behalf of the society, and (ii) unlike individuals who face the consequences of their own decisions, the consequences of judges’ decisions are born by the society at large or other persons. This makes adjudicating hard cases more delicate, difficult, and prone to error. Indeed, judges are entrusted with decision tasks that people seldom face in their private lives. Proportionality approach forces the judge to explicitly articulate whether and why she evaluates the desired effects of her decision as proportional with respect to the undesired limitations the decisions implied for the other rights, using a common, if abstract, metric for such a comparison and to do so in a manner that is methodical and consistent across cases (see Klatt and Meister 2012).

Proportionality test does not need to deliver a rock solid solution in every case in which it is conducted. Indeed, the test may sometimes yield an ambiguous result. Referring to our weighting formula, this would happen when the resulting weight is equal to one or when alternative but plausible assumptions behind the individual entries (i.e., uncertainty) produce different outcomes of the test. In such stalemate situations, there is no way out other than court’s discretion to decide either way (Alexy 2014, see also Chang 2002; Da Silva 2011; Raz 1999, Chap. 3). It goes without saying that where the test is valuable is precisely in cases where it points one direction

with enough certainty. It is these cases where it helps the court to make a sound decision and avoid a bad trade-off.

Incommensurability criticisms raise valid points about the fundamental challenges associated with the proportionality test. Indeed, detailed awareness about the problems concerning the determination of the relative value of rights and infringements may improve the quality of courts' deliberation. However, pointing out that two rights are not commensurable is not very helpful when the court has to decide between them somehow. In the absence of availability of a better alternative decision algorithm, the proportionality appears to be the go-to solution. The value of the proportionality project seems to be in that it forces courts to articulate their arguments in an explicit, structured, and transparent manner so that the court's deliberation and final decision can be understood (Klatt and Meister 2012, see also Winter 2016), even if one might disagree with the weights assigned to individual elements in the weighting exercise.

Cross-References

- [Conflict of Laws](#)
- [Cost–Benefit Analysis](#)
- [Judicial Decision-Making](#)

References

- Alder J (2001) Incommensurable values and judicial review: the case of local government. *Public Law* 4:717–735
- Aleinikoff TA (1987) Constitutional law in the age of balancing. *Yale Law J* 96:943–1005
- Alexy R (2002) *A theory of constitutional rights*. Oxford University Press, Oxford
- Alexy R (2014) Constitutional rights and proportionality. *Revus* 22:51–65
- Barak A (2012) *Proportionality: constitutional rights and their limitations*. Cambridge University Press, Cambridge, UK
- Chang R (1997) Introduction. In: Chang R (ed) *Incommensurability, incomparability, and practical reason*. Harvard University Press, Cambridge, MA
- Chang R (2002) The possibility of parity. *Ethics* 112:659–688
- Chang R (2012) Are hard choices cases of incomparability? *Philos Issues* 22:106–126
- Chang R (2013) Incommensurability (and incomparability). In: LaFollette H (ed) *International encyclopedia of ethics*. Blackwell Publishing, Malden, pp 2591–2604
- Da Silva VA (2011) Comparing the incommensurable: constitutional principles, balancing and rational decision. *Oxf J Leg Stud* 31:273–301
- Edmonds D (2014) Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong. Princeton University Press, Princeton
- Endicott T (2014) Proportionality and incommensurability. In: *Proportionality and the rule of law: rights, justification, reasoning*. Cambridge University Press, Cambridge, UK, pp 311–342
- Frank RH (2008) Why is cost-benefit analysis so controversial? In: Hausman DM (ed) *The philosophy of economics: an anthology*. Cambridge University Press, New York, pp 251–269
- Klatt M, Meister M (2012) *The constitutional structure of proportionality*. Oxford University Press, Oxford
- Raz J (1986) *Morality of freedom*. Oxford University Press, Oxford
- Raz J (1999) *Engaging reason: on the theory of value and action*. Oxford University Press, Oxford
- Winter J (2016) Alexyho vážící formule. *Právník* 5:446–461

Prosocial Behaviors

Sébastien Roussel

Department of Economics, CEE-M, Université Montpellier, CNRS, INRA, SupAgro, Université Paul Valéry Montpellier 3, Montpellier, France

Abstract

Prosocial behaviors contribute to the well-being of others following activities like charitable giving or volunteering. Theories that explain prosocial behaviors are pure altruism and impure altruism, inequality aversion reciprocity and conditional cooperation. These theories are linked to a system of motivations, that is intrinsic motivation, extrinsic motivation, and image motivation. Social norms play a key part in one's motivations to behave prosocially, in particular with regard to image motivation and likely sanctions when one's deviates from these norms. The expressive power of law leads to norms legitimacy. Prosocial behaviors are mainly studied in the literature in economics and psychology, especially in behavioral and experimental economics.

Definition

Prosocial behaviors refer to behaviors contributing to the interests of others and not to its own self-interest. Antisocial behaviors stand for the opposite of prosocial behaviors.

Introduction

From the seminal example of blood gift (Titmuss 1970) to the field of charitable giving (Bekkers and Wiepking 2011), many situations show that people act in a prosocial manner. Prosocial behaviors contribute to the well-being of others following activities like charitable giving (monetary giving or sharing for example), volunteering, voting, etc. List (2011) highlights that, in a regular year, total charitable gifts of money represent 2% of the US Gross Domestic Product (GDP), e.g., worth \$314 billion in 2007. This latter figure has been increasing to reach consecutive records in 2014 and 2015, with an estimated \$373.25 billion in 2015 (Giving USA 2016).

In terms of the history of economic thought, prosocial behaviors have been originally considered through Adam Smith's visionary overview of one individual's rationale. Smith (1759) stated that: "How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it" (cited by Meier 2007) with regards to "*the passions and motives which influence it*" (cited by Bénabou and Tirole 2006). This interpretation of rationality already suggested that decision-makers are not always egoistic. Consequently, one may analyze the role of social preferences/other-regarding preferences and motivations leading to prosocial behaviors (section "[Social Preferences and Motivations](#)"), as well as the part played by the institutions and the legal framework through social norms and the expressive function of law (section "[Social Norms and Laws](#)"), and how economics model and assess these behaviors (section "[Modeling and Assessing Prosocial Behaviors](#)").

Social Preferences and Motivations

In a detailed survey, Meier (2007) underlines the economic theories that explain prosocial behaviors through social preferences. These theories are respectively: pure altruism (Eckel and Grossman 1996) and impure altruism (Andreoni 1990), inequality aversion (Fehr and Schmidt 1999), reciprocity (Rabin 1993), and conditional cooperation (Frey and Meier 2004). While pure altruism and inequality aversion refer to situations where individuals act themselves selflessly, impure altruism, reciprocity, and conditional cooperation suggest that individuals expect to get something back from their prosocial activities.

The theories involving social preferences are linked to a system of motivations that result in prosocial behaviors in different ways. Ariely et al. (2009) describe the set of motivations that leads people to adopt prosocial behaviors. These motivations are in a broad sense: intrinsic motivation, extrinsic motivation, and image motivation. Intrinsic motivation refers to prosocial preferences like altruism, while extrinsic motivation indicates how external factors like monetary rewards matter in acting prosocially. Image motivation deals with the impacts of the others' perception and the individual's self-esteem on prosocial behaviors, and how this may frame one's behavior. However, the interplay between these motivations is critical. For example, many studies in economics and psychology show how intrinsic and extrinsic motivation may be complement or substitute, leading to crowding in – crowding out phenomena (Frey and Oberholzer-Gee 1997). This is also linked to the key role played by social norms.

Social Norms and Laws

Social norms play a key part in one's motivations to behave prosocially, in particular with regard to image motivation and likely sanctions when one's deviates from these norms. Social norms rely on a consensus towards acceptable behaviors (Elster 1989; Fehr and Schmidt 1999). Viscusi et al. (2011) characterize social norms "in terms of what is normatively appropriate rather than what

is the conventional mode of behavior,” and study how personal and external norms matter in recycling behavior. Another classical distinction refers to injunctive norms and descriptive norms, where the former characterizes the perception of what should be socially approved, and the latter characterizes the perception of what is effective. What is at stake is to conform or not to these norms, along with the individual valuation of what is morally good or wrong, and how these norms are enforced. Besides, identity plays a meaningful role while individuals follow prescriptions or social norms to secure their self-concepts (Akerlof and Kranton 2000).

The institutionalization of social norms into laws has been underlined by McAdams (2015) through the expressive powers of law, allowing for the recognition of good behavior versus bad behavior, leading to injunctive norms legitimacy. Moreover, he shows that law contributes to a coordination function, discloses information and beliefs, and provides compliance rules as explicit incentives. This was previously highlighted in the literature by the expressive function of law (Sunstein 1996a, b). Public authorities disclose the social meaning of one’s activities to help people make desired choices in a so-called norm management. In addition, Funk (2007) provides an empirical analysis of voting turnout and identifies the greater impact on behavior when law targets at the civic duty.

Modeling and Assessing Prosocial Behaviors

Prosocial behaviors are mainly studied in the literature in economics and psychology, especially in behavioral and experimental economics. Participation in prosocial activities is then traditionally modeled as a contribution to a public good (Andreoni 1990). Regarding the expressive content of laws, it has been shown how this can be modeled in the presence of norms or reputational rewards following Bénabou and Tirole (2006, 2011). In the literature, economic games are used in experiments to test hypotheses on social preferences and motivations that lead to prosocial

behaviors. In these experiments, a prosocial behavior like giving more is desirable (monetary or non-monetary contribution), and economic incentives relying (or not) on intrinsic, extrinsic, and image motivations can be implemented.

Traditional economic games used in the experiments on prosocial behaviors are mainly the ultimatum game, the (repeated) public good game, or the dictator game. For example, the dictator game allows testing individual generosity (List 2007), where the recipient must accept the sum offered by the dictator; consequently, the dictator’s behavior is then not strategic. It has also been shown that a social link between the dictator and the recipient increases the incentive to donate (Engel 2011), in particular when the recipient is a nongovernmental organization (NGO) like the Red Cross (Eckel and Grossman 1996). Furthermore, psychological and biological factors drive prosocial behaviors in expressing the motivations stated above. Moods and emotions do influence prosocial behaviors in various ways. For instance, gratitude plays a key role in building relationships and shaping costly prosocial behaviors (Bartlett and DeSteno 2006), awe increases donation amounts (Ibanez et al. 2017), whereas anger can impact negatively these behaviors (Drouvelis and Grosskopf 2016). With regards to the impact of laws on behavior, Galbiati and Vertova (2014) show from alternative public good games that obligations (laws and other formal rules) and economic incentives are complementary to incentivize cooperative behaviors.

Cross-References

- [Altruism](#)
- [Intrinsic and Extrinsic Motivation](#)
- [Public Goods](#)

References

- Akerlof GA, Kranton RE (2000) Economics and identity. *Q J Econ* 115(3):715–753
- Andreoni J (1990) Impure altruism and donations to public goods: a theory of warm-glow giving. *Econ J* 100: 464–477

- Ariely D, Bracha A, Meier S (2009) Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am Econ Rev* 99(1):544–555
- Bartlett MY, DeSteno D (2006) Gratitude and prosocial behavior. *Psychol Sci* 17(4):319–325
- Bekkers R, Wiepking P (2011) A literature review of empirical studies of philanthropy: eight mechanisms that drive charitable giving. *Nonprofit Volunt Sect Q* 40(5):924–973
- Bénabou R, Tirole J (2006) Incentives and prosocial behavior. *Am Econ Rev* 96(5):1652–1678
- Bénabou R, Tirole J (2011) Laws and norms. The National Bureau of Economic Research (NBER) working paper no. 17579
- Drouvelis M, Grosskopf B (2016) The effects of induced emotions on pro-social behaviour. *J Public Econ* 134:1–8
- Eckel CC, Grossman PJ (1996) Altruism in anonymous dictator games. *Games Econ Behav* 16:181–191
- Elster J (1989) Social norms and economic theory. *J Econ Perspect* 3(4):99–117
- Engel C (2011) Dictator games: a meta-study. *Exp Econ* 14(4):583–610
- Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. *Q J Econ* 114(3):817–868
- Frey BS, Meier S (2004) Social comparisons and prosocial behavior: testing conditional cooperation in a field experiment. *Am Econ Rev* 94(5):1717–1722
- Frey BS, Oberholzer-Gee F (1997) The cost of price incentives: an empirical analysis of motivation crowding-out. *Am Econ Rev* 87(4):746–755
- Funk P (2007) Is there an expressive function of law? An empirical analysis of voting laws with symbolic fines. *Am Law Econ Rev* 9:135–159
- Galbiati R, Vertova P (2014) How laws affect behavior: obligations, incentives and cooperative behavior. *Int Rev Law Econ* 38:48–57
- Giving USA (2016) The annual report on philanthropy for the Year 2015. The Indiana University Lilly Family School of Philanthropy
- Ibanez L, Moureau N, Roussel S (2017) How do incidental emotions impact proenvironmental behavior? Evidence from the dictator game. *J Behav Exp Econ* 66:150–155
- List JA (2007) On the interpretation of giving in dictator games. *J Polit Econ* 115(3):482–493
- List JA (2011) The market for charitable giving. *J Econ Perspect* 25(2):157–180
- McAdams RH (2015) The expressive powers of law: theories and limits. Harvard University Press, Cambridge, MA
- Meier S (2007) A survey of economic theories and field evidence on pro-social behavior. In: Frey BS, Stutzer A (eds) *Economics and psychology: a promising new cross-disciplinary field*. MIT Press, Cambridge, MA, pp 51–88
- Rabin M (1993) Incorporating fairness into game theory and economics. *Am Econ Rev* 83(5):1281–1302
- Smith A (1759) *The theory of moral sentiments*. Prometheus Books, Amherst
- Sunstein CR (1996a) On the expressive function of law. *Univ Pa Law Rev* 144:2021–2053
- Sunstein CR (1996b) Social norms and social roles. *Columbia Law Rev* 96:903–950
- Titmuss RM (1970) *The gift relationship: from human blood to social policy*. Allen and Unwin, London
- Viscusi WP, Huber J, Bell J (2011) Promoting recycling: private values, social norms, and economic incentives. *Am Econ Rev* 101(3):65–70

Prostitution

► Sex Offenses

Prostitution, Demand and Supply of

Øystein Hernæs^{1,2}, Niklas Jakobsson³ and Andreas Kotsadam⁴

¹Department of Economics, European University Institute, Florence, Italy

²Norwegian Social Research, Oslo, Norway

³Oslo and Akershus University College of Applied Sciences, Norwegian Social Research and Karlstad University, Oslo, Norway

⁴Department of Economics, University of Oslo, Oslo, Norway

Abstract

The market for sex is a contentious one, and has often been subject to heavy regulation. This chapter goes through factors that are important with regards to the demand for and supply of prostitution. A particular focus is on the relationship between laws and the quantity of sex bought and sold. The most common way that laws have been used to affect the demand for prostitution is by outright criminalization, which may lead to less prostitution, but may also drive the activity further underground. The effect of criminalizing prostitution on trafficking is ambiguous since criminalization may also lead to a substitution effect towards more trafficked prostitutes. Scarcity of reliable data is one of the main challenges for the study of prostitution.

Definition

The demand and supply of prostitution

Introduction

Buying and selling of sex can be analyzed as any regular economic exchange. The demand side is made up of customers who compensate the suppliers, prostitutes, for the commodity, or service, of sex. This all takes place within some regulatory framework.

The market for sex is a contentious one and has often been subject to heavy regulation. Both these facts have spurred academic research, as this chapter will show. However, the field has until recently not been very well developed, mostly because it has been plagued by a scarcity of reliable data. The lack of data is due to both the fact that prostitution in many societies is illegal or just at the margin of being illegal and the social stigma attached to being associated with the trade.

This chapter goes through factors that have found to be important with regard to the demand for and supply of prostitution. A particular focus is on the relationship between laws and quantity of sex bought and sold. It ends with an identification of gaps in the literature and a discussion of some underexplored questions and fruitful avenues for further research.

Demand

Amenable to Change Through Laws

The most common way that laws have been used to affect the demand for prostitution is by outright criminalization. The possibility of a penalty is an unambiguous (expected) cost for a potential buyer and will lead to less prostitution. This mechanism is confirmed empirically by Jakobsson and Kotsadam (2012), who investigate the effect of criminalization of buying sex on the amount of sex bought. Using longitudinal data from Norway and Sweden and employing a difference-in-differences approach, they estimate that the criminalization of buying sex in Norway in 2009 leads

to a downward change in the amount of sex bought of 3.5% points. They do not find an effect on attitudes in this sample and conclude that the most likely mechanism is an increased risk of getting caught. A serious problem in this type of studies is that they rely of self-reporting, in particular self-reporting of activities that are not only highly stigmatized but also change legal status during the period investigated. The authors also use different types of measures, one direct whereby people are asked if they had bought sex during the last 6 months and one more indirect where the respondents are asked if they know someone having bought sex during the last 6 months. While it is likely that respondents are more likely to report on people they know, it is still an inherent problem that the people they know perhaps keep such information more to themselves after it has become illegal.

If demand is relatively inelastic, criminalization of clients may to a large extent drive the activity further underground, with a corresponding increase in risk and worsened working conditions for prostitutes.

Laws may also have a normative function. In particular, it may have an expressive function in signaling a society's beliefs about what is considered acceptable behavior. Jakobsson and Kotsadam (2011) analyze whether the criminalization of buying sex in Norway in 2009 affected attitudes to prostitution. Again using longitudinal data from Norway and Sweden and employing a difference-in-differences approach, they find that the law made people in the capital more negative toward buying sex, but no effect on moral attitudes in the aggregate. They explicitly relate this to a framework whereby people that are most affected by legal changes are most likely to change their attitudes and argue that since people in the capital area were more exposed to street prostitution before the law and saw a large immediate reduction afterward, they should be more likely to be affected. They also find that young people were more affected in their attitudes by the law, which they link to socialization theory arguing that young people are more adept to changing their moral values. On the other side, passing laws that are difficult to enforce or against which there

is significant disagreement may weaken the respect for laws in general.

Della Giusta et al. (2009b) model reputation and stigma explicitly. Loss of reputation may be incurred by both buyers and sellers of sex and reduce the amount of sex exchanged. As described in the studies above, moral attitudes may not be very responsive to the law, but criminalization certainly carries social stigma in itself. Although Jakobsson and Kotsadam (2012) do not find an effect on moral attitudes of Norway's prostitution law, it is possible that the law increased the stigma related to buying sex and thereby added to the formal penalty. Della Giusta (2010) suggests that politicians adopt overly repressive regimes to please conservative swing voters.

Della Giusta et al. (2009a) delve empirically into the demand side and find that stigma does play an important role. They also find that demand may be multifaceted, e.g., in terms of risk preferences, and stress the importance of recognizing this when designing policy.

Important Factors Less Responsive to Legal Change

Insofar as sex is a normal good, the amount bought depends positively on an individual's income. However, prostitution is more common in poor countries. This will often have to do with supply, discussed below, but higher income of the customers may also decrease prostitution. Edlund and Korn (2002) propose a theoretical model in which increases in male (customer) income decrease the number of prostitutes because men prefer women as wives rather than as prostitutes. Relatedly, men's marriage opportunities may to a large extent determine their demand for prostitution and itself be determined by income, although married men constitute a large share of the demand. This suggests that policies and regulations promoting stable employment opportunities and a social safety net are relevant tools for authorities concerned with prostitution.

Noncommercial nonmarital sex and commercial nonmarital sex are typically thought of as substitutes for marital sex (Edlund and Korn 2002), but the relation between commercial and

noncommercial nonmarital sex is ambiguous. On factors such as pornographic material, a society's sociosexual culture, and the distinctness of gender roles, there is also little consensus.

Sex ratios in the population are also thought to be important. Places where men have outnumbered women to a significant degree have largely seen more prostitution. As Edlund and Korn (2002) point out, these places have tended to consist of men in transit, who participate only in a locality's sex market and not in its marriage market.

Supply

In some countries, selling sex is illegal while buying is legal, but a number of countries maintain the illegality of both sides of the market. When selling sex is illegal, the potential prostitute faces formal sanctions in addition to social stigma. The occupation of prostitute is thus made less appealing, and the woman in question is perhaps led to consider other even less desirable options. If prostitution does not entail any negative externalities and criminalizing selling sex is bad for women in the business, lawmakers should of course think twice before enacting such laws. What women do, and even if they do change occupation, after it becomes illegal is, however, unknown, but there are indications of severe externalities in terms of sexual trafficking, so a firm policy conclusion cannot be reached at this stage.

The most obvious factor affecting the supply of prostitution is income relative to other occupations and in general employment opportunities for women. As with any other type of work, the better potential labor market elsewhere, the less supply there will be of workers in a particular trade. As with the demand side, this suggests a role for providing employment opportunities and a social safety net in reducing prostitution. Other factors that impact the profitability of prostitution are sex ratios and inequality. To the best of our knowledge, there are no empirical studies investigating these issues.

Another standard factor in occupational choice is working conditions. As an effect of being

undertaken outside the formal economy, prostitution has often lacked regulation and oversight. In general, prostitutes face a more risky working environment than other laborers. In Europe, a significant minority of countries have legalized prostitution and tried to regulate the industry (Jakobsson and Kotsadam 2013). Safe working conditions, access to health services, and the rights to public benefits are all factors that make prostitution relatively more tempting. Regulations denying these to an occupation, as is largely the case for prostitution, can therefore be expected to lead to less selection into that profession. Any form of criminalization will likely lead the mere advocacy of basic rights and labor standards more difficult.

Prostitution is distinct from most other professions in that it carries a significantly higher opportunity cost in terms of social stigma. Della Giusta et al. (2009b) model reputation explicitly. Criminalization will mostly affect the stigma attached to the side or activity that is criminalized; however, it might spill over to other sides as well. For instance, outlawing buying of sex could increase the stigma not only of buying sex but also of selling. Edlund and Korn (2002) stress foregone marriage opportunities as the most important opportunity cost of prostitution. This is due to limited mobility between the marriage and prostitution sectors, a feature that also explains prostitutes' relatively high (up-front) wages. However, Arunachalam and Shah (2008) find no empirical support for the marriage-based explanation of the earnings premium in sex work.

There have been many instances in which women have been coerced to work as prostitutes. Forced labor is now outlawed in almost all countries – prostitution legislation aimed at alleviating this problem has therefore often been aimed at restricting the market, in the hope that this will hurt the potential profits of traffickers. Related to this, there has recently been developed a separate literature on trafficking. Although any kind of criminalization will reduce prostitution, the effect on trafficking is ambiguous, since criminalization may also lead to a substitution effect toward more trafficked prostitutes (Cho et al. 2013). Jakobsson and Kotsadam (2013)

and Cho et al. (2013) investigate this question using cross-country data and case studies and find that harsher sex laws are associated with less trafficking.

Notwithstanding the problem of forced labor in the sex trade, a significant part of international sex labor migration can be seen as voluntary and economic in motivation. Thus, policies that promote freer movement of labor and fewer restrictions on travel can be expected to increase the amount of prostitution. On the other hand, better opportunities to freely travel and search for other jobs may reduce the possibilities of human traffickers to exploit or deceive vulnerable migrant and force them into the sex trade.

A general issue with regard to trafficking is the tension between the welfare of potential trafficking victims and that of current prostitutes. Restricting the market for prostitution, either through the denial of benefits to prostitutes or some form of criminalization, is likely to lead to prostitution being less profitable and thus less tempting for traffickers but is also likely to hurt those currently in the profession. Lee and Persson (2013) propose a regulatory framework that provides a solution to this conundrum: licensing prostitutes and criminalizing buying sex from unlicensed prostitutes. This model deters demand for nonvoluntary prostitutes and thus reduces the potential profit for traffickers while at the same time safeguarding the interests of licensed prostitutes.

Technological change has impacted the prostitution market. Cunningham and Kendall (2011) examine the use of the Internet in the prostitution business. They find that online soliciting has displaced some of the off-line market but that the net effect is an increase in supply. They also find that prostitutes soliciting online are largely engaging in lower-risk behaviors than their street-level counterparts, although former street workers may retain their higher-risk behavior also online.

Gaps in the Literature

Scarcity of reliable data is one of the main challenges for the study of prostitution. Collection of

new, reliable data is one of the most worthwhile initiatives with regard to prostitution research. This could be in the form of surveys, anonymous if necessary, but in societies where prostitution is legal, the ambitions should be even higher. For instance, if the framework of mandatory registration of legal prostitutes proposed by Lee and Persson (2013) is adopted, data from those registration procedures should be collected, kept safe, and be available for researchers. In general, regulation will help data collection, due to the fact that one is not collecting data on something criminal and that there will be data emanating from the enforcement of regulations, whether they are about working conditions, health, or taxes.

Even where prostitution is not legal, there are opportunities to collect more and better data and develop better measures, for instance, through the use of list experiments (also called list treatment or the item count technique).

There has been little research on how different forms of criminalization or regulation affect the working conditions of prostitutes. Di Tommaso et al. (2009) use data from the Anti-Trafficking Unit of the International Organization for Migration and find that the well-being of trafficked women is (further) worsened when having to work in secluded spaces.

Criminalization legislation is often the result of an alliance between one side consisting of people who are opposed to prostitution as a matter of principle and another that is concerned about coercion of involuntary prostitutes. The stability of this alliance remains an open question.

Whether demand for voluntary and non-voluntary prostitutes is uniform is an important question that has not been much studied. For instance, if a substantial share of demand is for the voluntary product, that would be valuable to take into account when designing regulation, as shown theoretically by Lee and Persson (2013). Della Giusta et al. (2009a) take some steps toward disaggregating the demand side, but they do not specifically focus on if demanders care about voluntariness.

There are also ways in which the supply side can be disaggregated further. Globalization has made movement across borders easier. Cho

et al. (2013) and Jakobsson and Kotsadam (2013) investigate the relation between sex laws and trafficking, but it would be valuable to know also about the more direct relation between the enforcement of laws against trafficking and coerced labor on the one side and involuntary trafficking on the other.

Topics on which some hold strong moral attitudes can easily become politicized and subject to limited inquiry. So also with prostitution; it is therefore particularly important that researchers in this field are allowed to remain independent when addressing the topics described in this chapter.

Cross-References

- [Crime and Punishment \(Becker 1968\)](#)
- [Gender Diversity](#)
- [Liberty](#)
- [Sex](#)
- [Sex Offenses](#)
- [Shadow Economy](#)
- [Underground Economy](#)

References

- Arunachalam R, Shah M (2008) Prostitutes and brides? *Am Econ Rev* 98(2):516–522
- Cho S-Y, Dreher A, Neumayer E (2013) Does legalized prostitution increase human trafficking? *World Dev* 41:67–82
- Cunningham S, Kendall TD (2011) Prostitution 2.0: the changing face of sex work. *J Urban Econ* 69(3):273–287
- Della Giusta M (2010) Simulating the impact of regulation changes on the market for prostitution services. *Eur J Law Econ* 29(1):1–14
- Della Giusta M, Di Tommaso ML, Shima I, Strøm S (2009a) What money buys: clients of street sex workers in the us. *Appl Econ* 41(18):2261–2277
- Della Giusta M, Di Tommaso ML, Strøm S (2009b) Who is watching? the market for prostitution services. *J Popul Econ* 22(2):501–516
- Di Tommaso ML, Shima I, Strøm S, Bettio F (2009) As bad as it gets: well-being deprivation of sexually exploited trafficked women. *Eur J Polit Econ* 25(2):143–162
- Eldlund L, Korn E (2002) A theory of prostitution. *J Polit Econ* 110(1):181–214

- Jakobsson N, Kotsadam A (2011) Do laws affect attitudes? an assessment of the norwegian prostitution law using longitudinal data. *Int Rev Law Econ* 31(2):103–115
- Jakobsson N, Kotsadam A (2012) Shame on you, john! laws, stigmatization, and the demand for sex. *Eur J Law Econ* 37(3):1–12
- Jakobsson N, Kotsadam A (2013) The law and economics of international sex slavery: prostitution laws and trafficking for sexual exploitation. *Eur J Law Econ* 35(1):87–107
- Lee S, Persson P (2013) Human trafficking and regulating prostitution. New York University School of Law, Law & economics research paper series, working paper no. 12–08

Protective Factors

Norbert Hirschauer¹ and Sebastian Scheerer²
¹Agribusiness Management, Institute of Agricultural and Nutritional Sciences, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

²Institute for Criminological Social Research, Faculty of Economics and Social Sciences, University of Hamburg, Hamburg, Germany

Abstract

Protective factors help people achieve positive outcomes despite exposure to potentially negative influences (in medicine, good health in spite of smoking cigarettes; in the social sciences, productive adult lives as law-abiding citizens despite adverse conditions during childhood and adolescence). In Law and Economics, the concept, in conjunction with control theories, contributes to a better understanding of why economic actors obey the law and comply with rules despite exposure to material incentives to the contrary. Protective factors can result from external sources (social control) and internal sources (internalized conventional values). They generate nonmaterial benefits in the case of compliance and/or non-material costs in the case of noncompliance. The inclusion of protective factor into regulatory analysis – in combination with the consideration of risk factors that work in the opposite direction – helps to understand why some

regulatory regimes work better than others. The concept can be especially useful when it comes to designing regulatory innovation based on a better understanding of why some economic actors obey the law and others don't.

Definition

Protective factors are individual and/or environmental characteristics that reduce people's vulnerability to adversities such as natural, health, or moral hazards.

Protective Factors in Law and Economics and Criminology

Conceptual Background

The concept of *protective factors* owes its existence to the attempt to explain the curious fact that some people experience positive outcomes in the face of highly adverse conditions. If some heavy smokers, for instance, manage to live in good health until their 90th birthdays and beyond, then it is plausible to assume that there is something – some kind of protective factor – that shielded them from contracting any of the diseases commonly associated with inhaling the toxic mixture of over 4,000 chemicals called cigarette smoke. Analogous phenomena of *resilience* can be found in relation to delinquent life trajectories. The quest to explain surprisingly positive life courses of people from dysfunctional families and highly delinquent neighborhoods will reveal characteristics and/or resources that worked in favor of these individuals and guarded them against negative influences. Protective factors may also play a role in the context of *Law and Economics* and, more specifically, rational choice theories. Here, the concept can explain why (some) people obey rules in spite of *moral hazards* and an abundance of profitable and convenient opportunities to break them, i.e., despite an environment with a high degree of adversity to *compliance* with rules.

When using the term “protective factor” in *Law and Economics*, one implicitly resorts to

principal agent theory and adopts the dichotomous perspective of a *principal* (e.g., a lawmaker or regulator) who has defined two types of *agents'* behavior with regard to a rule: noncompliance (undesired behavior) as opposed to compliance (desired behavior). The concept's analytical potential is especially useful when the term "protective factors" is limited to designate nonmaterial factors associated with social control and internalized norms that encourage compliance in an environment where material factors would favor noncompliance. This requirement is met by the following definition:

Protective factors are characteristics in individuals and/or their socio-economic environments that discourage actors from rule-breaking by causing non-material benefits (utility) in the case of compliance and nonmaterial costs (disutility) in the case of non-compliance.

This definition is a useful tool to transcend an all too *narrow rational choice* conception with its restrictive assumption of *utility* hinging exclusively on material wealth. It turns our attention to the fact that not only law works "as a means for changing relative prices attached to individual actions" (Parisi 2004, p. 262), but that there are also nonmaterial factors influencing those "prices" for the individual actor. It corresponds with the understanding that people often pursue multiple goals and strive not only for wealth but also for *social recognition* and distinction as well as for consistency with their *internalized values* and identity (Lösel and Bender 2003; Akerlof and Kranton 2010). Depending on the situation, utility gains from complying with rules may, or may not, outweigh temptations to break them (Pinstrup-Andersen 2005). In other words, the utility of multi-goal decision-makers with *bonds to social norms* is not in all cases monotonically increasing

in expected material wealth. Instead, such actors are prepared to pay an *ethical premium* that shields them from deviance if – but only if – it exceeds the economic temptation to break the rules. The ethical premium is the utility outcome produced by protective factors. It provides an operational definition of resilience which, in turn, is the antonym of *vulnerability*: The stronger an individual's protective factors, the higher her/his ethical premium and the higher (lower) her/his resilience (vulnerability) to moral hazards.

Figure 1 illustrates that we can distinguish four classes of protective factors: On the one hand, there are factors that provide nonmaterial benefits to the individual in case of compliance. These benefits may result from (1) external sources (e.g., social respect) or (2) internal sources (e.g., self-respect). On the other hand, there are factors that cause nonmaterial costs in the case of non-compliance. These costs may also have (3) external sources (e.g., ostracism) or (4) internal sources (e.g., conflict with self-image).

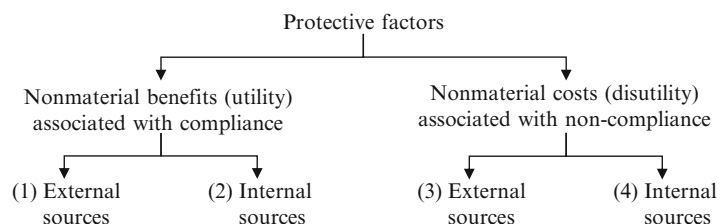
Adopting the perspective that, besides economic wealth, people may seek social respect and inner contentment necessarily requires juxtaposing protective factors with their opposite that has been labeled *risk factors* (and is sometimes also referred to as *criminogenic factors*). Risk factors can be defined as follows:

Risk factors are characteristics in individuals and/or their socio-economic environments that encourage actors to break rules by causing non-material benefits (utility) in the case of non-compliance and nonmaterial costs (disutility) in the case of compliance.

Analogous to protective factors, four classes of risk factors can be distinguished: An individual's integration into a deviant subculture, for instance, may provide (1) nonmaterial external benefits for

Protective Factors,

Fig. 1 A decision-theoretic classification of protective factors



the rule breaker (e.g., respect by deviant peers). A lack of conformity with the values behind the rules (termed conventional in the rest of this essay) or an outright internalization of deviant norms may generate (2) nonmaterial internal benefits for the deviant individual (affirmation of the deviant self). Prominent examples are illicit behavior due to *reactance* (Brehm 1966; Miron and Brehm 2006) and *altruistic rule breaking*. (3) Compliance in such settings, in contrast, may give rise to ostracism by the deviant peer group and (4) cause conflict with the individual's deviant self-image. In brief, nonmaterial costs for obeying rules and nonmaterial benefits for disobeying can be conceived as individual-level consequences of risk factors that would cause illicit behavior even in the absence of economic temptations to break the rules (hence, the alternative label "criminogenic factors").

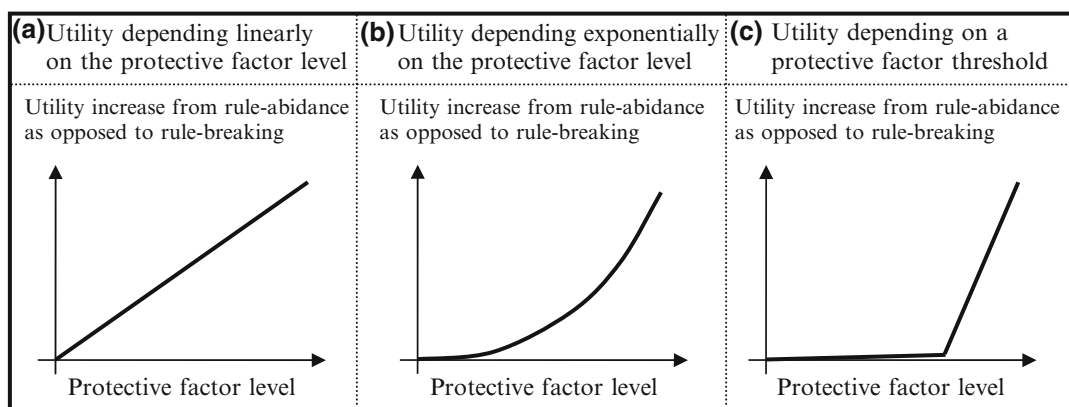
Including the term "risk" into the label that has been attached to the nonmaterial factors encouraging deviance reflects three important conceptual aspects: First, risk factors are conceived as random variables. Second, they impact the probability distribution of the behavioral outcome (i.e., the *behavioral risk*). Third, there is an unambiguous direction of influence: Risk factors increase the probability of the undesired behavior on the part of the agents. While not semantically reflecting the aspect of risk, protective factors share the first two qualities of risk factors: They are random variables and they impact the behavioral risk.

They work in the opposite direction, however, and decrease the probability of the undesired behavior.

In empirical research on how protective factors and risk factors jointly impact behavioral risks, we will encounter three unknowns: the protective factors and risk factors that are at work in a specific context, the level of these factors, and, equally important, the interactions and functional forms that link them to utility. It will thus be doubtlessly difficult to quantify how much utility people derive from nonmaterial sources in given contexts. We know, however, that they will only abstain from profitable rule breaking if their willingness to renounce profits (ethical premium), as resulting from the balance of protective factors and risk factors, exceeds the monetary benefits from illicit behavior.

Adopting a *ceteris paribus* perspective and thus abstracting once again from risk factors (and their interaction with protective factors), Fig. 2 illustrates how protective factors work: An individual's utility for rule abidance as opposed to rule breaking increases monotonically in the level of some upstream variable acting as a protective factor. The precise form of the functional relationship varies depending on factor and context. In the simplest case, it will be linear as depicted in (a). If the functional relationship is to reflect a self-reinforcing mechanism, it will be exponential as depicted in (b). If a threshold needs to be exceeded before a protective factor generates a significant

P



Protective Factors, Fig. 2 Functional relationships linking protective factors to utility and behavior

impact, the function will include a clear discontinuity and may be divided into two linear segments as depicted in (c).

Finally, it should be noted that even a *broad rational choice* conception that includes utility derived from social recognition and inner harmony is often insufficient to explain and predict – with reasonable success – the behaviors of people facing specific rules and specific contextual conditions. Instead, one needs situation-specific models or “minitheories” (Korobkin and Ulen 2000) that facilitate a reconstructing understanding of idiosyncratic or group-specific decision-making processes in particular real-life contexts. Such models do not have to be isomorphic with respect to the “objective” rules of the world but to the world as subjectively perceived and evaluated by the individual or specific subpopulation under consideration (Rubinstein 1991). In other words, one needs models that tackle the gap between narrow rational choice predictions and actual behavior (Garoupa 2003). Such models need to reflect behavior as the result of *multi-goal* decision-making by (potentially) rule breaking and *bounded rational* actors (cf. Simon 1957; Gigerenzer and Selten 2001) who *subjectively* form expectations and evaluate likely outcomes at the time they make their choices.

Levels of Analysis

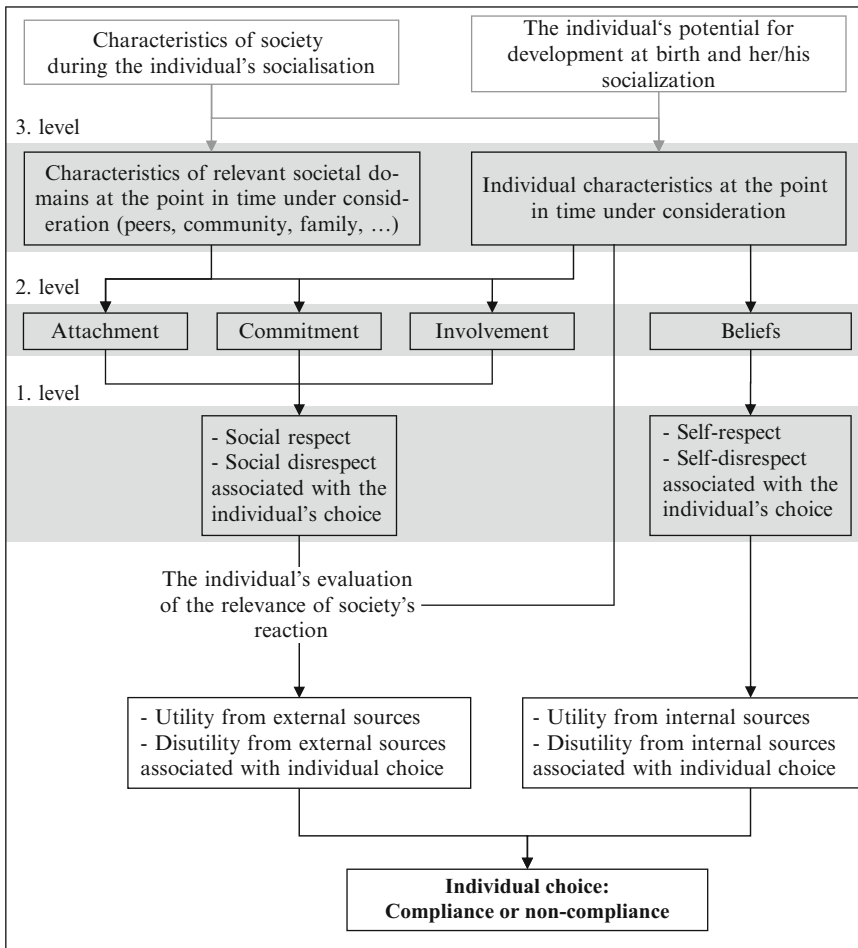
The decision-theoretic definition of protective factors provided above facilitates an operational understanding of their utility outcomes and a clear separation from risk factors. Focusing on the utility outcomes of the individuals under consideration, however, does not yet tell us which real-life characteristics produce these outcomes and thus make people resilient to the seductive potential of illicit behavior. Figure 3 describes the different levels of the *cause-effect chain* on which these factors can be found.

On a first upstream level in the cause-effect chain, we can distinguish external causes of utility (social respect and recognition, social disrespect and ostracism) from internal causes (self-respect and self-affirmation, self-disrespect and self-loathing). Regarding external utility sources, a

conceptual distinction between a given level of respect/disrespect shown by others and the individual’s evaluation of that level may be useful. It does not seem to make sense, however, to attempt an operational distinction between a given level of one’s own self-respect/self-disrespect and one’s own evaluation of that level. At best, it would require a schizophrenic separation of the self in two parts.

On a second upstream level of the cause-effect chain, four major types of factors acting as protective mechanisms and determining the factors on the first level have been distinguished (Hirschi 1969): first, social integration into conventional society and emotional bonds to relevant others (elders and peers) who cherish rule abidance and repudiate rule breaking (*attachment*); second, hostage posting to conventional society through reputation as an honest business person or holder of public office that has been accumulated from earlier “investments” in conventional life and that is lost if outlawed behavior is disclosed (*commitment*); third, engagement in conventional activities that consume time and mental energy and thus help the individual abstain from the often complex and time-consuming actions of rule breaking (*involvement*); and fourth, identification with the conventional system of values that are “behind the rules,” including the recognition of authority, the emulation of role models, and an appreciation of rule abidance in itself (*beliefs*).

Going further back the causal chain, attempts have been made to systematize protective factors “behind” the factors on the second level. They all list and classify variables believed to be positive influences still further upstream. Commonly used classifications categorize these upstream factors into domains such as *peers*, *community*, *family*, and *individual*. Accordingly, variables such as long-lasting relationships and trust in the (business) community, a functioning social fabric in neighborhoods and families, low existential stress, and emotional stability are listed as protective factors. All these compilations face the difficult task to avoid gaps, duplications, and inconsistent listings of variables that, while representing behavioral correlates such as gender



Protective Factors, Fig. 3 Protective factors: cause-effect chain and levels of analysis

or ethnic affiliation, are either not causes of behavior in the sense of this word or situated on different levels of the cause-effect chain. As indicated in Fig. 3, one could go even further back in time and causation and compile protective factors on still further upstream levels. Whatever the focal level of analysis, inconsistent listings are futile descriptive exercises with little practical potential to indicate the variables that preventive strategies could, and should, tackle to produce more compliant behavior.

Relation to Control Theories

The decision-theoretic concept of protective factors can be linked to the view of human nature in criminological *control theories* (cf. Hirschi 1969;

Gottfredson and Hirschi 1990; Tittle 2000). Contrary to conventional criminological theories, which used to explain crime and deviance by searching for criminogenic factors in the delinquent subject and her/his environment, control theories contend that it is not the breaking of rules that needs explanation in the first place but rather compliance. In the view of control theorists, humans are basically free to deviate from rules following their own interest. It is therefore the challenge of every society to gradually and effectively “deprive” individuals of their original freedom to deviate by creating bonds that make it increasingly difficult for a well-socialized member of society to act against social and legal restraints.

In their attempts to understand the processes by which societies deprive people of their original freedom to deviate, control theorists analyze how social bonds are created through both the attachment to others and the internalization of value systems in line with social expectations. While the former generates external (social) control and rewards/sanctions, the latter leads to internal (self-)control and rewards/sanctions. If society fails to create sufficient social bonds, man’s self-interest and original freedom to deviate, in conjunction with economic temptations, continue to induce rule breaking. Table 1 summarizes the core concept of control theories using an illustrative numerical example.

Our point of departure is an individual who faces a misdirected incentive of –80 (economic temptation) to break some rule. Due to protective factors – both in the form of bonds to relevant others and internalized values – the individual is prepared to pay an ethical premium totaling 55. Since this premium does not exceed the economic temptation, we would conclude that the individual is not sufficiently shielded from the natural inclination of self-interested rule breaking. Reflecting the basic concern of control theories with the gaps

and holes in the system of social bonding and control, the numerical example in Table 1 is based on the assumption that deviant behavior (e.g., *white-collar crime*) is caused by economic temptations that coincide with lacking protective factors rather than by risk factors. This is why nonmaterial cost from obeying and nonmaterial benefits from disobeying, while being mentioned, are assumed to be zero.

Assuming risk factors to be inexistent may be an adequate analytical simplification in some situations; it may be completely inadequate in others. The degree of resistance to a given material temptation results, in principle, from the balance of protective factors and risk factors. An individual may, for instance, simultaneously be a member of social groups with conflicting value systems. As a result, she/he will strive not only for the social recognition within conventional society but also for the respect of deviant elders and peers, thus possessing external protective factors and external risk factors at the same time. Even two souls may be dwelling in an individual’s breast. An example would be a generally law-cherishing individual who, for some reason or another, feels an internal pressure to engage in altruistic rule

Protective Factors, Table 1 An illustrative example for the concept of control theories^a

	Utility for obeying		Utility for disobeying		Utility differential
Material benefits (+) and costs (–)		+700		+780	–80^b
External benefits (+) and costs (–) from social respect and disrespect in conventional society	<i>Protective factors</i>	+10	<i>Protective factors</i>	–25	+35 ^c
External benefits (+) and costs (–) from social respect or disrespect in deviant subculture	<i>Risk factors</i>	0	<i>Risk factors</i>	0	0
Internal benefits (+) and costs (–) from consistency and conflict with internalized conventional values	<i>Protective factors</i>	+5	<i>Protective factors</i>	–15	+20 ^d
Internal benefits (+) and costs (–) from reactance or internalized values of deviant subculture	<i>Risk factors</i>	0	<i>Risk factors</i>	0	0
Expected benefits (+) and costs (–) from nonmaterial sources		+15		–40	+55

^aFor risk-averse actors who deduct risk premiums from risky benefits and add risk premiums to risky costs, the figures are to be conceived as units of utility. For risk-neutral actors, they can be directly interpreted as payoffs and payoff equivalents.

^bThe negative sign of the differential indicates a misdirected incentive (economic temptation).

^cEthical premium caused by external sources.

^dEthical premium caused by internal sources.

breaking in a certain context. Another example would be a situation of reactance in which an actor finds inner satisfaction by breaking an imposed rule while simultaneously feeling uneasy or even guilty for breaking the law. In both cases, internal protective factors and internal risk factors are at work at the same time. While risk factors are conventionally considered in criminological research on juvenile delinquency and illicit drug use, they are often omitted from the analysis of economic deviance. However, reactance, deviant cultures, and the internalization of deviant norms can be important in business contexts too. Unless we have contrary contextual knowledge, risk factors should thence not be *a priori* ignored in the analysis of economic deviance.

The adequacy of a behavioral model depends on the context and the level of analysis (cf. Hess and Scheerer 2004). On the *macro-level* where one considers an industry's response to (a change in) rules, a narrow rational choice approach with its exclusive focus on economic incentives may, or may not, facilitate a realistic view. For an analysis on the *meso-level* of company decisions made by respected members of the business community, applying the rationale of control theories and abstracting from risk factors can be adequate but may not always be so. There may be more causes of white-collar crime, for instance, than just economic temptations and gaps in social bonds and social control (Katz 1988, pp. 310–324). *A priori* disregarding risk factors, however, is most plausibly inadequate not only when trying to understand juvenile delinquency but also in a *micro-level* analysis of individual employees who may face strong group cohesion and deviant corporate subcultures with informal social pressures and rewards to bend or break the law.

The extension of the narrow rational choice conception provided by the inclusion of protective factors and risk factors into the set of potential behavioral determinants can be related to Ostrom's (2005) "Institutional Analysis and Development Framework." Adopting an institutional economics focus on rule-governed social life, Ostrom attaches the label *delta parameters* to the nonmaterial benefits and costs that an

individual derives from her/his choices. That is, besides the utility derived from economic incentives (payoffs), emphasis is put on the utility (payoff equivalents) that people gain from external sources in terms of social approval or disapproval (positive and negative external deltas) and internal sources in terms of affirmation of, or conflict with, internalized values and self-image (positive and negative internal deltas). Given the obvious communication problems caused by the lack of a generally acknowledged terminology between the various behavioral science disciplines (here as well as in many other cases), one can but agree with Ostrom's own words, "If every social science discipline or subdiscipline uses a different language for key terms and focuses on different levels of explanation as the 'proper' way to understand behavior and outcomes, one can understand why discourse may resemble a Tower of Babel rather than a cumulative body of knowledge" (Ostrom 2005, p. 11).

Prevention of Rule Breaking

How people respond to rules is one of the core concerns of *Law and Economics* and regulatory theory. First of all, regulators have to understand existing compliance problems as resulting from the joint effects of economic incentives, protective factors, and risk factors that are at work at a given point in time. Regarding prevention and the *management of behavioral risks*, a reconstructing understanding of the status quo is but a starting point, however. To successfully manage behavioral risks and reduce the probability of first-time rule breaking as well as of recidivism, regulators need to identify effective and cost-efficient regulatory strategies. Their success in this endeavors – be they directed at certain target groups within risk-based preventive schemes (secondary prevention) or at individual offenders within personalized reintegration approaches (tertiary prevention) – depends on the ability to forecast behavioral changes that are likely to be brought about by hypothesized regulatory action.

While the protective factor concept is geared toward the analysis of compliance with given rules, one should keep in mind that, depending on the type of the behavioral determinant they are

primarily aimed to impact, four ideal types of regulatory regimes can be distinguished in the first place:

1. *Hierarchical command and control*, such as tight law enforcement, impacts the set of choices that are available to regulatees (e.g., through license withdrawal) as well as their relative economic competitiveness (e.g., through sanctioning).
2. *Incentive-oriented regimes*, such as steering taxes and payments handed out to those who undertake specified actions, increase the relative economic competitiveness of the socially desired behavior without imposing mandatory rules.
3. *Value-oriented regimes*, such as the promotion of corporate social responsibility and professional ethics, are based on persuasion and finally aimed to increase protective factors that shield regulatees from socially undesired behavior.
4. *Human capacity building*, such as counseling and training, is to enhance the actors' abilities to perform socially desired behavior.

The law- and control-based systems that criminology is traditionally concerned with can clearly be associated with the first regulatory type. However, even within regimes based on mandatory rules, regulators can go beyond command and control and target the full set of behavioral determinants to produce compliant behavior. Accordingly, regulatory enforcement can be related to the three basic regulatory strategies known in criminology: (1) *incapacitation and target hardening*, (2) *deterrence*, and (3) *accommodation* (Picciotto 2002). Two definitional settings are required for this relational positioning: First, we need to equate sanctioning and deterrence not only with punitive measures but with the production of incentives that reduce the relative competitiveness of rule breaking (including damage compensations, recall costs, contractual fines, litigation costs, losses of sales, etc.). Second, we need to understand accommodation as not only addressing personal values but including other ways of human capacity building.

Regulatory and criminological scholars from different normative schools attribute different degrees of importance to material and nonmaterial

motivations. Some believe in deterrence only; others believe that accommodation and gentle persuasion (i.e., the promotion of protective factors) work to secure compliance (Ayres and Braithwaite 1992). Contrary to this antithetic pairing, evidence from a wide range of fields – from occupational safety (Scholz and Gray 1990) over standards in nuclear plants (Rees 1994) to tax compliance (Braithwaite 2009) – indicates that successful strategies avoid the dysfunctional effects of pure deterrence, such as defiance and reactance, and the negative effects of pure accommodation and leniency, such as a blurring of standards. Successful strategies are able to simultaneously reduce economic temptations and strengthen bonds to social norms by generating *value correspondence* between regulator and regulatees.

Focusing on tertiary prevention and the reduction of recidivism, Braithwaite (2002) provides a related systematization of regulatory strategies and stresses that they need to be combined and applied in a systematic progression. He distinguishes *restorative justice coaching* (persuasion/counseling) from increasing levels of *deterrence* (warning letters → civil penalties → criminal penalties) and finally from different levels of *incapacitation* (license suspension → license revocation). Accounting for the pros and cons of persuasion as opposed to the various degrees of punishment, he advocates that one should attempt to reintegrate offenders into the rule-abiding society by using transparent and graduated responses to misconduct, with escalating/de-escalating measures contingent on the regulatee's degree of bad/good conduct (*responsive regulation*). According to his *enforcement pyramid*, noncompliance should be met with a clear disapproval of the fact and increasingly punitive measures, but regulators should always start softly by using the cooperative measures of persuasion and counseling aimed at reintegrating offenders into the law-abiding community. This is not to be confused with naïve trust. According to the concept of responsive regulation, the harsher the available ultimate sanctions, the more likely compliance will be achieved through persuasion. The recommendation to regulators would therefore be to “speak softly, while carrying very big sticks” (Ayres

and Braithwaite 1992, p. 40, paraphrasing a well-known quotation from Theodore Roosevelt).

Technically speaking, the responsive regulation approach moves away from an overly simplistic partial perspective according to which one might ostensibly conclude, for instance, that the more controls and sanctions, the higher the actors' inclination to abide by the rules. Instead, responsive regulation represents a holistic approach geared to generate and exploit positive interactions between nonmaterial and material behavioral determinants and, above all, to avoid so-called ironic or perverse effects of sanctions and control, including *self-fulfilling prophecies of distrust* (Luhmann 2000). It can thus be related to the concern with motivation in law and behavioral science. When regulators introduce new measures to steer behavior, be they new rules or new enforcement approaches, a *ceteris paribus* assumption with an exclusive focus on economic incentives is hardly adequate to predict the behavioral impacts. This is due to the fact that the following interactions may occur (cf. Frey 1997):

- Even primarily incentive-oriented regulatory action – be it based on monetary rewards for desired or on sanctions against undesired behavior – may have a positive impact on internal protective factors as well. This will be the case, for instance, if incentives/disincentives and accompanying controls are being appreciated by the individual as a clear affirmation of what is right and wrong in a society. Such desirable synergies of regulatory action have been termed *motivation crowding in*.
- Regulatory change such as tightening of controls and/or an increase of sanctions may reduce the economic temptation to break rules but impair internal protective factors. In such a case, the positive impact on the behavior in question will be lessened, and the desired motivational change could even be reduced to (or below) zero. Such dysfunctional dynamic shifts have been termed *motivation crowding out*.
- Regulatory change aimed at reducing economic temptations might even generate risk factors. When regulatees experience a new regulation as an illegitimate interference with their basic

values and/or freedom of action, they may derive a genuine nonmaterial internal benefit from standing up against the new rule. This reactance makes them even accept an economic disadvantage of rule-breaking (a *negative ethical premium*) in exchange for the satisfaction of asserting their autonomy in the face of what they see as an outrageous interference.

Frey and Jegen (2001, p. 590) describe motivation crowding out as “one of the most important behavioral anomalies in economics, as it suggests the opposite of the most fundamental economic ‘law,’ that raising monetary incentives increases supply.” One would certainly have to integrate reactance into this statement. The *behavioral anomaly* obviously dissolves as soon as one adopts a more realistic and broad utilitarian view instead of maintaining narrow rational choice assumptions. Regulators should realize that regulatory change may cause interconnected impacts and consequently search for strategies that are “smart” in that they avoid dysfunctional effects and, at best, generate synergies. The conception of economic man underlying the change from the famous *get incentives right* to the more adequate *get utilities right* – as subjectively expected by those who are subject to rules – is the key to the understanding of what behavioral economic analysis and the regulatory issue are essentially about (Hirschauer et al. 2012).

Last not least, looking beyond the enforcement of given rules, it should be noted that producing compliant behavior makes only sense if the rule itself makes sense. If a rule is not meaningful in the first place (i.e., if the social costs associated with compliance exceed the social benefits of compliance), then even the most effective production of compliance will not contribute to a positive outcome in the real world. If a rule makes sense, however, adverse social outcomes resulting from malpractice can be conceived as *negative externalities* that are caused by the breaking of rules that had been designed to prevent them in the first place. In this latter case, we can indeed limit our concerns to the question of how society can cost-efficiently make individuals abide by its (presumably social efficient) rules.

Cross-References

- [Criminal Sanctions and Deterrence](#)
- [Externalities](#)
- [Rationality](#)

References

- Akerlof G, Kranton R (2010) *Identity economics*. Princeton University Press, Princeton
- Ayres I, Braithwaite J (1992) *Responsive regulation: transcending the deregulation debate*. Oxford University Press, New York
- Braithwaite J (2002) *Restorative justice and responsive regulation*. Oxford University Press, New York
- Braithwaite V (2009) *Defiance in taxation and governance: resisting and dismissing authority in a democracy*. Edward Elgar, Northampton
- Brehm JW (1966) *A theory of psychological reactance*. Academic, New York
- Frey BS (1997) *Not just for the money: an economic theory of motivation*. Edward Elgar, Cheltenham
- Frey BS, Jegen R (2001) Motivation crowding theory: a survey of empirical evidence. *J Econ Surv* 15:589–611
- Garoupa N (2003) Behavioral economic analysis of crime: a critical review. *Eur J Law Econ* 15:5–15
- Gigerenzer G, Selten R (eds) (2001) *Bounded rationality: the adaptive toolbox*. MIT Press, Cambridge
- Gottfredson MR, Hirschi T (1990) *A general theory of crime*. Stanford University Press, Stanford
- Hess H, Scheerer S (2004) *Theorie der Kriminalität*. In: Oberwittler D, Karstedt S (eds) *Kölner Zeitschrift für Soziologie und Sozialpsychologie. Sonderheft 43. Soziologie der Kriminalität*, Wiesbaden, pp 69–92
- Hirschauer N, Bavorová M, Martino G (2012) An analytical framework for a behavioural analysis of non-compliance in food supply chains. *Brit Food J* 114:1212–1227
- Hirschi T (1969) *Causes of delinquency*. University of California Press, Berkeley
- Katz J (1988) *Seductions of crime*. Basic Books, New York
- Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. *Calif Law Rev* 88:1051–1144
- Lösel F, Bender D (2003) Resilience and protective factors. In: Farrington DP, Coid J (eds) *Prevention of adult antisocial behaviour*. Cambridge University Press, Cambridge, pp 130–204
- Luhmann N (2000) *Vertrauen*, 4th edn. Lucius & Lucius, Stuttgart
- Miron AM, Brehm JW (2006) Reactance theory: 40 years later. *Z Sozialpsychol* 37:9–18
- Ostrom E (2005) *Understanding institutional diversity*. Princeton University Press, Princeton
- Parisi F (2004) Positive, normative and functional schools in law and economics. *Eur J Law Econ* 18:259–272
- Picciotto S (2002) Introduction: reconceptualizing regulation in the era of globalisation. In: Picciotto S, Campbell D (eds) *New directions in regulatory theory*. Blackwell, Oxford, pp 1–11
- Pinstrup-Andersen P (2005) Ethics and economic policy for the food system. *Am J Agric Econ* 87:1097–1112
- Rees JV (1994) *Hostages of each other: the transformation of nuclear safety since three Mile Island*. Chicago University Press, Chicago
- Rubinstein A (1991) Comments on the interpretation of game theory. *Econometrica* 59:909–924
- Scholz JT, Gray WB (1990) OSHA enforcement and workplace injuries: a behavioral approach to risk assessment. *J Risk Uncertainty* 3:283–305
- Simon HA (1957) *Models of man: social and rational*. Wiley, New York
- Tittle CR (2000) Theoretical developments in criminology. *Crim Justice* 1:51–101

Further Reading

- Braithwaite J (2008) *Regulatory capitalism: how it works, ideas for making it work better*. Edward Elgar, Northampton
- Coleman JS (1987) Norms as social capital. In: Radnitzky G, Bernholz P (eds) *Economic imperialism. The economic method applied outside the field of economics*. Paragon House Publisher, New York, pp 133–155
- Englerth M (2010) *Der beschränkt rationale Verbrecher: Behavioral Economics in der Kriminologie*. LitVerlag, Berlin
- Thomas N, Baumert A, Schmitt M (2012) Justice sensitivity as a risk and protective factor in social conflicts. In: Kals E, Maes J (eds) *Justice and conflicts: theoretical and empirical contributions*. Springer, Berlin/Heidelberg, pp 107–120
- Tittle CR (1995) *Control balance. Towards a general theory of deviance*. Westview Press, Boulder
- Tyler TR (1990) *Why people obey the law*. Princeton University Press, Princeton

Public Choice: The Virginia School

Rosolino A. Candela

Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

Abstract

Public Choice, broadly defined, is the economic analysis of nonmarket decision-making. My primary focus in this chapter will be on the central importance of the “Virginia School” of

Public Choice, specifically in two aspects. First, Public Choice was central to countering the presumption of market failure among economists, beginning in the 1950s and 1960s. Second, yet less emphasized, is the role in which Public Choice played in analyzing the economics of anarchy, beginning in the 1970s. Although seemingly disconnected, both aspects of the Virginia School are linked by the fundamental question of political economy: Under what institutional conditions can social order emerge unintendedly from the self-interest of individuals? Therefore, in both respects, the Virginia School has been central to the rearticulation of 19th political economy during the twentieth and twenty-first centuries.

Introduction

Public Choice, broadly defined, is the economic analysis of nonmarket decision-making. More specifically, James Buchanan, who cofounded the field of Public Choice with Gordon Tullock, referred to Public Choice as “politics without romance” (Buchanan 1979). Public Choice first emerged as a discipline within the broader tradition of political economy at the Thomas Jefferson Center for Studies in Political Economy at the University of Virginia (UVA), which Buchanan co-founded in 1957 together with G. Warren Nutter. Since then, several other branches of Public Choice have emerged. These include the “Bloomington School” of Vincent and Elinor Ostrom; the “Chicago School” developed by Gary Becker, Sam Peltzman, and George Stigler; and the “Rochester School” of William Riker. To compare and contrast each of these branches of Public Choice, let alone provide a comprehensive survey of the entire field, goes beyond the scope of this entry (see Mitchell 1988 and Mueller 1976).

My primary focus in this chapter will be on the central importance of the “Virginia School” of Public Choice, specifically in two aspects. First, more than any other branch of public choice, its explicit motivation “is the rejection of all of the pillars of the Samuelsonian revolution” (Boettke

and Marciano 2015: 54), which had begun to emphasize the tendency for markets to be a sub-optimal mechanism of resource allocation. By applying the logic of economic decision-making to political settings, Public Choice economists countered the Samuelsonian presumption of market failure by pointing out that governments may also be suboptimal in their attempt to address market failures. Second, yet less emphasized, is the role in which Public Choice played in analyzing the economics of anarchy. Beginning in the 1970s, the Virginia School began to analyze the mechanisms by which institutions can emerge to define and enforce property rights and contractual arrangements without government. This has led to a flowering of theoretical and historical case studies providing evidence of social cooperation without government.

At first glance, these two aspects of the Virginia School seem to be disconnected. However, both the study of government failure and anarchy are flips sides of the same question grounding political economy: Under what *institutional* conditions can social order emerge unintendedly from the self-interest of individuals? Therefore, in both respects, the Virginia School has been central to the rearticulation of 19th political economy during the twentieth and twenty-first centuries. In both the study of government failure and anarchy, the uniqueness of the Virginia School has been its ability to “marry the property-rights, law-and-economics, public-choice, Austrian subjectivist approaches” of economics (Buchanan 2015: 260). For the remainder of this chapter, I will trace out the origins and contributions of the Virginia School to the broader tradition of political economy.

A Brief Overview of the Virginia School

The origins of Public Choice can be traced back to Adam Smith and the classical political economists of the nineteenth century. What all branches of Public Choice, in particular the Virginia School, shared with classical political economy is their comparative analysis of alternative processes of decision-making and their respective results under

alternative institutional arrangements, particularly between market and nonmarket institutional settings. Public Choice, and the Virginia School in particular, bases its analysis of political decision-making on three basic assumptions: (1) rational choice; (2) methodological individualism; and (3) politics as an exchange process.

To say that individuals choose rationally implies simply the following: When an individual faces a set of alternatives, he or she will choose the alternative they expect to give them the greatest satisfaction. In other words, the pursuit of one's rational self-interest implies simply that individuals choose more rather than less of whatever they prefer, or that the individual *strives* to maximize utility. Moreover, Public Choice assumes *behavioral symmetry*, meaning that individuals are utility maximizers in both market and political settings; the only difference between market and political actors is the manner in which utility-maximizing behavior manifests itself. For example, in the marketplace, firm owners strive to maximize monetary profit, but in nonprofit setting, such as politics, there are four groups of decision-makers, each of which are maximizing utility as well. They include voters and interest groups, who want goods and services from elected officials, politicians (i.e., elected officials), who strive to maximize votes and financial support from interest groups and voters, and bureaucrats, who wish to maximize their budgets from elected officials (Simmons 2011: 52).

Unlike in the market place, where individuals express their preference by "voting" with their dollars for different goods sold simultaneously by competing firms, the same individual in politics must vote on a single, bundled policy, which constitutes many different goods, "sold" to them by political officials they've elected. Given the bundled nature of public policy, and its implementation, voters will be rationally ignorant, given the high costs of gathering information about the different "goods" in a policy bundle, which may include defense, health care, import tariffs, etc. Moreover, the bundle nature of public policy also implies that individuals will prefer different "goods" *within the bundle* with different intensities. For example, given the concentrated benefit

from implementing tariffs on imported sugar, domestic sugar producers will no doubt express greater preference for such a policy, in the form of votes as well as campaign contributions to elected officials, unlike the common voter, who will tend to remain relatively ignorant of such a policy. Therefore, the logic of political decision-making will be for vote-maximizing politicians to concentrate benefits on well-organized and well-informed interest groups, which include budget-maximizing bureaucracies, and disperse costs on unorganized and ill-informed voters. Such a political outcome is rational, once we've traced such an outcome back to the choices made by utility-maximizing individuals under a nonmarket setting, where individuals do not bear the full costs and benefits of their decision-making. This brings us to methodological individualism.

By rejecting an organic holistic vision of the state, Public Choice employs methodological individualism, meaning that the outcomes of collective action are traced back to the choices and interactions of rational, utility maximizers. This does not imply that individual preferences can be aggregated into a social welfare function, which the state maximizes on behalf of the electorate. Rather, collective action under a methodologically individualist view will take place if two or more individuals find it mutually beneficial to accomplish certain common purposes jointly with others, rather than separately through bilateral market interactions, an example being the drainage of a mosquito-infested swamp (Buchanan 1964: 219–220). This example reveals that, in the Virginia School, social phenomena are not simply the result of individuals passively responding to constraints.

Unlike in a constrained maximization problem, defined by fixed constraints, Public Choice theorists of the Virginia School direct their analytic attention to *choice among constraints*, where politics is the artifact of an exchange process, whereby individuals strive to agree to mutually beneficial rules that constrain their behavior in a Pareto-optimal manner. The Virginia School political economists therefore take a *constitutional perspective*, which focuses on analyzing "the rules of game" that governs political

interaction. Relevant political choices, according to the study Constitutional Political Economy, are not choices among alternative distributions or allocations of resources within a set of political rules. Rather, it postulates that relevant political choices are among an alternative set of political rules that generate different patterns of distribution and allocation of resources. Constitutional Political Economy is tied to Public Choice, broadly speaking, in two respects. Assuming rational self-interest on the part of political officials, alternative decision-making rules, such as majority rule or unanimity, will generate different policy outcomes within those sets of rules. For example, requiring more than majority rule for road repair will tend to reduce public expenditure on roads, since more people will have to be included in the political exchange. Increasing the number of individuals as a decision-making rule will consequently increase the cost to the beneficiaries of road repair, and, therefore, low costs will be spilled over and dispersed to parties not agreeing with the exchange. Second, political reform cannot occur by changing the “players” within the political game, but only by changing the rules by which the game is played. It is no accident that the study of Public Choice, and the Virginia School, in particular, emerged when the rule level of analysis began to recede to the background of political and economic analysis.

The birth of the Virginia School cannot be understood outside the historical context it inherited. The Samuelsonian revolution changed the tacit presupposition of economists with regard to the government’s role in a market economy, from a “laissez-faire presumption” to a “market failure presumption” (Boettke and Lesson 2015: xiii–xviii). This presumption can best be understood as *epistemological*, rather than political or ideological. Under the laissez-faire presumption, the analytical description of the market economy was complemented by a default presumption to limit government’s role in the marketplace, generally speaking, to the protection of private property and contract enforcement. However, under the market failure presumption, economists assume that the presence of macroeconomic instability, monopoly power, externalities, and public

goods will merit, by default, government correction of the market’s allocative process. The distinction here is not simply a presumption of “more government” or “less government,” but fundamentally a presumption of what knowledge government officials possess to correct market failures. Buchanan best states this distinction as follows:

The classical (Smithean) argument for control (or depoliticization) and the welfare economists’ argument for control (for politicization) are on all fours *only if we presume the existence of the same underlying evaluative standard in the two cases*. To suggest, with the welfare economists, that market failure supports politicization, there must be not only departures from the necessary conditions for efficiency, but also some presumption that political action is informed by a knowledge of what the allocatively efficient solution is, quite apart from the operation of politics itself. By contrast, to suggest, with Adam Smith, that regulatory failure supports market liberalization does not require any presumptive knowledge about what particular outcome is likely to produce maximal value. *There is a categorical epistemological difference between the two comparative exercises, a difference that many modern economists still do not understand*. (emphasis added, Buchanan 1996 [2001]: 292)

Central to this “market failure presumption” and the Samuelsonian paradigm was the presumed existence of an objective social welfare function, which government officials would maximize by delivering an optimal amount of public goods, eliminating monopoly power and externalities, and establishing macroeconomic stability in the economy through aggregate demand management. From this context, Buchanan, Tullock, and the Virginia School established Public Choice, based on the notion of *homo economicus* and politics as an exchange process.

Public Choice and Government Failure

Beginning in the late 1940s and early 1950s, Public Choice began to counter the presumption of market failure by challenging the existence of an objective social welfare function. In his 1949 article, “The Pure Theory of Government Finance: A Suggested Approach,” Buchanan contrasted between an “organismic” theory of

the state and an “individualistic” theory of the state. The basis of this contrast would not only define Buchanan’s intellectual trajectory (Wagner 2017), but also that of Public Choice as well. According to the organismic theory, the state is considered as a single decision-making unit, or “a fiscal brain” as Buchanan puts it, which acts for society as a whole by seeking to maximize “general welfare” or “social utility” (1949: 497). In the individualistic theory of the state, the state embodies no ends other than those of individuals in society, who find it in their self-interest to pursue a certain portion of their wants collectively. The individualistic theory typifies the nature of interaction between individuals and the state as an “exchange” of government services paid out of the economic resources of individuals. The basis of the individualistic theory of the state, and the Public Choice perspective, is an extension of two related aspects from economic theory: (1) based on the incentive structure of an institutional setting, individuals will act according to their own self-interest (i.e., *homo economicus*); and (2) individuals pursue their self-interest by engaging in mutually beneficial exchange. By applying the behavioral postulate of *homo economicus* and the principle of mutually beneficial exchange to political decision-making, Public Choice challenges the main pillars of the Samuelsonian paradigm by arguing threefold: (1) there is no objective welfare function that a benevolent despot maximizes; (2) even if such a social welfare function existed, only individuals choose, not “society” or “the state”; and (3) individuals acting in a market setting or a political setting will pursue their self-interest based on their subjective assessment of costs and benefits (Boettke 2012: 249).

In countering the paradigm of Samuelsonian economics, Public Choice emerged “as a ‘theory of government failure’ that offsets ‘the theory of market failure’ that emerged from theoretical welfare economics” (Buchanan 1983 [2000]: 113). Under market failure theory, perfect competition became the normative benchmark from which the inefficiencies of real-world market outcomes were compared. Any deviations from perfectly competitive equilibrium, such as the presence of public

goods, externalities, monopoly power, or asymmetric information, imply government intervention as a public policy prescription to correct for such market failures. Consistent with the evaluative standard of perfect competition, political actors are presumed to possess perfect information to address market failures. However, this was a categorical epistemological difference in understanding market processes as well as political processes, not only for the Virginia School, but also for economists prior to the mid-twentieth century. Ronald Coase, a pioneer in the field of law and economics, and one of the early faculty members of the Thomas Jefferson Center at the UVA, best articulated this presumption of government failure:

This “novel theory” (novel with Adam Smith) is, of course, that the allocation of resources should be determined by the forces of the market rather than as a result of government decisions. Quite apart from the malallocations which are the result of political pressures, an administrative agency which attempts to perform the function normally carried out by the pricing mechanism operates under two handicaps. First of all, it lacks the precise monetary measure of benefit and cost provided by the market. Second, it cannot, by the nature of things, be in possession of all the relevant information possessed by the managers of every business. . . to say nothing of the preferences of consumers for the various goods and services (Coase 1959: 18).

Rather than view economics in terms of states of equilibrium, the Virginia School has always regarded economics as a science of exchange and the institutions within exchange takes place. By anchoring political choices in price theory and methodological individualism, Buchanan, Tullock, and the Virginia School as a whole were able to turn the presumption of market failure on its head, namely, by marrying the best insights of the property-rights economists, law-and-economics, and Austrian subjectivism. For example, following the subjectivism of Austrian economists Carl Menger, Ludwig von Mises, and F.A. Hayek, Buchanan saw the notion of cost as inherently tied to the act of individual choice and understood a subjective assessment of trade-offs by individuals if it would have any meaning in a theory of decision-making (Buchanan 1969 [1999]; see also Buchanan and

Thirby 1981). Although a seemingly elementary insight into price theory, its public policy implications for market failure theorists are devastating. For example, in the face of externalities, such as pollution, market failure economists would propose, à la Pigou, the use of taxes so that individual polluters would bring the full social costs of polluting into their decision-making. However, returning to the earlier quote by Coase, this presumes that political actors have both the incentives as well as the knowledge to address these market failures in accordance with ideal conditions of general competitive equilibrium. However, under such conditions, such a policy is either possible and redundant, or impossible to set because the institutional conditions presupposed for their establishment either eliminate their necessity or preclude the availability of the knowledge necessary to calculate the correct tax to levy.

In the field of public finance, Buchanan challenged the Samuelsonian orthodoxy in fiscal policy by questioning the incentives that political actors face in balancing budgets, namely by demonstrating the intergenerational transfer of government debt (Buchanan 1958 [1999]). In his 1948 edition of his seminal text *Economics*, Paul Samuelson argued that the “interest on internal debt is paid by Americans to Americans; there is no *direct* loss of goods and services” (Samuelson 1948: 427). In other words, government debt is not a burden because “we owe it to ourselves.” By disregarding the notion of a social welfare function, let alone its maximization, what Public Choice shows is that political actors will only assess costs as subjective trade-offs in the maximization of their own utility functions. Understood this way, politicians will only have the incentive to run ever increasing budget deficits, not to smooth government spending over the business cycle. This is because the economic logic that political officials face in their decision-making is to concentrate benefits among small and well-organized interest groups in their constituency, and disperse costs among the ill-informed masses of the population. Alternatively speaking, if we regard fiscal responsibility as a public good, political actors face a free-rider

problem in the elimination of budget deficits, namely because of the cost of eliminating a spending program will be concentrated on a specific interest group, while the benefits of eliminating fiscal irresponsibility will only be dispersed throughout the population. Given the absence of property rights in a political institutional setting, it only makes political sense that a “tragedy of the fiscal commons” will prevail (see Wagner 2012). Thus, it is the interest of political actors to shift the debt burden to future generations. This fiscal insight, however, has broader public policy implications. Given the logic of political decision-making, government attempts to address alleged market failures, not only in the form of macroeconomic instability, but also in the presence of monopoly, externalities, and public goods, are prone to rent-seeking and regulatory capture for the private benefit of special interest groups, which also include government bureaucracies who benefit from the expansion of the scale and scope of government. As we show in the next section, this general concern regarding the ever increasing size and scope of government due to government failure naturally evolved into a presumption of anarchy in the Virginia School (see Powell and Stringham 2009).

Public Choice and Anarchy

Beginning in the 1970s, Public Choice economists began to analyze the capability of individuals to engage in peaceful social cooperation without government. Although seemingly new and radical, this was an inquiry in political economy dating back to the nineteenth century. Carl Menger proposed this question in his 1883 book, *Investigations into the Method of the Social Sciences*, where he asked the following: “*How can it be that institutions which serve the common welfare and are extremely significant for its development come into being without a common will directed toward establishing them?*” (emphasis original, 1883 [1985]: 146). Due to the civil unrest that emerged during the Vietnam War and the Civil Rights movement, James Buchanan, Gordon Tullock, and Winston Bush undertook a

radical re-examination of alternative institutional arrangements for governing society at Virginia Polytechnic Institute and State University, resulting in the publication of *Explorations in the Theory of Anarchy* (1972) and *Further Explorations in the Theory of Anarchy* (1974). As Winston Bush stated, “[i]t is not surprising that ‘anarchy’ and ‘anarchism’ have reemerged as topics for discussion in the 1960s and the 1970s, as tentacles of government progressively invade private lives and as the alleged objectives of such invasions receded yet further from attainment” (1972 [2005]: 10). However, the early investigations into the prospects of anarchy were generally more pessimistic that they would later become among scholars of the Virginia School. Given the historical context in which they were writing, Buchanan, Bush, and the other contributors regarded anarchism with skepticism. “The anarchists of the 1960s,” according to Buchanan, “were enemies of order, rather than proponents of any alternative organizational structure” (2005: 192). Anarchy, as it was understood in *Explorations in the Theory of Anarchy* and *Further Explorations in the Theory of Anarchy*, referred to a state in society characterized by the absence of law, leading to banditry, violence, and social disorder. The common assumption held by these scholars was an identification of government with governance itself.

Since the 1970s, however, Public Choice scholars have extended the original work on government failure to show in fact that anarchy operates more effectively than previously believed. Applying the same logic developed by the earlier generation of Public Choice economists to develop the theory of government failure, later these scholars simply pushed the theory to its logical conclusion, arguing that governments would not always improve upon conditions of anarchy. The work of Public Choice economist Bruce Benson illustrates this well. For example, Benson has shown historically how international commercial law, known as the Law Merchant, emerged spontaneously in Medieval Europe to facilitate international trade, and operated without government enforcement. Disputes between merchants were settled in private merchant courts, and

these court decisions were accepted by winners and losers because they were backed by the discipline of repeated dealings as well as the threat of ostracism by the merchant community at large (Benson 1989). In *The Enterprise of Law* (1990), Benson also illustrates how the centralization of law enforcement by government in England later crowded out private mechanisms of law enforcement for the purpose of using the legal system to collect revenue. The failure of anarchy to provide a private market for law enforcement can be attributed then to government failure.

Another important publication in the transition from a pessimistic presumption of anarchy to a more optimistic presumption of anarchy was *Anarchy, State and Public Choice* (2005), which was ostensibly written in response to the contributions written in *Explorations in the Theory of Anarchy* and *Further Explorations in the Theory of Anarchy*. From *Anarchy, State and Public Choice*, a burgeoning literature has further developed the economic analysis of anarchy. Whereas the earlier generation of Virginia School economists saw the gains from trade and innovation being limited by the extent to which governments secured property rights and enforced contracts, the empirical challenge of this newer generation was to show that the existence of such potential gains from trade and innovation presented an entrepreneurial profit opportunity for the endogenous formation of norms and rules.

The economic analysis of anarchy as it evolved in the 1990s and 2000s, like the economic analysis of government failure in the 1950s and 1960s, cannot be understood outside the historical context in which such scholarship emerged. Specific events, such as the collapse of communism in Eastern and Central Europe, ethnic and religious fractionalization in the Balkans and the Middle East, and the exportation of liberal democracy to failed and weak states in the developing world, have demonstrated that governance requires the endogenous formation of rules, rather than their imposition by governments exogenously (Coyne 2008). Moreover, the economic analysis of anarchy, just like the theory of government failure that preceded it, emerged to counter the policy

implications of utilizing perfectly competitive equilibrium as a normative benchmark of analyzing markets in the developing world. The standard neoclassical model populated by fully informed and homogenous agents, in which property rights are well-defined and well-enforced, is unreliable to understanding the situation of failed and weak states in the developing world for two reasons. First, governments in the developing world provide poor enforcement of property rights, or are outright predatory. Second, precisely because the situation in failed and weak states is one in which individuals are heterogeneous, have imperfect information, and exhibit high discount rates, collective action problems that may exist under anarchy may prove to be even worse under a dysfunctional government. For example, the Virginia School scholars have looked at Somalia as a case study, analyzing development economic indicators before and after the collapse of the Barre regime in 1991. While it may be the case that Somalia under anarchy remains one of the poorest parts of the world, it does not automatically follow that the re-establishment of government would be an ideal solution for the provision of governance. In Somalia, data on standard indicators of economic development suggests that statelessness has improved Somali development substantially in terms of lower rates of infant mortality, higher life expectancy, and lower percentage of individuals living on less than one dollar per day (See Leeson 2007: 697). Scholars of anarchy in the Virginia School, like their predecessors analyzing government failure, have attempted to overturn the existence of a “Nirvana Fallacy” (Demsetz 1969) by comparing imperfect and real institutional alternatives in history between the existence of anarchy and the state. Such analysis can be traced back not only to the roots of Public Choice broadly defined as the economic analysis of nonmarket decision-making, but also more specifically as the study of the economic role of the state without romance. In both respects, Public Choice, particularly the Virginia School, has progressively brought the central inquiry of political economy back to the foreground of economics in the twentieth and twenty-first centuries.

Conclusion

Given the breadth and development of Public Choice since the 1950s, the author has focused on the intellectual origins and evolution of the Virginia School of Public Choice. The author has highlighted the uniqueness of the Virginia School in developing a theory of government failure to counter the presumption of market failure in economics. Moreover, in developing this presumption of government failure, the author has highlighted that the economic analysis of anarchy has followed logically from the initial skepticism among Public Choice economists to use public policy measures to address market failures. Although seemingly separate enterprises, they are both rooted in comparative institutional analysis and an understanding of the price mechanism as offering the possible approach to understanding the emergence of peaceful social cooperation without government command.

Cross-References

- [Anarchy](#)
- [Constitutional Political Economy](#)
- [Gordon Tullock: A Maverick Scholar of Law and Economics](#)
- [Government Failure](#)
- [Market Failure: Analysis](#)
- [Market Failure: History](#)
- [Political Economy](#)
- [Public Interest](#)

References

- Benson BL (1989) The spontaneous evolution of Commercial law. *South Econ J* 55(3):644–661
- Benson BL (1990) *The Enterprise of law: justice without the state*. Pacific Research Institute for Public Policy, San Francisco
- Boettke PJ (2012) *Living economics: yesterday, today, and tomorrow*. The Independent Institute, Oakland
- Boettke PJ, Lesson PT (2015) Introduction. In: Boettke PJ, Leeson PT (eds) *The international library of critical writings in economics 304: the economic role of the state*. Edward Elgar, Northampton

- Boettke PJ, Marciano A (2015) The past, present, and future of Virginia political economy. *Public Choice* 163(1–2):53–65
- Buchanan JM (1949) The pure theory of government finance: a suggested approach. *J Polit Econ* 57(6):496–505
- Buchanan JM 1958 (1999) The collected works of James M. Buchanan Vol. 2, Public principles of public debt: a defense and restatement. Liberty Fund, Indianapolis
- Buchanan JM (1964) What should economists do? *South Econ J* 30(3):213–222
- Buchanan JM 1969 (1999). The collected works of James M. Buchanan Vol. 6, Cost and choice: an inquiry in economic theory. Liberty Fund, Indianapolis
- Buchanan JM 1979 (1999) Politics without romance: a sketch of positive public choice theory and its normative implications. In: The collected works of James M. Buchanan Vol. 1, The logical foundations of constitutional liberty. Liberty Fund, Indianapolis
- Buchanan JM 1983 (2000) The achievement and the limits of public choice in diagnosing government failure and in offering bases for constructive reform. In: The collected works of James M. Buchanan volume 13: politics as public choice. Liberty Fund, Indianapolis
- Buchanan J 1996 (2001) Adam Smith as inspiration. In: The collected works of James M. Buchanan volume 19: ideas, persons, and events. Liberty Fund, Indianapolis
- Buchanan JM (2005) Reflections after three decades. In: Stringham E (ed) *Anarchy, state and public choice*. Edward Elgar, Cheltenham
- Buchanan JM (2015) Notes on Hayek – Miami, 15 February 1979. *Rev Austrian Econ* 28(3):257–260
- Buchanan JM, Thirby GF (1981) *L.S.E. Essays on cost*. New York University Press, New York
- Bush W 1972 (2005) Individual welfare in anarchy. In: Stringham E (ed) *Anarchy, state and public choice*. Edward Elgar, Cheltenham
- Coase RH (1959) The Federal Communications Commission. *J Law Econ* 2:1–40
- Coyne CJ (2008) *After war: the political economy of exporting democracy*. Stanford University Press, Stanford
- Demsetz H (1969) Information and efficiency: another viewpoint. *J Law Econ* 12(1):1–22
- Leeson PT (2007) Better off stateless: Somalia before and after government collapse. *J Comp Econ* 35(4):689–710
- Menger C 1883 (1985) *Investigations into the method of the social sciences*. New York University Press, New York
- Mitchell WC (1988) Virginia, Rochester, and Bloomington: twenty-five years of public choice and political science. *Public Choice* 56(2):101–119
- Mueller DC (1976) Public choice: a survey. *J Econ Lit* 14(2):395–433
- Powell B, Stringham EP (2009) Public choice and the economic analysis of anarchy: a survey. *Public Choice* 140(3–4):503–538
- Samuelson PA (1948) *Economics*. McGraw-Hill, New York

- Simmons RT (2011) *Beyond politics: the roots of government failure*. The Independent Institute, Oakland
- Tullock G (ed) (1972) *Explorations in the theory of anarchy*. Center for the Study of Public Choice, Blacksburg
- Tullock G (ed) (1974) *Further explorations in the theory of anarchy*. University Publications, Blacksburg
- Wagner RE (2012) *Deficits, debt, and democracy: wrestling with tragedy on the fiscal commons*. Edward Elgar, Cheltenham
- Wagner RE (2017) *James M. Buchanan and liberal political economy: a rational reconstruction*. Lexington, Lanham

Public Enforcement

Iljoong Kim

Department of Economics, SungKyunKwan University (SKKU), Jongno-Gu, Seoul, South Korea

Abstract

This essay starts with discussions regarding what public enforcement is and why it is necessary. We explain economic rationales under which public enforcement becomes a superior sanctioning mode, in controlling many undesirable acts, to a wide variety of non-public sanctioning counterparts. Nonetheless, given that a large portion of the literature considers the high-cost aspect of public enforcement, the essay emphasizes the importance of lowering administrative costs and overcoming bureaucracy. From a similar perspective, we also examine the combination of public and non-public enforcement as well as the joint use of different modes of public enforcement.

Synonyms

[Law enforcement by government](#)

Definition

Public enforcement (PE) is a sanctioning mode involving a wide variety of government people such as police, prosecutors, and regulators.

Although society can rely on a range of non-public sanctioning modes to control undesirable acts, certain situations occur where economic rationales exist for PE. The social harm from the act, the probability of detection, and the severity of the sanction are critical for its optimality. It is also important to lower administrative costs and to overcome bureaucracy for successful PE.

Introduction

Public enforcement is a sanctioning mode sometimes with the use of physical force, involving a wide variety of people in the government sector such as police, prosecutors, and various regulators. In fact, the phenomenon of public enforcement playing such a predominant role is very recent in human history terms.

Believing that an efficient sanctioning mode varies across different situations, a number of scholars have examined the issue of enforcement for some time under themes such as “system of social control,” “structure of enforcement,” “modalities of regulation,” or “methods of public control” (Ellickson 1973; Shavell 1993; Posner 2011). In particular, since Becker (1968), scholars have paid special attention to public enforcement and have produced numerous articles.

Why Public Enforcement and How Is It Undertaken?

Society relies on a range of nonpublic sanctioning modes to control many undesirable acts. A representative example is “self-help” such as reputation, self-protection, and purchase of insurance (Ehrlich and Becker 1972). Another example is civil litigations associated with torts, contracts, and other private-law doctrines, sometimes called “judicial regulations” (Posner 2011). Society has also developed remedial methods to support these nonpublic sanctioning modes, such as the property rule, the liability rule, or inalienability (Calabresi and Melamed 1972). However, certain situations occur where these private sanctioning

modes do not work or become very inefficient in coping with undesirable actors.

Since Becker and Stigler (1974), Shavell (1984a), and others, it has been well established that at least four economic rationales exist for public enforcement. Public enforcement becomes a superior mode as government is better equipped in gathering information about the production of harm. Also, upon being caught, injurers may have insufficient assets to compensate the victim (i.e., the judgment-proof problem). It is also superior when the potential injurer seldom faces the threat of civil suits and can thus escape liability. Finally, public enforcement should not incur too high administrative costs to secure its superiority.

Given these situations, based on Polinsky and Shavell (2000) and others, let us consider briefly how public enforcers could set the severity of the sanction (s) at an optimal level. A risk-neutral injurer (A) obtains a benefit (b) from an act which can incur social harm (h), with a probability (q). If h occurs, A is caught with a probability (p). p is less than one, reflecting the reality of imperfect enforcement. h differs across acts and is assumed to be fixed in the short run.

The optimality requires that the expected sanction to A , ps , should equal the expected social harm, qh . Since A then performs the act only when b exceeds ps , the optimality condition warrants the social efficiency of the act in question (i.e., the condition whereby b should exceed qh). In other words, A 's private incentive is consistent with that of the social planner. Further, since p is less than one, the optimal level of s is qh divided by p . Thus, the inverse of p plays the role of a multiplier, so that A 's private decision can be internalized appropriately. Note that the optimal level of s should be reduced if A is risk averse. Otherwise, A 's act will be over-deterred because the expected net benefit (i.e., b contracted by qh) is discounted by risk aversion. Finally, noteworthy is that this model is independent of A 's benefit, b . In fact, a model heavily focusing on b inevitably forces public enforcers to focus on the information associated with b that can inherently be more easily fabricated by A , consistently making the arrow of public enforcers land wide on the target of optimal sanctioning.

Additional comments on this simple model might shed beneficial insights, even for more complicated models that were later developed. First, public enforcers, of the three variables in the model, need to have precise information particularly regarding h across different acts, in order for this system to work (p and q are assumed to be fixed at least in the short run, and information on b is unnecessary). Therefore, the aforementioned superiority of government in gathering information about the harm is a core prerequisite for public enforcement.

Second, although the (low) level of p has often been treated as exogenous, public enforcers must maintain a certain level of p ; otherwise, the system will suffer from the ex post equity problem between the detected injurers and those who escaped successfully. Further, a prohibitively high s with a very low p (close to zero) will too easily induce the judgment-proof problem, leading to nullification of public enforcement. More importantly, such a system will hamper the essential (constitutional) principle in enforcement of “marginal deterrence,” which postulates that the severity of sanction should increase proportionally to the harm level. This notion, in fact, was emphasized even in earliest writings such as Bentham’s *Principles* in 1789. For example, if the death penalty or a one million dollar fine is imposed to a driver who hits a pedestrian and causes a slight wound, the driver would have an incentive to run away or even to kill the wounded pedestrian to escape the penalty. Therefore, public enforcers are required to invest in raising p to an acceptable level, which would consume real resources.

Lowering High Costs of Public Enforcement and Overcoming Bureaucracy

As theoretical inquiries have shown, public enforcement certainly offers an advantage under many circumstances. Nonetheless, as with many other government operations, it incurs generally high cost. If the assumptions about the government as the benevolent and omniscient planner are

released, the cost further increases. A large portion of the literature considers this cost aspect of public enforcement.

Firstly, scholars realized that, while public enforcement is needed for a certain number of undesirable acts, for various reasons, such acts cannot be efficiently controlled by it alone. For example, the marginal information cost increases with respect to the degree of enforcement sophistication. In fact, such recognitions motivated scholars to examine the allocation of public enforcement resources, even from the early stage of research. It is perhaps in this context that research on the combination of public and non-public enforcement captured their attention. In retrospect, these early studies were meaningful contributions, particularly in the sense that the researchers were developing more hands-on normative theory built on positive observations.

Consider the case of using traffic signals to maintain order on public roads. Enforcement by police in terms of whether drivers violate them can be done reasonably easily. However, police cannot assign a tailor-made speed limit that reflects each driver’s value of time, level of urgency, driving skill, etc. The limit is uniformly enforced (e.g., 100 km/h on highways and 30 km/h in downtown areas). If a driver mistakenly hits another car while driving over the limit and causes harm, liability is imposed on top of issuing a ticket (i.e., “per se negligent”). However, the opposite rule (i.e., “compliance defense”) does not hold. That is, even if the injurer is driving under the limit, liability is not automatically exempted.

This joint use of public enforcement and liability is used in most jurisdictions. Shavell (1984b) attempted the first theorization to draw its efficiency implication. He highlighted the imperfect nature of public enforcement (e.g., enforcing the speed limit) due to imperfect information particularly about different magnitudes of harm. Also, the insufficient precaution when using liability alone was emphasized. Thus, a popular proposition followed that it is socially cheaper to employ regulation and liability jointly, with a lower regulatory standard than if liability were not used. Later, the role of liability to

ameliorate the high costs of public enforcement, primarily in terms of reducing information costs, was further confirmed. Also, the role of public enforcement to support liability was sometimes underlined instead. Overall, these studies were endeavors in search of efficient public enforcement, given its high costs (Kolstad et al. 1990; De Geest and Dari-Mattiacci 2007; Bhole and Wagner 2008).

Scholars subsequently applied these theoretical studies to broad arenas, ranging from product safety to transportation, hazards, environment, health, etc. La Porta et al. (2006) provided an exemplary inquiry. The authors demonstrated that the proper use of civil liability standards together with public enforcement is necessary for the successful operation of securities markets. Meanwhile, extensions were made to explore the joint use of different modes of public enforcement, such as that of administrative penalties and criminal punishment (Garoupa and Gomez-Pomar 2004; Bowles et al. 2008). Given that such joint use is ubiquitous across countries, this approach is differentiated from the earlier dichotomous-choice models. Furthermore, research on the combined criminal sanctions of fines and imprisonment was launched already in the 1980s (Polinsky and Shavell 1984). A major implication is that the use of imprisonment, the highest-cost sanction, should be confined to cases where the insufficient-asset problem or the need to incapacitate offenders prevails.

Finally, a brief but significant caveat deserves mention. Although overall features of actual public enforcement are roughly consistent with the theories, substantial discrepancy routinely occurs, i.e., “bad equilibrium,” primarily due to the public enforcer’s incentives which differ from the public-interest mindsets that ordinary citizens would expect them to have (Becker and Stigler 1974). Public enforcement can be characterized as being dominated by “entrepreneurial competition,” wherein bureaucrats pursue their subjective goals such as wealth, promotion, and discretion (Breton and Wintrobe 1982). Numerous researchers have already examined public enforcement from this perspective. Nonetheless, in summary, it should be emphasized that the details of public

enforcement must be steadily scrutinized through these critical lenses in order to facilitate a change in such bad equilibria. Also, in attempting to further lower high costs, incentive-compatible rules must be implemented that require much more use of “pricing” in the allocation of enforcement resources.

Cross-References

- ▶ [Becker, Gary S.](#)
- ▶ [Criminal Sanctions and Deterrence](#)
- ▶ [Government](#)

References

- Becker G (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Becker G, Stigler G (1974) Law enforcement, malfeasance, and compensation of enforcers. *J Leg Stud* 3:1–18
- Bhole B, Wagner J (2008) The joint use of regulation and strict liability with multidimensional care and uncertain conviction. *Int Rev Law Econ* 28:123–132
- Bowles R, Faure M, Garoupa N (2008) The scope of criminal law and criminal sanctions: an economic view and policy implications. *J Law Soc* 35: 389–416
- Breton A, Wintrobe R (1982) The logic of bureaucratic control. Cambridge University Press, New York
- Calabresi G, Melamed D (1972) Property rules, liability rules, and inalienability: one view of the Cathedral. *Harv Law Rev* 85:1089–1128
- De Geest G, Dari-Mattiacci G (2007) Soft regulators, tough judges. *Supreme Court Econ Rev* 15:119–140
- Ehrlich I, Becker G (1972) Market insurance, self-insurance and self-protection. *J Polit Econ* 80:623–648
- Ellickson R (1973) Alternatives to zoning: covenants, nuisance rules, and fines as land use controls. *Univ Chic Law Rev* 40:681–714
- Garoupa N, Gomez-Pomar F (2004) Punish once or punish twice: a theory of the use of criminal sanctions in addition to regulatory penalties. *Am Law Econ Rev* 6:410–433
- Kolstad C, Ulen T, Johnson G (1990) *Ex-post* liability for harm vs. *ex-ante* safety regulation: substitutes or complements? *Am Econ Rev* 80:888–901
- La Porta R, Lopez-de-Silanes F, Shleifer A (2006) What works in securities laws? *J Financ* 61:1–32
- Polinsky M, Shavell S (1984) The optimal use of fines and imprisonment. *J Public Econ* 24:89–99
- Polinsky M, Shavell S (2000) The economic theory of public enforcement of law. *J Econ Lit* 38:45–76

- Posner R (2011) *Economic analysis of law*, 8th edn. Aspen Publishers, New York
- Shavell S (1984a) Liability for harm versus regulation of safety. *J Leg Stud* 13:357–374
- Shavell S (1984b) A model of the optimal use of liability and safety regulation. *RAND J Econ* 15:271–280
- Shavell S (1993) The optimal structure of law and enforcement. *J Law Econ* 36:255–287

Public Goods

Tim Friehe¹ and Florian Baumann²

¹Public Economics Group, Philipps-University Marburg, Marburg, Germany

²Center for Advanced Studies in Law and Economics (CASTLE), University of Bonn, Bonn, Germany

Definition

A public good is a good that simultaneously features both nonexcludability and nonrivalry in consumption.

Delimitation and Examples

A pure public good is a good that simultaneously features both nonexcludability and nonrivalry in consumption. Nonexcludability implies that excluding individuals from making use of the good is prohibitively costly, such that all individuals can benefit from the provision of the good. Nonrivalry means that the consumption of the good by one individual does not reduce the consumption possibilities of other individuals. Classic examples include national defense or a lighthouse. In contrast, a private good is characterized by both excludability and rivalry in consumption. For example, if an apple is consumed by one individual no other individual can make use of the apple. In addition, excluding other individuals from the benefits of the apple is associated with only minor costs. There are also goods that show rivalry in consumption but nonexcludability. Such goods are denoted commons

and are exemplified by fish in a given area. While it is difficult and costly to exclude individuals from fishing, every fish caught by one person can no longer be caught by any other individual and may endanger the sustainability of the stock of fish. Finally, pure public goods have to be distinguished from club goods. Club goods are characterized by nonrivalry in consumption and the possibility of exclusion. Broadcasting can be regarded as a club good, because watching the program does not influence the option for others to do the same while exclusion is possible through the use of scrambling.

Theory

From an economic point of view, the property that the consumption of the public good by one individual does not infringe on the consumption benefits of other individuals leads to a first question: how is the socially optimal supply of a pure public good determined? Moreover, given nonrivalry and nonexcludability, it is also interesting whether markets or voluntary contributions by individuals can implement the optimal supply of the public good. The basic problem may best be illustrated by a simple example resembling the prisoners' dilemma.

Suppose there are two individuals, A and B, who both hold an endowment that generates private utility equal to one. Each individual decides once whether or not to use the endowment to provide one unit of a public good. Due to the nonrivalry of the public good, both individuals gain from the provision of one unit of the public good by either of the individuals. Suppose that the utility derived per unit of the public good equals 0.75 for each individual. Table 1 summarizes the possible choices and resulting utility levels, where the first (second) entry refers to individual A's (B's) utility level.

Public Goods, Table 1 Actions and payoffs in a simple public-good game

Individual A/Individual B	Don't supply	Supply
Don't supply	1;1	1.75;0.75
Supply	0.75;1.75	1.5;1.5

The outcome that maximizes the sum of utilities is the one in which both individuals supply one unit of the public good (in which case the joint payoff equals three). The costs for each unit supplied are equal to one while the joint payoff is given by 1.5 ($0.75 + 0.75$). However, if the individuals A and B decide independently, each of them fares better by not supplying one unit of the public good given any decision by the other individual. The supply of one unit of the public good only yields a private benefit of 0.75 for each individual and is not sufficient to balance the utility cost equal to one. As a result, the equilibrium in strictly dominant strategies is that neither A nor B supplies a unit of the public good; consequently both end up with a utility level of one. This outcome is Pareto inferior to the optimal solution.

The example illustrates two ideas regarding public goods: (1) To decide on the socially optimal level of supply of the public good, the costs of supplying an additional unit should be compared to the sum of gains achieved by all individuals through the additional supply (due to the property of nonrivalry). (2) Without cooperative decision-making, the voluntary supply of the public good may not yield the optimal outcome.

More generally, the seminal condition that characterizes the Pareto-efficient supply of a public good was first established by Samuelson (1954, 1955). For any two private goods, a Pareto-efficient allocation is achieved when the marginal rates of substitution between the two goods (i.e., the rates at which individuals are willing to exchange the two goods without a change in utility) are equalized for all individuals and, in addition, are equal to the marginal rate of transformation (i.e., the rate at which the two goods can be exchanged according to the production possibility set). In contrast, the optimal allocation regarding a public good and a private one requires that the sum of the marginal rates of substitution over all individuals is equal to the marginal rate of transformation. As in the example above, increasing the supply of the public good is socially desirable as long as the marginal costs (i.e., the marginal rate of transformation) are weakly less than the sum of the gains enjoyed by all

individuals (i.e., the sum of the marginal rates of substitution).

An important contribution examining the theory of the noncooperative provision of public goods is provided by Bergstrom et al. (1986). They confirm that public goods are underprovided and establish that a redistribution of wealth among individuals will change the amount of the public good supplied only if it changes the set of individuals actually contributing to the public good. Furthermore, additional supply of the public good by the state (and financed by taxes) will at least partly (if not fully) be offset by reductions in private contributions.

A cooperative provision of public goods may be attained under some circumstances, especially when the group considered is small. For example, a Lindahl equilibrium constitutes an allocation resulting from bargaining which results in a Pareto-efficient supply of the public good with individual-specific contributions based on each individual's valuation of the public good (Lindahl 1919). However, it must be acknowledged that individuals might have an incentive to misrepresent their preferences. The later literature related to the topic of cooperative public good provision has been greatly influenced by the work on collective action by Olsen (1965).

An alternative to the private provision of public goods is the provision by the state that can collect contributions by force in the form of taxes. In this case, if only distortionary taxes can be used to raise revenue, the additional costs from these distortions should be taken into account when deciding about the optimal level of the public good (see, e.g., Atkinson and Stern 1974). An even more fundamental problem arises in the case with asymmetric information about individuals' preferences. If individual payments are to be based on the stated preferences for the public good, individuals might understate their willingness to pay. Otherwise they might overstate it if they know that the burden of financing the additional supply will be shared by the society as a whole. Clarke (1971) and Groves (1973) present a mechanism which leads to a truthful statement about preferences. The amount of the public good may also be determined by majority voting

instead of by a benevolent government, a scenario for which Bowen (1943) offers an early analysis (Bergstrom and Goodman 1973 provide an empirical investigation).

The theory of public good provision based on standard preferences has been complemented by predictions derived from alternative specifications for individual preferences. For example, in order to better understand the sizable contributions to charities, nonstandard motives such as impure altruism may be considered. In the case of impure altruism, individuals obtain utility by increasing their contribution not only from a higher total level of the public good but also from having a higher individual contribution since these create a so-called warm glow (Andreoni 1989, 1990). The predictions that follow from frameworks enriched in such a way are often more intuitive and more in line with empirical observations, such as the result that government contributions to charity will only incompletely crowd out private donations.

Experiments

For decades, the so-called public-good game has been a workhorse for experimental economists interested in social dilemmas and their potential resolution through institutions (e.g., Ledyard 1995). Classically, individuals in groups of n subjects simultaneously determine the split of their symmetric endowments between a public account and a private one, where the return from the public account per subject is determined by the product of the marginal per capita return ($MPCR$) and the sum of contributions and $0 < MPCR < 1 < MPCR * n$ applies. This is the so-called linear public good game and has a unique Nash equilibrium in which no subject (with standard preferences) contributes. The experimental finding that subjects contribute on average 40–60% of their endowment to the public good in the one-shot version or in the first round of the repeated variant of the game is in striking contrast to the prediction based on standard preferences. However, the initially high contribution rates, which decline over time, are consistent with

the idea that there are many conditionally cooperative subjects who are willing to contribute more when they expect others to contribute, although they do not match an increase in contributions by the others in full (e.g., Croson 2007; Fischbacher et al. 2001). Turning to institutions that remedy free-riding incentives, allowing participants to punish peers based on their contributions seems effective (although not necessarily efficient when punishment costs are taken into account). Intuitively, costly peer punishment is on average predominantly chosen by subjects with above-average contributions and addressed at subjects with below-average contributions. Such a mechanism seems to be associated with not only higher overall contributions but in the case of low costs of punishment and/or a long time horizon with real efficiency gains (e.g., Chaudhuri 2011). Interestingly, there is experimental evidence that the possibility of peer punishment attracts subjects when they can migrate between a regime with the possibility to sanction and one without and that sanctioning institutions emerge endogenously (e.g., Gürer et al. 2006; Kosfeld et al. 2009). Other means to improve contributions to the public good include communication and ostracism (e.g., Chaudhuri 2011; Maier-Rigaud et al. 2010).

Applications

The theory of public goods is relevant to a number of applications in the field of law and economics. First, the private enforcement of norms, for example, by assigning punishment points to peers in experimental settings (as just described) or by ostracizing and shaming offenders in the field may become subject to free-rider incentives, as the deterrence benefit is diffused (e.g., Posner 1997). Relatedly, consider the case of unobservable private precautions against crime. When a household increases investments, this lowers the expected return for the potential offenders, decreasing the involvement of thieves. The benefit accrues to all households equally, whereas the full costs are borne by the household in question. This implies that decentrally determined private precautions

against crime fall short of what is optimal for the group of households (e.g., Shavell 1991). In another realm, it is often argued that the knowledge created by innovative activity is nonrival and may present difficulties regarding excludability. Patents ensure excludability for some time, thereby incentivizing the creation of knowledge while limiting the extent of the use of the knowledge. Accordingly, there is a discussion of the relative merits of different instruments in this context (e.g., Shavell and van Ypersele 2001). In addition, the divergence of private marginal benefits and social marginal benefits that commonly arises with public goods may also induce socially inefficient private incentives to proceed to trial. Court decisions may be socially valuable by creating public information about plaintiffs or precedents, aspects which usually do not feature prominently in the private trade-off (e.g., Hua and Spier 2005; Shavell 1999).

Cross-References

- [Altruism](#)
- [Commons, Anticommons, and Semicommons](#)
- [Efficiency](#)
- [Experimental Law and Economics](#)
- [Externalities](#)
- [Market Failure: Analysis](#)
- [Market Failure: History](#)

References

- Andreoni J (1989) Giving with impure altruism: applications to charity and Ricardian equivalence. *J Polit Econ* 97:1447–1458
- Andreoni J (1990) Impure altruism and donations to public goods: a theory of warm-glow giving. *Econ J* 100:464–477
- Atkinson A, Stern N (1974) Pigou, taxation and public goods. *Rev Econ Stud* 41:119–128
- Bergstrom T, Goodman R (1973) Private demands for public goods. *Am Econ Rev* 63:280–293
- Bergstrom T, Blume L, Varian H (1986) On the private provision of public goods. *J Public Econ* 29:25–49
- Bowen H (1943) The interpretation of voting in the allocation of economic resources. *Q J Econ* 58:27–48
- Chaudhuri A (2011) Sustaining cooperation in laboratory public good experiments: a selective survey of the literature. *Exp Econ* 14:47–83
- Clarke E (1971) Multipart pricing of public goods. *Public Choice* 11:17–33
- Croson R (2007) Theories of commitment, altruism and reciprocity: evidence from linear public good games. *Econ Inq* 45:199–216
- Fischbacher U, Gächter S, Fehr E (2001) Are people conditionally cooperative? Evidence from a public good experiment. *Econ Lett* 71:397–404
- Groves T (1973) Incentives in teams. *Econometrica* 41:617–631
- Gürerk Ö, Irlenbusch B, Rockenbach B (2006) The competitive advantage of sanctioning institutions. *Science* 312:108–111
- Hua X, Spier KE (2005) Information and externalities in sequential litigation. *J Inst Theor Econ* 161:215–232
- Kosfeld M, Okada A, Riedl A (2009) Institution formation in public good games. *Am Econ Rev* 99:1335–1355
- Ledyard O (1995) Public goods: a survey of experimental research. In: Kagel J, Roth A (eds) *Handbook of experimental economics*. Princeton, Princeton University Press, pp 111–194
- Lindahl E (1919) *Die Gerechtigkeit der Besteuerung*. Gleerup, Lund
- Maier-Rigaud FP, Martinsson P, Staffiero G (2010) Ostracism and the provision of a public good: experimental evidence. *J Econ Behav Organ* 73:387–395
- Olsen M (1965) *The logic of collective action*. Harvard University Press, Cambridge, MA
- Posner R (1997) Social norms and the law: an economic approach. *Am Econ Rev Pap Proc* 87:365–369
- Samuelson PA (1954) The pure theory of public expenditure. *Rev Econ Stat* 36:387–389
- Samuelson PA (1955) Diagrammatic exposition of a theory of public expenditure. *Rev Econ Stat* 37:350–356
- Shavell S (1991) Individual precautions to prevent theft: private versus socially optimal behavior. *Int Rev Law Econ* 11:123–132
- Shavell S (1999) The level of litigation: private versus social optimality of suit and settlement. *Int Rev Law Econ* 19:99–115
- Shavell S, van Ypersele T (2001) Rewards versus intellectual property rights. *J Law Econ* 44:525–547

Further Reading

- Batina RG, Ihori T (2005) *Public goods: theories and evidence*. Springer, Berlin
- Cornes R, Sandler T (1996) *The theory of externalities, public goods and clubs*. Cambridge University Press, Cambridge
- Oakland WH (1989) *Theory of public goods*. In: Auerbach AJ, Feldstein M (eds) *Handbook of public economics*, vol 2. North Holland, New York, pp 485–535
- Sandmo A (2008) *Public goods*. In: Durlauf SM, Blume LE (eds) *The new Palgrave dictionary of economics*, 2nd edn. Palgrave Macmillan, Basingstoke

Public Interest

Michael Hantke-Domas

Centre for Water Law, Policy and Science,
University of Dundee, Scotland, United Kingdom
Third Environment Court, Valdivia, Chile

Abstract

The public interest is a concept that can be traced back to the late XVIII century. Ever since the concept has been used to refer to a goal to be obtained by actions of governments and public officials alike. Such view has been strongly contested by important economic schools such as Public Choice and the Chicago School of Regulation. Based on neoclassical economics, they contend the altruism needed to deliver in the public interest, as public officials pursue their own interest and regulations are captured to benefit regulated industries but recognizes the existence of equilibria that might balance the interests at stake (like Pareto or Kaldor-Hicks criteria). In contrast, the law is more optimistic when using public interest either to proceduralize collective interest or for courts adjudication. Even more, public interest serves as justification for government intervention in economic affairs. In the last decades, the promotion of integrity and prevention of corruption have tackled self-interest of public officials, leaving space for the public interest to be pursued by governments and public officials.

Definition

Public interest is the motivation of public officials in exercising their functions to achieve the common good of the collective, by improving democracy and social and economic welfare.

Introduction

Abraham Lincoln in his Gettysburg address (1863) coined a famous phrase defining

democracy as a “government of the people, by the people, for the people.” The idea that governments – or States, in continental European tradition – must serve the people is an expectation common to many people in modern history.

Modern States are inspired in the principles of the French Revolution that changed the Old Regime to a new social organization based on the values of liberty, equality, and fraternity. The State as a moral entity is – or should be – the guarantor of such values. The State has grown to become one of the most important institutions in many aspects of societies. Indeed, meanwhile in the nineteenth century, government expenditure amounted to 10% of GDP, and by 1996 it amounted to 45% in developed OECD countries (see Middleton 1996; Tanzi and Schuknecht 2000).

The “common good” – or the general welfare – is an important ideal for governments, as they must act in pursuance of public interest. However, what is the substance or content of such ideal?

John Rawls, in his book *A Theory of Justice* (1971), does not treat public interest as a common good but rather delineates a corresponding ethical ideal that is “justice as fairness” (Weisbrod and Benjamin 1978).

“For us the primary subject of justice is the basic structure of society, or more exactly, the way in which the major social institutions [political constitution and the principal economic and social arrangements] distribute fundamental rights and duties and determine the division of advantages from social cooperation...the major institutions [for example, the legal protection of freedom of thought and liberty of conscience, competitive markets, private property in the means of production, and the monogamous family] define men’s rights and duties and influence their life prospects, what they can expect to be and how well they can hope to do.” (p. 7)

Public interest or justice as fairness hinges upon an assessment made in the “initial situation” where an interpretation is made of the moment and the problem of choice it poses, and an agreement is entered into on a set of principles (Rawls 1971). These principles illuminate what public interest is as they serve as guidance for the initial

position, in the sense that (a) each person has rights equal “to the most extensive basic liberty compatible with a similar liberty of others” (p. 60) and “social and economic inequalities are to be arranged so that they both (a) reasonably expect to be to everyone’s advantage, and (b) attached to positions and offices open to all” (p. 60). Restating Rawls’ proposal, public interest or justice as fairness might refer to equal rights and fair access.

Economists have different takes, which range from equating public interest to general welfare to those that see no benefit to anyone except public officials. Lawyers feel more comfortable with the concept, and the concept of public interest is widely used with less skepticism.

Not surprisingly the concept of public interest has attracted criticism, as some think that it is too broad to mean something specific (Lewis 2006). Notwithstanding, the concept remains important, probably because it is an analytic tool or a heuristic device present in normative public administration theory, economic regulation, administrative law, and judicial procedures. It is also important in governmental and professional standards of practice (Lewis 2006). Furthermore, it is of relevance for law, as interests are objects of protection by constitutions and laws, and adjudicated by the judiciary (Schmidt-Assmann 2003). Public interest is also important in economics as market failures – and the role of the State to correct them – find in public interest a rationale for economic regulation.

Despite the controversy, what remains the same is the existence of a space where citizens can get together to, for example, entrust their protection either from foreign aggression or for getting medical assistance.

As public interest can be interpreted from a variety of disciplines, I will devote the rest of the entry to reviewing the discussions from the standpoints of economics and law.

Public Interest and Economics

The concept of public interest holds the attention of economists in at least two different areas. First,

it is a criterion used in collective choice, an area of economics that cares about the relationships between the preferences of individuals and the choices made by the government (Brown and Jackson 1998). Second, it is used for State regulation of economic activities when markets fail.

In liberal democracies, citizens select their representatives (e.g., members of parliaments or presidents) to make decisions and choices on their behalf. For economists, the aggregation of all individual preferences produces multiple combinations that are efficient to adopt by governments. However, this aggregation does not specify how societies formulate and express collective value judgments (Brown and Jackson 1998).

This is the point where disagreements start to appear, as ethical issues arise regarding collective choice rules (is the rule ethically acceptable?). Economists have argued over the idea of the rational, disinterested, and benevolent public official, which altruistically takes office for the good of the whole community.

Public choice scholars blow off the foundations of the commonly accepted idea that government officials pursue the interests of the whole community. They sustain that human beings act on a rational, self-interested, and utility maximizing basis (Mueller 1989) meaning that government officials act pursuing their own interest rather than promoting the general welfare.

Anthony Downs (1957), in *An Economic Theory of Democracy*, sustained that human beings are selfish. This is simply the application of a tenet of neoclassical economic theory of consumers. As individuals maximize their own utility, Downs could not fathom why government officials should depart from such truism to be altruists devoting their work to enhance the general welfare. If everyone is selfish, then government officials should be the same. Downs also studied political parties and concluded that they were mere vehicles to foster private interests (Lewin 1991).

In the same line as Downs, Buchanan and Tullock in *The Calculus of Consent* (1962) sustained that politics is an exchange between parties that aim to maximize individual interest. Common good was discarded as naive, since

public interest could only mean the satisfaction of many individual desires, facilitated by political activity (Lewin 1991).

Later, Gordon Tullock elaborated on government officials and concluded that an expression of their self-interest was the expansion of public sector irrespective of costs. According to Tullock, bureaucrats wanted to enhance their position either by increasing their own salaries, esteem, or influence. Bureaucrats disregarded the intentions of legislature – similar to public interest – as they wanted to foster their own private interests (Lewin 1991).

Public interest, viewed from public choice theory, is simply the aggregation of individual interests. Paradoxically, it is impossible, under mild conditions, to aggregate individual interests as Kenneth Arrow (1951) demonstrated in his *general possibility theorem*. In short, it is impossible to create a social order of preferences based on individual interests that meet certain criteria: unrestricted domain, non-dictatorship, Pareto efficiency, and independence from irrelevant alternatives (Sola 2004).

Leif Lewin (1991) adds that such impossibility is reinforced by the prisoner's dilemma, which shows that self-centered individuals driven by their own individual preferences usually end up worse off not better off from a collective point of view.

Even if it were possible to aggregate individual preferences, economics only presents a narrow frame of reference, as it does not include in its analysis different social and political values inherent to our democracies, which usually are part of the legal basis for the existence of the State, represented in their constitutions (Feintuck 2010). One clear example is the environment, as it has legal recognition providing for regulation that stretches beyond economic analysis, for example, the precautionary principle that allows governments to adopt decisions despite the existence of incomplete scientific data to warrant them, in order to protect the environment (Feintuck 2010).

Economic regulation is connected to collective choice as both are motivated by public and/or private interests. Government decisions are

present in the regulatory process, from ascertaining the need for State intervention to its implementation. Thus, the collective choice discussion is part of economic regulation. The assessment of whether a market fails or not and the decisions to correct externalities, to complete information, to promote competition, to solve principal-agent problems, and to provide public goods are all governed by collective choice.

Economic regulation studies have shown that its implementation might not satisfy general welfare, as regulation in itself might fail (for reviews, see Hägg 1997; Hantke-Domas 2003). The Chicago Theory of Regulation, originated in the works of George Stigler (1971) and further developed by Sam Peltzman (1976) and Gary Becker (1983), holds that regulation is captured by industry and consequently is designed and operated for its benefit. Richard Posner (1974) added that empirical evidence showed the poor performance of the regulatory process usually benefited influence groups. In addition, regulation encourages rent-seeking behavior because interested parties are willing to invest in order to yield favorable decision-making by agencies or courts (Harnay and Marciano 2011).

Equally, economists have shown mechanisms allowing to correct market failures through the market without the need for governmental regulation, i.e., self-centered agents still maximize welfare even in case of market failures such as externalities or monopolies, for example, Coase's theorem regarding property rights and externalities or Demsetz's contestable markets showing that even if competition *within* the market is not possible, it still is possible to motivate competition *for* the market (Hägg 1997).

Institutional economics has shown that regulatory institutions (i.e., governments) can replace or assist private bargaining in the presence of externalities or risks born by groups facing high transaction costs or asymmetry of information endured by consumers prior to transactions. This insight might be associated with public interest, but it is not contradictory with public choice or the Chicago theory of regulation. Indeed, regulatory institutions might well be the result of pressure groups or the correction of a market failure (Hägg 1997).

Agency theory sustains that economic regulation fails as information is asymmetrical between regulators and regulated industry; hence, the latter can extract economic rents in the regulatory process. This phenomenon leads to agency capture. In other words, correction of market failures – aiming at enhancing general welfare – may not be possible as regulated industries drag out the regulatory process (Laffont and Tirole 1993). Again, this insight does not trump the public choice of the Chicago theory of regulation, but sheds light on the complexities of economic regulation, leaving space for public interests to be a driving force of regulation (Hägg 1997).

The discussion on collective choice shows that public decisions cannot satisfy every individual interest and the quest for public interest might be more complex than assuming that governments will promote it seamlessly. Notwithstanding, it is possible to find efficiency in cases of State intervention, despite the limitations annotated. Pareto criterion might be satisfied if at least one person is better off and none of the rest are worse off. Even in the former case, public interest decisions might be efficient if the well off can compensate the worse off, using the Kaldor-Hicks criterion (Coleman 1980).

Public Interest and Law

Public interest is ubiquitous within the realm of law. It proceduralizes collective interests, and courts adjudicate disputes to achieve justice. Public officials assess their actions against the interests of the whole community in implementing public policy by means of the law.

Traditionally, the law has used the concept of public interest as justification for government intervention in economic affairs. Black's Law Dictionary (2004) defines it as "The general welfare of the public that warrants recognition and protection" or "Something in which the public as a whole has a stake; esp., an interest that justifies governmental regulation."

In its regulatory justification, public interest originally appeared in the work of Lord Mathew

Hale, *The Portibus Maris* (1787), inspiring two important decisions, one in England (*Allnutt v. Inglis*) and one in the United States (*Munn v. Illinois*). In essence, Lord Hale sustained that private businesses become *juris publici* if they were licensed or chartered by the King to act as a monopoly and those services were available to the public (Hantke-Domas 2003). The importance of the concept lies in the fact that some economic activities were in the interest of the public, and correspondingly it was necessary to balance expectations with the economic activity to obtain mutual benefit (Craig 1991).

The law embraces the concept of public interest in many ways besides being a warrant for governmental regulation. The UK Enterprise Act (2002) introduced a "public interest test" in cases of takeovers and mergers, by which the Secretary of State may intervene in the latter if interests of the public are involved as, for example, with issues of national security, media quality, plurality and standards, and financial stability ("Takeovers," n.d.). The UK Public Interest Disclosure Act (1998) bears a test that grants protection to whistle-blowing employees if the matter brought to light is in the public interest. Another important UK law is the Freedom of Information Act 2000 that provides for access to government-held information only if the public interest in disclosing it outweighs the public interest in not disclosing it. In these three laws, public interest is liberally used either to justify by the government a ban on a merger or to withhold or publicize information.

Civil service is another area where public interest is present ("Public Interest in UK Courts" 2011). In the United Kingdom, Lord Nolan's 7 Principles of Public Life are part of the remit of the Committee on Standards in Public Life advising the British government. For example, "selflessness" principle affirms, "Holders of public office should act solely in terms of the public interest." In Spain, its constitution (Section 103.1) obliges its civil service ("*administración pública*") to work for "general interests."

British courts may deal with public interest in different opportunities as defined by the Public Interest in UK Courts Project (2014). First, the law refers expressly to the concept, as in the case

of the English Freedom of Information Act 2000. Second, it refers to legal proceedings brought by the Attorney General in the “public interest.” Third, the courts may invoke the concept to justify new developments in the law. Fourth, public interest immunity may be argued in court by the Crown to withhold documents from parties when such disclosure may be contrary to the public interest. Fifth, European Union law and the law of the European Convention on Human Rights recognize rights of European citizens but that these rights can be limited in the public interest in cases of national security, public health, and the prevention of crime.

Public officials’ decisions in England are subject to judicial review regarding lawful exercise of their statutory powers. Public interest may be argued by courts to extend this review to non-statutory decisions. In continental Europe administrative law holds public officials accountable for the same reason, but with a comparative more developed corpus of law known as contentious administrative.

In Spain concepts such as “just economic and social order” or “quality of life” are written in her constitution and are part of the common good; hence “. . . public interest happens when the aim that must be served by a political or administrative organisation on its entirety, can only be achieved by that entirety” (Parejo 2003). More precisely, the general interest is fused together with the aims of the State. In the Spanish Constitution general interest refers to the protection of legal interests belonging to the community, a safeguard duty that must be assumed by the State as it is in charge of managing the common interest (Parejo 2003).

From a German perspective, the idea of common good is equally present, and for Prof. Eberhard Schmidt-Assmann (2003) it is the guidance for Parliament, public officials, and judges, but to be defended as well by private entities, interest groups, and individuals. Public interest strives to ensure the general interest. Both concepts are not interchangeable, but since public interest is promoted by the community, it tends to become general interest. Private interests are not opposed to general ones and in their evolution

blend with the general interest. Public interest is not predefined and static, but it evolves throughout the administrative process. The administration – or government – is conceived as a structure representing multiple public interests. For Prof. Schmidt-Assmann general welfare is equivalent to common good, and public interest is the feasible fulfillment of common interests by means of the law. “The determination of what is general welfare is a matter depending above all on positive law, which normally offers for its achievement procedures and material criteria” (2003, p. 167). This proceduralization of public interest should be understood under the light of fundamental rights and the rules regarding the State function. Common good is the government-guiding ideal to act in favor of the community, and democracy is better served if common good is actively pursued.

The use of the concept of public interest as a heuristic tool is present in the United Kingdom and in Spain and Germany, but in Spain and Germany, it is a legal obligation to be followed by the State, whereas in the United Kingdom it guides a limited set of situations.

This simple overview of different uses of public interest in the law shows that institutions are always inspired to serve the community and public officials are bound to pursue the public interest. Economists, at least some of them, reject any possibility that a public official – or anyone – may act in any way but for his or her own interest; hence the public interest might be an excuse for advancement by protecting the interests of groups. Law and economics clash for being on the one hand idealist (law) whereas on the other hand pessimistic (economics), but the power of law guarantees that public officials will be restrained on their self-interest. The law carries penalties that cannot be defined as idealistic, such as jail terms or hefty fines.

Consistent with the Public Choice Theory and the Chicago Theory of Regulation that dispute the existence of a benevolent government always searching for maximum social welfare, in the last decades the promotion of integrity and the prevention of corruption in the public space have increased. Although studies and policies on

corruption do not focus on public interest, they tackle self-interest of public officials. The matter has taken a prominent position on the agenda of international organizations with the adoption of the United Nations Convention against Corruption (2003). Indeed, important efforts to oblige States to implement regulations to curb the phenomenon preceded the UN Convention, as the OECD Convention on Combating Bribery of Foreign Public Officials in International Business Transactions (1999) or the Organization of American States Inter-American Convention against Corruption (1996). National legislation has gathered pace (e.g., UK's Bribery Act 2010, US Foreign Corrupt Practices Act).

Under the light of integrity legislation, corruption emerges when a public official obtains a private gain or status from office. Indeed, corruption is defined as “the abuse of public office for private gain” (World Bank 1997). A more comprehensive definition is given by Robert Klitgaard (1988) “behaviour which deviates from the formal duties of a public role because of private regarding (personal, close family, private clique) pecuniary or status gains; or violates rules against the exercise of certain types of private regarding behaviour.”

Behind the international effort to curtail corruption is the belief – and evidence – that corruption distorts prices. Inefficient markets with covert and upward redistribution of wealth within a society increase costs and reduce investment and produce a decline in society's moral and ethics (Senior 2006). Corruption can be petty or grand, and even there are some authors that affirm that in some societies, corruption helps failed States to run properly.

Using Robert Klitgaard's (1988) corruption equation (corruption equals monopoly power plus discretion by officials minus accountability), if public officials are rational, then public interest – as opposed to personal interest – should be pursued by them if the expected gain by the corruption is less than the penalty of being caught and prosecuted. This conclusion leaves the possibility that public interest can be a motivation for public officials when corruption is not present.

Cross-References

- [Administrative Corruption](#)
- [Altruism](#)
- [Constitutional Political Economy](#)
- [Cost–Benefit Analysis](#)
- [Efficiency](#)
- [Governance](#)
- [Market Failure: Analysis](#)
- [Political Corruption](#)
- [Public Choice: The Virginia School](#)
- [Rent Seeking](#)
- [Rule of Law](#)

Acknowledgment The author would like to acknowledge the helpful research assistance of Ms Francisca Henriquez Prieto.

References

- Arrow K (1951) Social choice and individual values, 2nd edn. Wiley, New York
- Becker GS (1983) A theory of competition among pressure groups for political influence. *Q J Econ* 98(3):371. <https://doi.org/10.2307/1886017>
- Brown C, Jackson P (1998) Public sector economics, 4th edn 1990. Blackwell Publishers, Oxford
- Buchanan J, Tullock G (1962) The calculus of consent: logical foundations of constitutional democracy. University of Michigan Press, Ann Arbor
- Coleman J (1980) Efficiency, Utility, and Wealth Maximization. Yale Law School Legal Scholarship Repository. Retrieved from http://digitalcommons.law.yale.edu/fss_papers/4202
- Craig P (1991) Constitutions, property and regulation. *Public Law* 538–554
- Downs A (1957) An economic theory of democracy. Harper, New York
- Feintuck M (2010) Regulatory rationales beyond the economic: in search of the public interest. In: Baldwin R, Cave M, Lodge M (eds) The Oxford handbook of regulation. Oxford University Press, Oxford
- Garnier B (ed) (2004) Public interest. In: Black's law dictionary, 8th edn. Thompson West, St. Paul
- Hägg PG (1997) Theories on the economics of regulation: a survey of the literature from a European perspective. *Eur J Law Econ* 4(4):337–370. <https://doi.org/10.1023/A:1008632529703>
- Hale M (Sir) (1787) A treatise, in Three Parts. Pars Prima, De jure maris et brachiorum ejusdem. Pars Secunda, De portibus maris. Pars Tertia, Concerning the custom of goods imported and exported. Considerations touching the amendment or alteration of laws. A

- discourse concerning the Courts of King's Bench and Common Pleas. In: Hargrave A (ed) *Collection of tracts*, vol I
- Hantke-Domas M (2003) The public interest theory of regulation: non-existence or misinterpretation? *Eur J Law Econ* 15(2):165–194. <https://doi.org/10.1023/A:1021814416688>
- Harnay S, Marciano A (2011) Seeking rents through class actions and legislative lobbying: a comparison. *Eur J Law Econ* 32:293–304
- Public Interest in UK Courts (2011) Retrieved 30 July 2014, from <http://publicinterest.info/>
- Klitgaard R (1988) *Controlling corruption*. University of California Press, Berkeley
- Laffont J-J, Tirole J (1993) *A theory of incentives in procurement and regulation*. MIT Press, Cambridge, MA
- Lewin L (1991) *Self-interest and public interest in western politics* (trans: Lavery D). Oxford University Press, Oxford
- Lewis C (2006) In pursuit of the public interest. *Public Adm Rev* 66(5):694–701
- Middleton R (1996) *Government versus the market*. Edward Elgar, Cheltenham
- Mueller D (1989) *Public choice*. Cambridge University Press, Cambridge
- Parejo L (2003) *Derecho Administrativo*. Ariel Derecho, Barcelona
- Peltzman S (1976) Toward a more general theory of regulation. *J Law Econ* 19(2):211–240
- Posner R (1974) Theories of economic regulation. *Bell J Econ Manag Sci* 5(2):335–358
- Rawls J (1971) *A theory of justice*. Harvard University Press, Cambridge, MA
- Schmidt-Assmann E (2003) *La Teoría General del Derecho Administrativo como Sistema: Objeto y fundamentos de la construcción sistemática*. (trans: Bacigalupo M, Barnés J, García Luengo J, García Macho R, Rodríguez de Santiago JM, Rodríguez Ruiz B, ... Huergo A). Marcial Pons, Madrid
- Senior I (2006) *Corruption – the world's big C: cases, causes, consequences, cures*. The Institute of Economic Affairs, London
- Sola JV (2004) *Constitución y Economía*. LexisNexis Abeledo-Perrot, Buenos Aires
- Stigler GJ (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 2(1):3. <https://doi.org/10.2307/3003160>
- Takeovers: the public interest test – Commons Library Standard Note. (n.d.). Retrieved 30 July 2014, from <http://www.parliament.uk/business/publications/research/briefing-papers/SN05374/takeovers-the-public-interest-test>
- Tanzi V, Schuknecht L (2000) *Public spending in the twentieth century: a global perspective*. Cambridge University Press, Cambridge
- Weisbrod BA, Benjamin CB (1978) APPENDIX: Some concepts of the public interest and public interest law. In: Weisbrod BA, Handler JF, Komesar NK (eds)

Public interest law an economic and institutional analysis. University of California Press, Berkeley

World Bank (1997) Annual Meetings, WorldBank Group Issue Brief Corruption and Good Governance. Retrieved from www.worldbank.org/html/extdr/am97/br_corr.htm

Public Investments: Broadband

Nicola Matteucci
Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy

Abstract

After a long phase of liberalizations, privatizations, and deregulation, in telecommunications the policy pendulum has moved in the opposite direction, and the State is back in many countries. By mid-2000s, the deployment of first-generation broadband communications had reached high levels of diffusion and, thanks to market-friendly regulation, market competition has generally led consumer prices to decline. However, marked dynamics did not ensure universal service (especially in rural areas), nor the continuous upgrade of the networks, with the migration to higher capacity NGAN – two requisites increasingly demanded by downstream market developments (e-Services, cloud computing, IoT). This entry reconstructs the main market, Government and (cohesion) policy failures occurring in the sector, and the responses so far experimented in the EU, in an international perspective. In particular, it illustrates its State aid control system and analyzes the current trade-off that the new industrial policy is posing to the policy-maker in a liberalized sector.

Synonyms

[State intervention in the market](#)

Liberalization, Privatization, and Market-Friendly Regulation

Since the 1980s (in the USA) and then the 1990s in Europe, there was a mounting consensus toward market liberalization and privatization of public companies (see entries on ► [State-Owned Enterprises](#)) – especially those active in network industries and utilities. Telecommunications were one of the sectors most heavily interested, starting with the famous AT&T breakup in the USA. According to the received wisdom, the privatization of this sector should have brought the largest societal benefits – at least in terms of retail prices (Newbery 2004); a main motivation was that technological developments would have reduced the competitive bottlenecks that had been justifying public monopolies and heavy old-style regulation (such as rate of return); then, market liberalization would have spurred pro-competitive dynamics, be able to progressively deregulate the concerned sectors, and move from *ex ante* (regulation) to *ex post* (antitrust) norms. Such a received wisdom does inform the current European Regulatory Framework for electronic communications (ex Directive 2002/21/EC and its further amendments) (see entry on ► [Telecommunications](#) – if any has been planned for the European Regulatory Framework).

As a matter of fact, in parallel with the paradigm change unfolding at the political and institutional levels, a technological revolution was also occurring, characterized by the advent of digital technologies for signal coding, transmission, and reception, that would have been deeply transforming the economics of network communication industries – particularly that of television and telecommunications. In particular, thanks to digital technologies and growing equipment standardization, the legacy communication networks (*in primis*, copper, and cable TV ones) have been transformed in a multipurpose infrastructure and able to carry a converging array of distinct services: voice telephony, video content, and data. As a consequence of this technological breakthrough, telecommunication networks have progressively converged with traditional audiovisual media, and new hybrid communication platforms

and services have been introduced, such as IPTV (Internet protocol TV) (Matteucci 2016). Concerning the market consequences of the technological and institutional transformation, it is conventionally believed that the outcomes differ by market (services and countries), although presenting some broad regularities. In wireless services (*in primis*, mobile phones, where digital technology caused a relevant relaxation of the existing spectrum constraints), liberalization and privatization were unquestionably pro-competitive and elicited a unprecedented surge of new investments by entrants in both developed and developing countries, with the best results achieved where spectrum management was more effective. In fixed telephony services, the impact of the institutional transformation remained less clear-cut, due to the natural monopoly features still persisting along the distribution network (local loop), the varying effectiveness of the regulatory and policy mix, and the multifaceted implications of digital technologies. For example, Bacchiocchi et al. (2011) find that for EU-15 countries in voice services, regulation played a larger impact in driving down prices than privatization. On a similar vein, Florio (2004), after studying with cost-benefit analysis tools the famous British Telecom privatization in the UK, reconsiders the received wisdom and concludes that effective regulation may be much more important for the overall welfare than ownership regimes (public versus private) per se and that the latter had a small effect on the long-term dynamics of consumer prices and productivity – at least in the UK. Indeed, this evidence recalls what standard microeconomic theory predicts that an unregulated monopolist (or dominant firm) sets profit-maximizing prices, higher than competitive ones: hence, regulation may be the crucial aspect to look at.

In today's Internet economy, where consumer's usage is increasingly multipurpose and hybrid and requests higher bandwidth capacity, broadband services stand as the main reference product for the telecommunication industry and the policy-makers, replacing the older emphasis on voice services – despite the fact that no formal universal service provision currently covers the

first, differently from the second, according to the existing European Regulatory Framework. In first-generation broadband (conventionally classified as those services offering from basic to premium data transmission rates, respectively, from 2 to 20 Mbps in downloading, and epitomized by ADSL technologies), we have now accumulated a sufficient evidence on the relation between institutional setting and industry performance, especially for EU countries. Here, liberalization together with wholesale access regulation brought about a higher degree of static competition that generally resulted in lower consumer prices. In turn, a favorable retail price dynamics – prevalent in non-concentrated markets – translated into higher take-up: for example, Distaso et al. (2006) found that, in EU-14, lower local loop unbundling prices (henceforth, LLU, the most important type of wholesale access regulation) encouraged the take-up of broadband subscriptions, even though, on the side of the generation of new investment, the detected effects were generally negative (Cambini and Jiang 2009). In the EU, the widespread adoption of the “ladder of investment” regulatory model (henceforth LOI, setting access levels and terms in a way to progressively incentivize entrants to move from service-based to infrastructural-based competition) is believed by its proponents to have contributed to rapidly create competitive retail markets while hopefully stimulating gradual investment strategies by new entrants aimed at building alternative networks (Caves 2014). However, among the encountered drawbacks, wholesale access regulation and LOI could not cater for providing adequate stimuli to rapid network rollout, differently from what achieved with infrastructural competition, where the latter has been available (mostly between DSL copper and cable networks) (Briglauer et al. 2014).

Another fundamental stylized fact in the industry has been that, after liberalizations and privatizations, competition (both service and infrastructural) and private investment for rolling out digital networks first and foremost unfolded in urban areas, while less populated (rural) and socioeconomically disadvantaged areas across

the EU were left behind, because of the higher network deployment costs and of the insufficient revenues characterizing these territories: these conditions typically configure cases of market failures (a situation when an output or activity yields net social returns higher than private ones, so that its market production is suboptimal) and/or problems of social and territorial cohesion (the existence of large GDP differences and resource unbalances between citizens and between regions). In this case, regulation cannot help so much, since even the facilities of incumbent operators (those having significant market power, henceforth SMP) are frequently unavailable or insufficient in these laggard areas, and no universal service provision is in place for broadband services in the EU.

As a consequence, rural and marginal areas accumulated a substantial gap – both concerning the timing and the quality of the communication services offered and subscribed. In particular, in 2013 (the deadline for reaching the first infrastructural objective of the Digital Agenda for Europe, henceforth DAE – EC 2010), the broadband coverage of the EU homes was nearly completed when considered in nominal terms and including all the potential networks (wireless and wired, fixed and mobile) of the first generation; however, the coverage was sensibly lower, on average, if only the more performing fixed technologies (xDSL, cable, and the fixed wireless WiMax) are taken into account: nominally, the latter was at 97.2%, but in real terms, it was often substantially inferior (depending on country and legacy network characteristics). However, the gap of rural areas was still relevant: in fact, in the same year, the nominal fixed coverage of the EU rural areas was lagging at 89.8% of the population (79.9% in 2012) (data from Digital Agenda Scoreboard 2014). Then, if one considers the NGAN achievements (next-generation access networks, including at that time broadband technologies such as VDSL, Cable DOCSIS 3.0, and FTTP), figures were much more disappointing, with fast broadband (ensuring transmission rate >30 Mbps in downloading) covering on average only 62% of the EU homes, but again dropping to only 18.1% in rural areas (with the coverage mostly

attributable to VDSL deployments, involving the incremental upgrade of legacy copper-based networks with segments of fiber optic).

For these reasons, in the current phase of the sector, where a timely rollout of new access infrastructure assumes a paramount societal importance (especially for the wired ultrafast networks, which however involve the highest sunk costs for cabling the terminal link), the policy pendulum is shifting from old-type wholesale access regulation (LLU and LOI) to new models (incentive based and geographically differentiated) and, increasingly, to active industrial policies. The latter, involving an original role for investment with public money, is especially needed if one wants to foster the deployment of wired ultrafast networks to comply with the stated targets and deadlines of the DAE that envisage an ubiquitous and fast-increasing connectivity capacity of the territories as a way to enhance the market competitiveness and socioeconomic growth of the EU society.

The State Comes Back in Telecommunications

Industrial policies aimed at solving the digital divide started to be experienced with first-generation broadband, both in Europe and elsewhere, before igniting a hot economic policy debate with second generation (NGAN). Since 2003, a certain number of small-scale (regional or municipal) State aid (henceforth, SA) measures supporting network deployments had been authorized by the EU Commission and mainly financed with European structural and investment funds or national finances: these measures gradually unfolded across countries as soon as the market diffusion path of broadband services (supply-side coverage and demand-side subscriptions) was progressing and encountering the first difficulties and failures, in relation to the different country sector's maturity (the first measure approved was the Cumbria project, in the UK, EC 2003). These policies acquired momentum in coincidence with the US subprime mortgage financial crisis and the following worldwide macroeconomic

turbulences. In the EU, the first noticeable initiative came out in 2008: the European Economy Recovery Act (EC 2008) loudly announced the introduction of specific measures and funds (roughly EUR 1 billion) earmarked for solving the digital divide affecting rural areas, focusing on the supply-side (investment in infrastructure). Similar and sometimes more ambitious initiatives were adopted by other industrialized countries, along a clear "interventionist" industrial policy agenda. In the USA, in 2009, the extensive budget of the American Recovery and Reinvestment Act included a sum of \$7.2 billion for grants and loans targeting broadband unserved and underserved areas (LaRose et al. 2014); this Act had a clear protectionist flavor, arriving to mandate also a sort of "Buy American" condition for funding, that provoked strong opposition among NAFTA partners (Canada). Other developed countries, including Australia and New Zealand – famous for their usually orthodox neoliberal policy agendas – introduced similarly ambitious plans of public intervention; then, the countries experiencing the highest public support were Japan and South Korea, two recognized worldwide ICT (Information and Communication Technology) leaders. Finally, active industrial policies for broadband were also enacted in BRICS and developing countries – with mixed characteristics and results (for Latin America, Galperin et al. 2013).

In the EU, the awareness of the policy-maker about the real extent of the infrastructural digital divide grew slowly, in parallel with the improvements of the surveying methodologies and available statistics. As a matter of facts, several original initiatives of country territorial mapping began to show that the network coverage gaps were numerous, spotty and very granular, and far from being limited to marginal rural areas – as an uncritical "market failure" rationale had let to assume. On the contrary, in some member states like Italy, historically characterized by dispersed models of urbanization and business location, coverage gaps and service disruptions happened to be very frequent also in strategic socioeconomic areas (export-oriented industrial districts and leading touristic locations), besides being present at the

edges of metropolitan and other urban areas (Matteucci 2014).

At the European level, the draft of the 2009 Broadband Guidelines (EC 2009; updated by the 2013 version, EC 2013) and the DAE (EC 2010) marked a radical step in elaborating a less rhetoric and more pragmatic approach, with respect to the earlier volitive but rather unrealistic agendas for Information and Communication Technologies (henceforth ICT) and Research and Development (R&D) support – from the Lisbon strategy in 2000 to the i2010 action plan in 2005. In particular, the Broadband Guidelines and DAE first paid a special attention to the worrying sluggishness showed by the NGAN market transition in Europe, with respect to the reference areas (US and, above all, the Asian leaders). As a consequence, in this period, some EU member states started to frame systemic and more ambitious nationwide broadband plans, designed to solve the digital divide with mixed policies, financed by a variety of funds and instruments (including public-private partnerships). Finally, in 2015, acknowledging the continuing relevance of these policy priorities, the Commission confirmed the orientation of its policy issuing another “Investment Plan for Europe” (also called “Juncker Plan,” EC 2014a), aimed to stimulate the EU economy with an initial budget of EUR 315 billion devoted to new investment – including NGAN networks; then, in late 2016, the Commission President announced that the European Fund for Strategic Investments (EFSI) (to be employed for the Investment plan) would be doubled by 2022.

The more interventionist policy phase that began at the end of the 2000s was motivated by a series of reasons – both structural and short term. A main one reappraises the traditional macroeconomic benefits of countercyclical fiscal policies (typically advocated by neo-Keynesians), after years of massive consensus on the neoliberal rhetoric and the EU “austerity” economic policy orientation, increasingly put under pressure after the 2008’s financial crisis (Schmidt and Thatcher 2013). Then, more micro-founded explanations acknowledge the specific positive impact of broadband investment on driving innovation, GDP growth, productivity, and competitiveness

(Analysys Mason, Tech4i2 Ltd 2013) – an impact that is proved to be sizable, yielding estimates for the demand multiplier within the range 1.2–1.5. The micro-founded dynamics at work are several. Besides increasing directly the GDP through public expenditures, broadband networks, as many ICT, are “general purpose technologies,” thereby possessing key valuable characteristics enhancing their socioeconomic impact. First, broadband infrastructure is particularly strategic since it channels and exchange increasing amounts of codified information and knowledge that, being public goods and involving positive externalities, are a primary source of innovation and economic growth. Second, broadband investment both spurs large complementary expenditures in other goods and services in the downstream economy. Third, broadband services are highly pervasive, and they enable further innovation and productivity gains in the using sectors, also due to the powerful network externalities and coordination dynamics (of the type “chicken egg”) unfolding with downstream ICT and not-ICT sectors.

At the same time, investment in broadband networks creates its own demand and market (as often most types of infrastructure do, in a “Say’s law” vein); for this, infrastructural investment in a free market economy may well experience coordination failures and holdup behavior by risk-adverse network operators. As a matter of fact, in the post-liberalization phase, the generalized drop in the investment rate experienced in the sector can be interpreted as driven by a rational strategy of privatized fixed network incumbents (now subject to the stock markets dividend constraint): they first aim to avoid cannibalization between NGAN and existing services and, second, opt for incremental and less risky investment paths (e.g., using cheaper FTTC architectures, instead of FTTH ones) to migrate to the new NGAN paradigm. Hence, in network sectors severe market, failures can occur, leading to service under provision and delays contrary to the public interest, however expressed. As a main example, in the current process of digitalization of the public administration (Seri et al. 2014), no forced migration (sunset dates) to digital services like e-Government, e-Health, or e-Participation can be reasonably mandated by

public authorities until all the population had been serviced with adequate broadband access to the Internet. Similarly, promising market developments such as cloud computing and the Internet of Things (IoT) require ubiquitous and constantly upgraded broadband capacity, but this frontier can be severely retarded – or even compromised – by the extent of the old and new types of infrastructural digital divide.

Indeed, for the EU to compromise on reducing the digital divide while not interfering with the outcome of the post-liberalization phase has proved much more difficult than elsewhere, due to the constitutional status of the general prohibition of SA, ex Art. 107(1) TFEU, and its stringent exceptions. At the end, the target of completing the coverage of broadband, although not submitted to a formal provision of universal service under the current Regulatory Framework, was codified as two distinct infrastructural targets of the DAE (EC 2010): to complete the coverage of the EU population (1) with first-generation services by 2013 and (2) with second-generation ones

(NGAN) by 2020. These policy targets can be achieved through generous market incentives (various types of subsidies) and active policy-making; the latter, for some member states, initiated a new challenging season of direct public ownership of the built infrastructure.

EU State Aid Control for Broadband

The current version (2013) of the Broadband Guidelines (henceforth Guidelines) represents the *summa* of the relevant EU policy-making, detailing all the admissible forms of public intervention in the sector (“vertical” SA). Acknowledging that liberalization and privatization profoundly reshaped the telecom market, the Commission first illustrates (EC 2013: Artt. 9–54) the main principles and instruments that any measure involving public money has to follow, to comply with EU competition norms. Basically, there are three main instruments for public financial intervention (see Table 1), according to

Public Investments: Broadband, Table 1 Public support for broadband networks: instruments available in the EU

Instrument	EU founding norm (s)	Goal(s)	Main criterion of scrutiny	Example case
State aid	Art. 107-9 TFEU	Prohibition of any unlawful form of SA, subject to predefined exceptions	Use of State resources, economic selective advantage potentially distorting competition and affecting internal market trade	See the List of the Commission’s decisions (source)
Market economy investor principle	Art. 345 TFEU	Equal treatment between private and public regimes of property ownership and entrepreneurship	Public financial support to an undertaking paid under normal market conditions	Case C 53/06-NL <i>Citynet Amsterdam-Investment (FttH) network</i>
Service of general economic interest	Article 106 (2) TFEU, the “Altmark case” jurisprudence, and “SGEI package”	Guaranteed supply of a SGEI	Imposition of public service obligations in return of specific compensation	Case N 331/08-F <i>THD Hauts-de-Seine</i>
Other administrative and regulatory measures	Regulatory Framework, SA case law, Directive 2014/61/CE	Achieve better market transparency and coordination and investment costs savings	To provide more information, transparency, open access to and efficient construction of network infrastructure	National legislation and regulation implementing the Directive 2014/61/CE

Source: our compilation from various sources, including EC (2013, 2014), and the List of the Commission’s decisions (http://ec.europa.eu/competition/sectors/telecommunications/broadband_decisions.pdf)

the Treaty on the Functioning of the European Union (TFEU, EU 2012) and the associated hard and soft law norms: (1) the SA norms, (2) the application of the Market Economy Investor Principle (henceforth, MEIP), and (3) the imposition of the special regime of Service of General Economic Interest (SGEI). Further supporting initiatives may concern instruments not falling within the meaning of SA, such as administrative and regulatory measures aimed at facilitating market transparency and coordination. A main example is the “Cost Reduction” Directive (n. 2014/61/CE, EU 2014), due to be implemented by the beginning of 2016, that foresees four main pillars of normative harmonization and promotion of best practices: (1) mandating reuse of existing physical infrastructure to avoid inefficient duplication, on behalf of market entrants; (2) enforcing coordination and transparency of planned civil works, catering for efficient construction; (3) enabling faster, simpler, and more transparent concession of administrative permits; and (4) modernizing norms applicable to new buildings, to be equipped with NGAN-ready facilities.

MEIP and the SGEI are the residual cases: while MEIP concerns public intervention carried out “at market terms” (in commercially profitable areas), SGEI *de facto* regards not profitable areas (“white areas”; see *infra*). Both have been used rather infrequently, due to the application complexities and the potential institutional inconsistencies and normative loopholes involved. Basically, the MEIP case follows the (now somehow neglected) principle of equal treatment between private of public property ownership and entrepreneurship and involves the direct or indirect attribution of public capital (or other tool) to an undertaking under circumstances which mirror normal market conditions (replicating the case of a standard private investor), thereby not qualifying as SA under Art 107 TFEU. The SGEI case, in turn, requires to demonstrate that existing private operators cannot offer in the near-future adequate broadband coverage in a universal way, thereby excluding a portion of the population. Under this prerequisite, an undertaking may be entrusted with the operation of an appropriately defined SGEI, receiving in return a public service

compensation. However, the accompanying conditions are quite demanding, since they are aimed at excluding any possible market interference with existing operators: for example, the motivation of inadequate digital connectivity for businesses in the targeted area would not suffice nor would be possible to entrust the SGEI broadband undertaking with retail operations. In other terms, SGEI feasibility in this sector needs to be limited to the network construction and the supply of wholesale access services through the public offer of the built passive infrastructure, neutrally open (EC 2013, Art. 21–25), whenever such an offer has been demonstrated as being unavailable from private operators. It follows that the practical feasibility of setting a SGEI nowadays is severely limited in any sector that underwent a massive liberalization process, like in the case of telecom and broadband; equally, as it is now, it is very difficult to use the SGEI instrument to relaunch the paradigm of the public company, where this has been privatized.

In essence, the SA norms (see entry ► [State Aids and Subsidies](#)) are the default system of public intervention – also for broadband networks. Art. 107(1) TFEU states a general prohibition of any unlawful form of SA, subject to predefined exceptions (ex Art. 107(2) and (3)), aimed at safeguarding the normal functioning of the internal market. The latter goal is also targeted by the sector liberalization process, and in this sense, the Guidelines match the general SA norms with the principles of the Regulatory Framework, setting binding requisites for efficient and not-distorting state intervention. Accordingly, any measure financed with public money must be preceded by a detailed mapping of the territory, aimed at distinguishing three types of areas: “black,” where infrastructural competition between at least two network operators is present; “gray,” where the service is provided by a unique network; and “white,” where no service is available, now or in the foreseeable future (3 years). Potentially, public intervention is admissible in white areas, while it becomes more difficult to authorize in gray ones and normally is not admitted in black areas (except for ultrafast NGA networks (offering at least 100 Mbps speeds), where the 2013 Guidelines

opened a possibility upon proving the performance “step change” with respect to existing fast NGAN, EC 2013; Artt. 82–85). Mapping of the territory must be conducted at the national level through public consultations with telecom operators, and, once the measure is authorized, any public fund needs to be granted on the basis of open, transparent, and competitive procedures, where bidding operators cannot be discriminated on the basis of the solutions employed (“technological neutrality” principle). Then, once built, the publicly owned or the private infrastructure that got the subsidy (both arrangements are possible) is subject to a reinforced wholesale access regime – independently of the SMP qualification of the winner – and network operations must be monitored for a certain period, to avoid overcompensation (“claw back” clause on the extra profits).

Over time, the Commission’s practice on broadband has evolved and adopted a “more stringent economic approach,” aimed at enhancing its market-friendliness (in line with the 2005 State Aid Action Plan and continuing with the 2012 State Aid Modernization Initiative). This also implied a growing complexity of the relevant soft law that the Commission has somehow tried to streamline by pushing forward the convergence with other “horizontal” SA norms, for example, the recent inclusion of the easier cases (involving only broadband “white” areas) in the General Block Exemption Regulation or the solution of the previous loopholes contained in the 2007–2013 Regional Aid Guidelines, which had provided a more generous authorization pathway.

Since 2003, a large number of SA cases have been analyzed by the EU Commission: as of July 2016, 148 decisions (mostly positive) have been finalized, with a growing presence of country-wide measures (national plans) for NGAN deployment, involving higher budgets; the latter fact is attributable to the larger deployment costs involved in second-generation broadband – mostly, civil engineering works for digging and laying passive infrastructure along the access network (such as ducts, wires and antennas), arriving to account for 70–80% of the total investment

costs. For the largest EU member states (the “big 5”), the representative budget of the most recently authorized national NGAN plans ranges from EUR 2 to 4 billion. On overall, the received wisdom predicates that the net balance of the Commission’s activity is positive: for example, Chirico and Gaál (2014) notice that only a tiny minority of its numerous decisions was appealed.

As a matter of fact, the net balance of the SA activity is multifaceted and much more complex to gauge – even sticking to strict economic considerations. First, economic approaches used in antitrust theory and practice remain prevalently static, so that they miss to consider alternatives of public intervention that may score higher when examined in a truly dynamic setting. Sometimes, even some SA pro-competitive principles (such as that of technological neutrality) can yield unintended effects: while the first version of the Guidelines was inspired by a truly “driving” conception of SA policy (Gómez-Barroso and Feijóo 2012), there are good reasons to believe that the current *de facto* policy assimilation between wireline and wireless NGAN (adopted since EC 2013) can give incentive to less ambitious rollout plans and networks. Second, EU competition law continues to present cases of institutional conflict and normative loophole with other domains – *in primis* with regional and urban policies – and their possibly diverging objectives of social and territorial cohesion (Colomb and Santinha 2014). Third, while the standard Commission’s scrutiny is based on *ex ante* evaluation, there is a real scarcity of evidence and literature on *ex post* evaluation of the implemented measures, which in turn is aggravated by the enduring lack of statistical evidence surrounding the sector’s operations and the heterogeneous member states’ institutional performance – especially at the disaggregated territorial level. Indeed, some broadband SA measures (e.g., EC 2014b) have been encountering severe implementation retards and even problems of institutional capacity that led to a perverse mix of market and Government failures, delaying a prompt achievement of the intended goals (especially, from a cohesion point of view).

Conclusion and Future Directions

To sum up, taking into account the worldwide trend toward higher State intervention on the market and active industrial policy, and their stringent trade consequences (including cases of “beggar-thy-neighbor” policies), it is difficult to predict whether the delicate policy equilibrium so far achieved in SA practice for liberalized network sectors in EU can be firmly maintained in the future – especially in the post-Brexit era. In particular, SA policy could be affected by much stronger pressures and conflicts from member states than before. Moving to a larger picture, in the most pessimistic scenario, due to the constitutional nature of the SA norms, the potential imbalances arising from its application could even arrive to damage the social and institutional cohesion of the EU.

References

- Analysys Mason, Tech4i2 Ltd (2013) The socio-economic impact of bandwidth. Report for the European Commission, London. Available at: <http://ec.europa.eu/digital-agenda/en/news/study-socio-economic-impact-bandwidth-smart-20100033>
- Bacchiocchi E, Florio M, Gambaro M (2011) Telecom reforms in the EU: prices and consumers' satisfaction. *Telecommun Policy* 35(4):382–396
- Briglauer W, Fröbing S, Vogelsang I (2014) The impact of alternative public policies on the deployment of new communications infrastructure – a survey. *Rev Netw Econ* 13(3):227–270
- Cambini C, Jiang Y (2009) Broadband investment and regulation: a literature review. *Telecommun Policy* 33(10–11):559–574
- Cave M (2014) The ladder of investment in Europe, in retrospect and prospect. *Telecommun Policy* 38(8–9): 674–683
- Chirico F, Gaál N (2014) A decade of State aid control in the field of broadband. *Eur State Aid Law Q* 1:28–38
- Colomb C, Santinha G (2014) European Union competition policy and the European territorial cohesion agenda: an impossible reconciliation? State aid rules and public service liberalization through the European spatial planning lens. *Eur Plan Stud* 22(3):459–480
- Digital Agenda Scoreboard (2014). Available at: <https://ec.europa.eu/digital-single-market/en/digital-scoreboard>
- Distaso W, Lupi P, Manenti FM (2006) Platform competition and broadband uptake: theory and empirical evidence from the European Union. *Inf Econ Policy* 18(1):87–106
- European Commission (EC) (2003) Cumbria broadband – project access – advancing communication for cumbria and enabling sustainable services, Case N282/2003 – United Kingdom, 10.12.2003, C(2003) 4480fin, Brussels
- European Commission (2008) A European economic recovery plan, communication from the Commission to the European Council, 26.11.2008, COM(2008)800 final, Brussels
- European Commission (2009) Community guidelines for the application of State aid rules in relation to rapid deployment of broadband networks, Communication 2009/C 235/04, 30.9.2009, Brussels
- European Commission (2010) A digital agenda for Europe, COM(2010)245, 19 May, Brussels
- European Commission (2013) EU guidelines for the application of State aid rules in relation to the rapid deployment of broadband networks, Communication 2013/C 25/01, 26.1.2013, Brussels
- European Commission (2014a) An investment plan for Europe, Communication from the Commission, COM/2014/0903 final, Brussels
- European Commission (2014b) Prolongation of the National Broadband Plan Italy, Case SA.38025 (2014/NN) – Italy, 11.12.2014, C(2014) 9725 final, Brussels
- European Union (2012) Consolidated version of the Treaty on the Functioning of the European Union, 2012/C 326/01, 26.10.2012, Brussels
- European Union (2014) Directive 2014/61/EU of the European Parliament and of the Council of 15 May 2014 on measures to reduce the cost of deploying high-speed electronic communications networks, 23.5.2014, Brussels
- Florio M (2004) The great divestiture. Evaluating the welfare impact of the British privatisations 1979–1997. The MIT Press, Cambridge, MA
- Galperin H, Mariscal J, Vicens FM (2013) One goal, different strategies: an analysis of national broadband plans in Latin America. *Info* 15(3):25–38
- Gómez-Barroso J, Feijóo C (2012) Volition versus feasibility: state aid when aid is looked upon favourably: the broadband example. *Eur J Law Econ* 34(2): 347–364
- LaRose R, Bauer JM, DeMaagd K, Chew HE, Ma W, Jung Y (2014) Public broadband investment priorities in the United States: an analysis of the broadband technology opportunities program. *Gov Inf Q* 31(1): 53–64
- Matteucci N (2014) L'investimento nelle reti NGA a larga banda: la 'questione settentrionale' (The investment in NGA broadband networks: the 'Northern question'). *Econ Polit Ind J Ind Bus Econ* 41(4):9–25
- Matteucci N (2016) Standards, IPR and digital TV convergence: theories and empirical evidence. In: Lugmayr A, Dal Zotto C (eds) Media convergence handbook, vol I, 1st edn. Springer, Berlin
- Newbery D M. (2004) Privatising network industries. Cesifo working paper n. 1132, Feb

- Seri P, Bianchi A, Matteucci N (2014) Diffusion and usage of public e-Services in Europe: an assessment of country level indicators and drivers. *Telecommun Policy* 38(5–6):496–513
- Schmidt VA, Thatcher M (eds) (2013) *Resilient liberalism in Europe's political economy*. Cambridge University Press, Cambridge

Public-Private Partnerships

Luciano Greco

Department of Economics and Management and CRIEP, University of Padua, Padua, Italy

Definition

Public-private partnerships (PPPs) are organizational forms involving public and private institutions and aiming at the provision of assets, goods, or services that, to a large extent, are relevant in terms of public interest. Alternatively, the same tasks (e.g., construction, operation, maintenance, financing) can be pursued by full-fledged public organizations. The nature of tasks and involved institutions (e.g., for-profit firms, nongovernmental organizations (NGOs), governmental agencies) and the constraints shaping interactions among partners (e.g., contract or information incompleteness) determine different effects as regards the distribution of risks and payoffs among partners and allocative efficiency.

Economics and Institutions

Several concepts of PPPs have been introduced by practitioners and academics. A comprehensive perspective about PPPs can be found in the management literature. Kivleniece and Quélin (2012) define “*public-private ties [...] as any long-term collaborative relationships between one or more private actors and public bodies that combine public sector management or oversight with a private partner's resources and competences for a direct provision of a public good or service.*” In such a rather extensive definition, we may include

a quite large array of institutional arrangements and activities, such as procurement of public infrastructures, collaborations to face societal challenges (e.g., poverty or disease eradication in developing countries), or research-driven joint ventures (Perkmann and Schildt 2015).

In line with the purpose of this essay, we focus on a narrower concept of PPPs as “*an agreement by which the government contracts a private company [or a consortium of firms] to build or improve infrastructure works and to subsequently maintain [and operate] them for an extended period of time [...] in exchange for a stream of revenues during the life of the contract*” (Engel et al. 2014, p. 2). The motivation is twofold. First, the theoretical (particularly, contract theory) literature on PPPs has primarily focused on the optimal design of incentive schemes and on the assessment of costs and benefits of *outsourcing* of important phases of public investment projects. However, the bulk of findings of these contributions can be extended to non-infrastructure PPPs. Second, the most common and recurrent use of PPP concept among practitioners is related to public infrastructures (several initiatives of public institutions in different countries are going to change this: e.g., the EU Commission has recently introduced the concept of *research public-private partnerships* which are joint ventures between public and private institutions to pursue common interest tasks related to large investments in research).

In the public debate, PPPs are often confused with privatizations. In both cases, the government may transfer public assets and activities to private institutions (possibly, for-profit firms) who are then in charge of running them. Nevertheless, two differences between such organizational forms are crucial. First, privatized assets and activities are not (anymore) part of government-specific tasks, while PPP infrastructures and services are (still) part of them. For example, a prison or a barrack continues to be essential facilities of law-and-order or defense governmental policies, even though these are managed by a PPP agreement (instead of a governmental agency). On the contrary, a privatized firm or building serves almost exclusively private purposes (e.g., profit

maximization). In many cases, government continues to play a supervisory role on privatized firms that, on a normative point of view, is not very different with respect to that on other firms in the same sector (this is quite clear in the case of competitive markets, but the same argument holds also for regulated sectors, where governmental authorities play just a market-failure-correction role). Second, privatizations are intended to be forever (unless major political changes foster *renationalization*), while PPPs feature (long but) finished terms, whereby assets and activities are transferred back to government at the end of contracts.

The previous argument has important policy implications. Consider that the government outsources to private institutions activities featuring a strong public interest (e.g., construction and operation of a hospital in countries where healthcare is publicly provided). Independent of the specific organizational setting, such an arrangement is a PPP. Indeed, outsourcing does not divest government of its role, nor it diminishes its responsibilities as regards final outcomes. For example, in the case of healthcare PPPs, any pitfall in the functioning of a hospital implies legal, economic, and political responsibilities for the governmental authority in charge of healthcare provision.

Practitioners' View

PPPs – as above defined – are a well-established practice to construct and operate public infrastructures in several countries (Bezançon 2005; Engel et al. 2014). Practitioners from public institutions and private business have pointed out three channels by which PPPs may improve social welfare, as compared to traditional, government-led frameworks:

1. *Efficiency gains* may derive from enhanced management of tasks – e.g., design, construction, and operation – and risks (many contributions in economics and management literatures emphasize innovation as an autonomous source of dynamic efficiency gains; for the purpose of this essay, we classify it among efficiency gains from enhanced task management).

2. The government may be able to attract and channel private financing into initiatives of public interest, thus relieving the public finance (i.e., *financial leverage*).
3. Government infrastructure policies are affected by perverse political incentives (e.g., investments in “white elephants,” procrastination of maintenance expenditures, expenditure political cycles) that may be corrected by the involvement of private partners (i.e., *market discipline*).

The international experience has shown that such benefits are not systematically delivered by PPPs (Engel et al. 2014; Saussier 2015). Also, in the last 30 years, these organizational forms have been characterized by growing financial and institutional complexity. For these reasons, many countries have introduced public oversight mechanisms. A central task of the latter is the *ex ante* assessment of the balance between costs and benefits (i.e., the so-called value for money) of PPPs, as compared to traditional public organization (Burger and Hawkesworth 2011).

In the last 20 years, different strands of academic literature have investigated the conditions under which the abovementioned benefits are likely to materialize.

Contract Theory Literature

As argued, complexity has become a distinguishing feature of PPPs. For example, the most prominent form of such arrangements, the *project financing* technique, puts a *project company* – also called *special purpose vehicle* or *consortium* – at the center of “a web of contracts aimed at distributing tasks, risks, costs, and revenues among all (private) partners” (Greco 2015). This explains why, in the last decades, contract theory contributions have been quite useful to elicit the main drivers of potential benefits of PPPs.

The standard analytical framework to assess the role and the features of PPPs is based on a *sequential moral hazard* problem (Hart 2003; Martimort and Pouyet 2008; Iossa and Martimort 2015). In this essay, we abstract from issues related to the *selection* of firms, and we focus

only on the post-contractual problems that are typical of long-term interactions like PPPs. In the simplest case, the public project cycle can be represented by two sequential tasks: in the first phase, a public infrastructure (e.g., a highway) has to be built; in the second phase, the infrastructure has to be operated and maintained to provide services (e.g., freight and passenger transportation). The social value of the public infrastructure is determined by the investment effort, which is implemented in the building phase, and by the management effort, which is implemented in the operation phase.

As a matter of fact, governments need external resources and competences to build and, in many cases, to manage public infrastructures. In this context, the main policy issue is whether it is better to outsource the two tasks to different agents (i.e., a building firm, in the first phase, and an operating-and-maintenance firm, in the second phase) or to a single agent (i.e., a consortium of firms that are specialized in different tasks). In the literature, the first setting – based on sequential contracting – is often called *unbundling* of tasks, as opposed to the alternative *bundling* of tasks. The latter represents, in a very stylized fashion, the basic contractual structure of PPPs.

A Basic Model

To grasp the essential issues of the choice between bundling and unbundling – which explains efficiency gains from PPPs – let us consider a very simplistic model. The quality of public infrastructures is almost always multidimensional; however, for our purposes, we consider the monetary value of the capacity of the infrastructure to satisfy end users and, possibly, to produce positive externalities on the whole society. Thus, the investment in infrastructure (say, highway) quality can be either high – i.e., $I = 1$ – or low, i.e., $I = 0$, and generates a gross social benefit $B = \tilde{B} + bI$, where $\tilde{B}$ is the exogenous (possibly, random) benefit generated by a minimum-quality highway and $b > 0$ is the marginal benefit of increasing the quality of the highway.

During the management phase, the operation and maintenance costs are $C = \tilde{C} - \delta I$, where $\tilde{C}$ is the exogenous (possibly, random) cost

component and δ is the marginal impact (or externality) of the first-phase investment on operation-and-maintenance costs. In most contract theory models on PPPs, the latter plays a crucial role: $\delta > 0$ implies that the first-phase investment is also beneficial in terms of second-phase infrastructure management; such an impact is negative when $\delta < 0$. A couple of examples may clarify this point: investments in technologies for touchless tolling improve the service quality of a new (or newly refurbished) highway but also reduce operation costs (i.e., labor costs), hence $\delta > 0$; conversely, investments to increase the number of lanes of a highway reduce congestion and, thus, increase the service quality but also maintenance costs during the operation phase, i.e., $\delta < 0$. For the sake of simplicity, we abstract from investments and managerial efforts that are typically implemented in the second phase to improve the social value of services provided by the infrastructure.

The *first best optimal outcome* is obtained by choosing $I = 1$ if and only if $b + \delta > 1$ – this result derives by the maximization of the social welfare: $B - I - C = \tilde{B} - \tilde{C} + (b + \delta - 1)I$. In the following, we assume that this condition is met; hence, investing in quality is always optimal from the point of view of the whole society. A benevolent government could attain the first best optimal outcome in a world where contracting with firms to outsource tasks is not affected by any significant informational or contractual imperfection. More technically, information is symmetric – i.e., investments are fully observable – and contracts are complete, i.e., investments can be verified by a court without any cost. In such a “wonderful world,” the choice between bundling and unbundling is immaterial. Any contractual scheme can be used to reach the first best. On the contrary, the choice between PPPs and sequential contracting becomes relevant when outsourcing is affected by substantial informational or contractual imperfections.

Asymmetric Information and Incentive Design

If the government cannot observe the first-phase investment of the builder (i.e., a building firm, under unbundling, or a consortium, under

bundling), neither it can infer it by the observation of B or C (because of the noise created by the exogenous random components $\tilde{B}$ and $\tilde{C}$), a classic moral hazard problem arises. In the absence of any incentive mechanism, the builder has no interest to invest in infrastructure quality. However, the government can design contracts to provide (monetary) incentives to the builder.

Under unbundled contracting, the government contracts with a building firm, in the first place, and then with another firm for operation-and-maintenance tasks. In our simplified setting, there is no moral hazard problem for operation and management. Thus, the second-stage contract is optimally a cost-plus one, $T_o = C$. In turn, the operator earns a zero profit and does not face any operation risk. As regards the building phase, the government can design an incentive payment scheme that is based on observable variables, $T_b(B) = T + tB$, where T is the fixed-price component of the payment to the builder and t is the marginal reward for each unit of extra benefits generated by the infrastructure. The government (correctly) anticipates the reaction of the builder, that is, to invest in quality if and only if $tb > 1$ – i.e., the building firm maximizes its profit: $T - I = T + t\tilde{B} + (tb - 1)I$. Thus, the government can control the decision of the builder by regulating the power of incentives: for $t > \underline{t}_u = \frac{1}{b}$, the builder invests.

Under bundled contracting, the government outsources both tasks to a single consortium of firms. The payment function is now $T_p(B, C) = T + t(B - C)$. Also in this case, the government anticipates that the consortium invests in high quality if and only if $t(b + \delta) > 1$ – i.e., the consortium maximizes its profit: $T - I - C = T + t(\tilde{B} - \tilde{C}) + [t(b + \delta) - 1 + \delta]I$. Again, the government can control the decision to invest of the consortium by a suitable power of incentives: for $t > \underline{t}_p = \frac{1 - \delta}{b + \delta}$, the consortium invests.

Let us remark that the minimum power of incentives to induce investment in infrastructure quality is different with respect to what we have in the unbundling case. In particular, *the power of incentives has to be stronger (respectively, weaker) under unbundling than under bundling*

if $\delta > 0$ (respectively, $\delta < 0$). The previous statement follows by the comparison of $\underline{t}_u$ and $\underline{t}_p$. This difference is irrelevant as far as increasing the power of incentives, i.e., t , is costless, which is the case when moral hazard is the only constraint that affects contracting between government and firms. In turn, both contractual forms afford the first best optimal outcome, and again the choice between bundling and unbundling is immaterial.

In the real world, facing higher-power incentives usually involves additional costs for the agents. For example, *risk-averse* firms require a risk premium (i.e., larger expected returns) to accept riskier contracts (i.e., larger t). In turn, to provide appropriate incentives to firms, the government has also to accept to pay larger expected transfers. A similar effect is determined by *limited liability* (or wealth) constraints faced by firms, which introduce lower bounds to government payments. By the previous arguments, we have:

Proposition 1 *Under asymmetric information and costly incentives, the government prefers bundling over unbundling (or the reverse) if and only if $\delta > 0$ (or $\delta < 0$).*

The basic idea is that the internalization of the externality between first-phase investment and second-phase operation costs reinforces (or weakens) contractual incentives if the externality is positive (or negative). A first policy implication of this literature is that *the enhanced management of synergic tasks is a crucial driver of efficiency gains from PPPs, because of information and financial frictions in the economy*.

Incomplete Contracts, Incentives, and Property Rights

In many real-world situations, observable variables – such as the quality of a hospital or a prison – cannot be assessed with a sufficient degree of precision to be verified by a court at negligible costs (Hart 2003). These quite common situations are characterized by incomplete contracts.

Incomplete contracting has dramatic impacts on the functioning of (alternative) contractual

schemes between government and firms. In particular, a *holdup* problem arises: in the terms of our model, the government is unable to commit to reward investments on the basis of observed (but unverifiable) infrastructure quality; firms rationally anticipate this outcome and do not invest.

Assuming that the ownership of the public infrastructure belongs to the government, the outcome of unbundling and bundling is equivalent to what would happen in the asymmetric information model (that we considered in the last section), when the only contract that the government can sign is such that $t = 0$. Thus, we have:

Proposition 2 *Under incomplete contracts and government ownership, the government prefers bundling over unbundling if and only if $\delta > 1$; otherwise, the two contractual schemes determine the same outcome.*

Contract incompleteness destroys the capacity of government to design (monetary) incentives for firms. In this situation, PPPs may improve social welfare if the intrinsic incentive underlying bundling is sufficiently strong (see Proposition 2).

In this framework, another important driver of incentives is the *distribution of rights* between government and partners (Bennett and Iossa 2006). To clarify this point, let us consider a specific case. If the public infrastructure is a building for (government) offices, it may clearly have alternative uses such that the quality of the infrastructure can be (fully) rewarded on the private market (e.g., a building for private offices). In this case, awarding ownership over such infrastructure to the builder reinforces its incentives to invest in quality. Indeed, once the infrastructure is built and its (high) quality observable, if the government does not pay to reward such quality, the private firm owning the infrastructure can renegotiate the terms of the contract with the government and sell the building on the private market. Anticipating such a situation, the builder invests in quality.

Four remarks are in order. First, the result with private ownership is the opposite with respect to the case of government ownership where, as above discussed, the builder has no incentive to

invest. Thus, the distribution of rights is a (rough) tool to regulate the power of builder's incentives.

Second, the main driver of the power of incentives (under private ownership) is the market value of alternative uses of the public infrastructure. In some cases (as in the previous example), the market value is likely to fully reward investments in quality. However, in the generality of cases, public infrastructures are not perfectly fungible for private purposes. In turn, only a fraction of investments in quality can be rewarded, which dampens incentives to invest. In some cases, no reward can be found on the private market for investments in public infrastructure quality (e.g., electricity networks). As the market value of alternative uses of the public infrastructure goes down, private ownership loses the capacity to convey incentives to invest in quality.

Third, a result similar to Proposition 2 can be obtained in any case: the incentive that is embedded in bundled contracting simply adds to the (possible) incentive that is determined by private ownership over public infrastructure.

Fourth, the contract theory literature has highlighted that the optimal design of property rights should also take into account the nature of outsourced assets and activities that, in turn, drives the interests of government and private partners in final outcomes (Besley and Ghatak 2001; Schmitz 2015).

The policy implication of these arguments is that *the distribution of rights compounds the design of monetary incentives to determine the results of alternative contractual schemes, in frameworks affected by contract incompleteness and asymmetric information.*

Extensions and Future Directions

The contract theory literature has confirmed that PPPs may deliver relevant efficiency gains whenever some form of externality exists between sequential tasks that constitute a public project cycle. Such an externality has not necessarily a technological nature, but it may also arise by financial constraints affecting private partners.

The previous arguments highlighted another crucial point, i.e., the proper distribution of risks among partners. Relying on a simple asymmetric information model, we pointed out that contracts that transfer risk to firms are useful to enforce stronger incentives but also involve additional costs for the government aiming at inducing firms to bear more risks (Martimort and Pouyet 2008; Iossa and Martimort 2015). In turn, it is crucial to assess which risks private contractors are able to handle and at which cost. Also, contractual designs should be conceived to improve appropriate risk sharing between government and firms (Engel et al. 1997).

This literature has also provided clear indications about another alleged benefit of PPPs identified by practitioners. An important result is that the *financial leverage* motive to undertake PPPs is clearly wrong (Engel et al. 2013). To put it in a very simple way, *the private financing of public infrastructures is a sort of public debt*. The reason why we observe, in the real world, a strong correlation between fiscal restraints and PPP investments is often argued to be linked to political incentives – e.g., in many countries public finance regulations may create incentives to undertake PPPs in order to bypass fiscal rules (Maskin and Tirole 2008). But such motives are not proven by compelling empirical evidence (Buso et al. 2017), and this point still needs deeper investigation.

The international experience shows that PPPs tend to work quite well in some sectors (e.g., highways, ports), while they tend to fail in others (Engel et al. 2014). On top of this, performance across countries appears to be mixed: the same kind of project may work properly in some countries (e.g., hospitals in the UK) and fail in others. From these considerations, two main questions arise. What mechanisms could explain the observed heterogeneity of PPP performance across countries and in time? Related to the previous question: how PPP performance should be measured?

Contributions to different strands of economics and management literature are trying to delve into the first issue. The institutional setting certainly plays a role to explain heterogeneity of PPP performance in time and across countries. However,

recent contributions have highlighted how, within the same institutional setting, organizational schemes based on sound business models are important keys toward PPPs' success (Villani et al. 2017). Other contributions have pointed out the role of agency problems and transaction costs within the consortium of private firms (Hoppe et al. 2013; Greco 2015). Further issues that still need improved understanding are related to renegotiations, which tend to reduce efficiency gains from PPPs (Engel et al. 2014), and are related to the lack of flexibility that, over long time spans and in particular sectors, may affect specifically PPPs (Martimort and Straub 2016).

As highlighted in the previous discussion, practitioners' debate has conveyed the idea that PPPs may be beneficial because of the market discipline they are able to introduce in public infrastructure planning. Following this argument, the mere involvement of private firms is a signal that a public infrastructure is viable from the economic and financial point of view. Clearly, this is a wrong idea: government assessment needs to precede the decision to undertake PPPs (Engel et al. 2014). However, this failed argument brings us to the very important issue of *how to measure the ex post performance of PPPs*. This issue is crucial both to assess theoretical predictions of the economics and management literature and to run policy evaluation studies. Since panel data econometrics is hardly useful in this case, because of the large heterogeneity across PPP contracts – even in the same sector, case studies or alternative econometric techniques (e.g., methods based on synthetic counterfactual) are required. This is surely an avenue of future empirical research that is worth to pursue.

Cross-References

- Externalities
- Franchise
- Governance
- Government
- Incomplete Contracts
- Institutional Economics
- Organization

- [Political Economy](#)
- [Privatization](#)
- [Public Goods](#)
- [State-Owned Enterprises](#)

References

- Bennett J, Iossa E (2006) Building and managing facilities for public services. *J Public Econ* 90(10):2143–2160
- Besley T, Ghatak M (2001) Government versus private ownership of public goods. *Q J Econ* 116(4):1343–1372
- Bezançon X (2005) Histoire du Droit Concessionnaire en France. *Entrep Hist* 38(1):24–54
- Burger P, Hawkesworth I (2011) How to attain value for money: comparing PPP and traditional infrastructure public procurement. *OECD J Budg* 2011(1):1–56
- Buso M, Marty F, Tran PT (2017) Public-private partnerships from budget constraints: looking for debt hiding? *Int J Ind Organ* 51(C):56–84
- Engel E, Fischer RD, Galetovic A (1997) Highway franchising: pitfalls and opportunities. *Am Econ Rev* 87(2):68–72
- Engel E, Fischer RD, Galetovic A (2013) The basic public finance of public-private partnerships. *J Eur Econ Assoc* 11(1):83–111
- Engel E, Fischer RD, Galetovic A (2014) The economics of public-private partnerships. A basic guide. Cambridge University Press, New York
- Greco L (2015) Imperfect bundling in public-private partnerships. *J Public Econ Theory* 17(1):136–146
- Hart O (2003) Incomplete contracts and public ownership: remarks, and an application to public-private partnerships. *Econ J* 113:C69–C76
- Hoppe EI, Kusterer DJ, Schmitz PW (2013) Public-private partnerships versus traditional procurement: an experimental investigation. *J Econ Behav Organ* 89(C):145–166
- Iossa E, Martimort D (2015) The simple microeconomics of public-private partnerships. *J Public Econ Theory* 17(1):4–48
- Kivleniece I, Quélin BV (2012) Creating and capturing value in public-private ties: a private actor's perspective. *Acad Manag Rev* 37(2):272–299
- Martimort D, Pouyet J (2008) To build or not to build: normative and positive theories of public-private partnerships. *Int J Ind Organ* 26(2):393–411
- Martimort D, Straub S (2016) How to design infrastructure contracts in a warming world: a critical appraisal of public-private partnerships. *Int Econ Rev* 57(1):61–88
- Maskin E, Tirole J (2008) Public-private partnerships and government spending limits. *Int J Ind Organ* 26(2):412–420
- Perkmann M, Schildt H (2015) Open data partnerships between firms and universities: the role of boundary organizations. *Res Policy* 44:1133–1143
- Saussier S (2015) Économie des Partenariats Public-Privé. *Développements Théoriques et Empiriques*. De Boeck
- Schmitz PW (2015) Government versus private ownership of public goods: the role of bargaining frictions. *J Public Econ* 132(C):23–31
- Villani E, Greco L, Phillips N (2017) Understanding value creation in public-private partnerships: a comparative case study. *J Manag Stud* 54(6):876–905

Q

Qualitative Choice Models

- [Discrete Choice Models](#)

Quality Entrepreneurship

- [Codes of Conduct](#)

Raising Rivals' Costs

Naoufel Mzoughi¹ and Gilles Grolleau^{2,3}

¹INRA, UR 767 Ecodéveloppement, Avignon, France

²Supagro, UMR 1135 LAMETA, Montpellier, France

³Burgundy School of Business – LESSAC, Dijon, France

Abstract

We describe the different possibilities that a protagonist has to start a process to raise rival's costs (RRC). We present the general RRC mechanism and necessary conditions to make it successful. We also expose the strengths and weaknesses of RRC theory.

Synonyms

Non-price predation

Definition

It is a strategy aiming to increase the cost of an entity's competitors in order to disadvantage and even exclude them from the market.

Raising Rivals' Costs: How It Works!

The original cases that founded the raising rivals' cost (RRC) theory relate to famous monopolization cases faced by the US Federal Trade Commission (e.g., Alcoa, Dupont de Nemours, Kellogg, and Standard Oil) where firms interfere in input or upstream markets in ways that reduce rivals' profits. In most cases, the premise of the RRC theory goes as follows: the predatory firm increases their competitors' costs by developing exclusive relationships with strategic suppliers, such as input overbuying, naked exclusion – where the supplier is committed not to sell inputs to competitors – and controlling the whole supply chain in order to prevent rivals from accessing consumption markets (Granitz and Klein 1996; Carlton and Perloff 1998; Scheffman and Higgins 2003). It is worthy to notice that the RRC strategy differs from price predation, because to be profitable, it does not require initial investments that need to be recovered (Scheffman 1992).

In order to present the RRC mechanisms and main results, let us consider the following market structure (Salop and Scheffman 1983; Church and

The authors of this essay previously investigated the strategic uses of environmental-related standards and eco-labeling schemes in order to raise rivals' costs (e.g., Grolleau et al. 2007). Hence, several parts are inspired from the mentioned previous works.

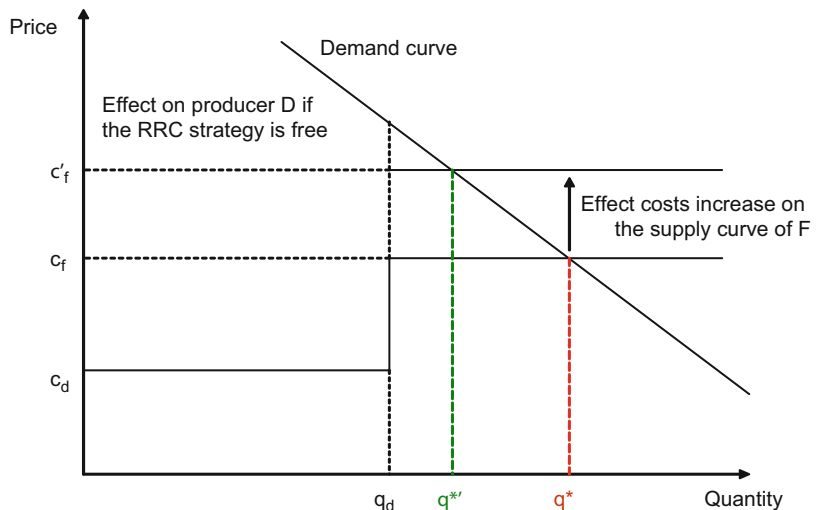
Ware 2000): a dominant firm (or group of firms), denoted D, a group of fringe rivals, denoted F, and a perfectly elastic supply for D. The marginal (and average) cost for the dominant firm c_d is lower than the marginal cost of the fringe firms, but this firm cannot produce beyond a given quantity, q_d . Moreover, assume the marginal (and average) costs for the fringe of competitive firms, denoted c_f , to be constant, but without any production constraint. The equilibrium price and quantity, respectively, denoted c_f and q^* , are presented in Fig. 1.

At equilibrium, the dominant firm is considered as inframarginal since its price is above average cost, despite a competitive market, which allows it to benefiting from inframarginal rents. Now assume that the dominant firm is able to increase the costs incurred by the fringe rivals. The equilibrium price will increase proportionally to the increase of the marginal cost of the fringe. However, the quantity produced by D will not change, while the quantity produced by F will decrease. In addition, assume that this cost is equally distributed across production units. The change in net profits of D can be written as $\Delta\pi_d^n = (\Delta c_f - \Delta c_d)q_d$. Hence, the profit of D increases if $\Delta c_f > \Delta c_d$. A sufficient condition for a profitable RRC strategy is that it increases the marginal cost of the fringe relatively more than it does for the dominant firm. Indeed, only an increase of the marginal cost of F leads to an increase of the equilibrium price. Consequently, any producer for

which an increase of the market price is beyond the increase of its average cost will benefit from a RRC strategy. Moreover, two other basic results (Scheffman 1992; Church and Ware 2000) can also be mentioned. First, an attempt to increase the dominant firm's costs does not influence the equilibrium price, although it can influence the profit of D. In other words, there is no strategic effect because the situation of F is not affected. Second, the demand has to be sufficiently inelastic in order to guarantee the profitability of the RRC strategy. Indeed, if the demand is highly elastic, the price will not increase despite a marginal cost increase of F. The fringe producers will rather exit the market.

Based on the seminal works of Director and Levi (1956), Nelson (1957), Williamson (1968), and Salop and Scheffman (1983), scholars examined the relevance of RRC strategies in various domains such as free trade agreements (Depken and Ford 1999), advertising, lobbying for product and/or environmental-related standards (Hilke and Nelson 1984; Grolleau et al. 2007), agro-food systems (Barjolle and Jeanneaux 2012), pollution regulation (Sartzetakis 1997; Lyon 2003), and stock exchange (Harris et al. 2014). In general, these studies provide, explicitly or not, evidence to the relevance of using the RRC theory in analyzing specific organizations of the considered markets. Moreover, Normann (2011) recently provided experimental evidence in favor of the

Raising Rivals' Costs,
Fig. 1 Raising the costs of
a competitive fringe
(Inspired from Church and
Ware 2000, 629)



hypothesis that vertically integrated firms have an incentive to foreclose the input market to raise its downstream rivals' costs. Nevertheless, several scholars developed arguments against the RRC theory, in particular regarding its "real" potential to analyze antitrust cases (e.g., Lopatka and Godek 1992; Coate and Kleit 1994). (Given the large literature, we only provide a general overview of the arguments against the RRC theory, without purporting to be exhaustive.) For instance, the RRC theory is often analyzed through its vertical aspects and consequently its contribution to the existing competition analyses can be considered rather weak (Brennan 1986). Indeed, a RRC strategy requires that the predator is assumed to look for a monopoly on a relevant input market, which makes it a particular case of preceding competition analyses. Similarly, Church and Ware (2000) argue that some RRC cases can be also analyzed through the lenses of other theories. Moreover, the required conditions for RRC strategies are so constraining that it turns unlikely to have a significant anti-competitive effects (Coate and Kleit 1994). In addition, the costs of excluding rivals could be also higher than the derived benefits. Furthermore, Boudreaux (1990) points out the fact that the RRC theory does not take into account competitors' counter-strategies. Interestingly, while S. Salop and D. Scheffman developed the RRC theory, they admit that their works have some limits mainly encompassing the following concerns:

1. The theoretical ambiguity of the RRC effects, especially because a RRC situation does not necessarily correspond to an anti-competitive behavior
2. The lack of a framework allowing to distinguish intentional RRC strategies and other types of competition that also alter rivals' situation
3. The focus on particular market structures that may not reflect real-world settings
4. The lack of an analysis regarding the whole effects in terms of well-being (Salop and Scheffman 1987; Scheffman and Higgins 2003)

Conclusion

We presented a general description of the RRC theory. We pointed out the necessary conditions for a profitable RRC strategy and the effects of its implementation, although such effects are often ambiguous. We believe that RRC theory offers promising insights in domains that were not initially considered or extensively studied such as environmental regulation or standards. RRC strategies are sometimes difficult to detect and can be justified by other issues such as environmental conservation or public health. Interestingly, these issues can also lead to "strange" coalitions, such as the one between Baptists and bootleggers described by Yandle (1983). Hence, given the multidimensional nature of RRC, the net welfare effect can remain ambiguous and makes it difficult to craft adequate remedies. Such a result can be discouraging for antitrust authorities.

References

- Barjolle D, Jeanneaux P (2012) Raising rivals' costs strategy and localised agro-food systems in Europe. *Int J Food Sys Dyn* 3(1):11–21
- Boudreaux D (1990) Turning back the antitrust clock: nonprice predation in theory and practice. *Regulation* 45–52
- Brennan T (1986) Understanding raising rivals' costs. *Antitrust Bull* 33:95–113
- Carlton D, Perloff J (1998) *Modern industrial organization*. Addison Wesley Longman, New York
- Church J, Ware R (2000) *Industrial organization: a strategic approach*. McGraw-Hill, Boston
- Coate M, Kleit A (1994) Exclusion, collusion, or confusion? The underpinnings of raising rivals' costs. *Res Law Econ* 16:73–93
- Depken G, Ford J (1999) NAFTA as a means of raising rivals' costs. *Rev Ind Organ* 15(2):103–113
- Director A, Levi E (1956) Law and the future: trade regulation. *Northwest Univ Law Rev* 51:281–296
- Granitz E, Klein B (1996) Monopolization by "raising rivals' costs": the standard oil case. *J Law Econ* 39:1–47
- Grolleau G, Ibanez L, Mzoughi N (2007) Industrialists hand in hand with environmentalists: how eco-labeling schemes can help firms to raise rivals' costs. *Eur J Law Econ* 24(3):215–236
- Harris F, Hyde A, Wood R (2014) The Persistence of dominant-firm market share: raising rivals' cost on the New York Stock Exchange. *South Econ J* 81(1): 91–112

- Hilke J, Nelson P (1984) Noisy advertising and the predation rule in antitrust analysis. *Am Econ Rev* 74(2):367–371
- Lopatka J, Godek P (1992) Another look at ALCOA: raising rivals' costs does not improve the view. *J Law Econ* 35:311–329
- Lyon T (2003) "Green" firms bearing gifts. *Regulation* 4:36–40
- Nelson R (1957) Increased rents from increased costs: a paradox of value theory. *J Polit Econ* 65:287–294
- Normann H (2011) Vertical mergers, foreclosure and raising rivals' costs – Experimental evidence. *J Ind Econ* 59(3):506–527
- Salop S, Scheffman D (1983) Raising rivals' costs. *Am Econ Rev* 73:267–271
- Salop S, Scheffman D (1987) Cost-raising strategies. *J Ind Econ* 1:19–34
- Sartzetakis E (1997) Raising rivals' costs strategies via emission permits markets. *Res Law Econ* 12:751–765
- Scheffman D (1992) The application of raising rivals' costs theory to antitrust. *Antitrust Bull* 37:187–206
- Scheffman D, Higgins R (2003) 20 years of raising rivals' costs: history, assessment, and future. *George Mason Law Rev* 12(2):371–387
- Williamson O (1968) Wage rates as a barrier to entry: the Pennington case. *Q J Econ* 82:85–117
- Yandle B (1983) Bootleggers and Baptists: the education of a regulatory economist. *Regulation* 7(3):12–16

severity of punishment, the kidnapper's willingness to kill the hostage, and the value of the hostage life from the point of view of the family. Limiting the family's ability to pay reduces the frequency of the offense but opens the possibility of unintended consequences in terms of fatalities and duration of abduction.

Definition

Kidnapping in its widest sense occurs when a person carries away another person by force or fraud with the intent to exploit the abduction for a variety of purposes. Ransom kidnapping refers to a situation in which the overriding purpose for the act is a payment (usually a sum of money) for the release of the hostage and the enrichment of the perpetrators. It is a serious crime causing not only economic losses but also pain and suffering, often in the form of post-traumatic stress disorder (PTSD) and major depression (MDD), to the victims and their families. Occasionally it ends in death.

Ransom Kidnapping

Marco Vannini¹, Claudio Detotto² and Bryan McCannon³

¹Department of Economics and CRENoS, University of Sassari, Sassari, Italy

²DiSea – CRENoS, University of Sassari, Sassari, Sardinia, Italy

³Saint Bonaventure University, St Bonaventure, NY, USA

Abstract

The practice of kidnapping for ransom, a predatory crime carried out mostly by criminal organizations, is a salient phenomenon in many regions of the world. It causes serious harm not only to victims and their families but also to private and social capital. As a paradigmatic rational crime involving negotiations, the incentives to commit the crime and the way it ends change with the probability and

Introduction

Kidnapping for ransom, also referred to as economic kidnapping or profit kidnapping, is a predatory crime carried out mostly by criminal organizations, rather than single offenders, usually after careful planning of the various stages of the (illegal) production process. The latter, in perfect business style, begins with a market analysis designed to identify the most profitable opportunities and to evaluate the benefits and costs of the different options in light of the strengths and weaknesses of the organization. Specialized gang members/units are required for carrying out the activity: e.g., for targeting victims (so-called spotters), holding the hostage (a tricky task often performed by someone difficult to track, like an absconder), communicating with the parties involved (media, families, police, negotiators: both formally and through the grapevine), providing food to hostages and guardians, and collecting

and cleaning the ransom (be it hard cash or any other resource with financial and/or social value). Depending on the nature of the criminal partnership and its specific market segment (see below), the production chain may include reinvestment of the proceeds in higher-value illegal markets, like the international drug industry, and/or further actions designed to secure a dominant position for the parent organization both in the underworld and the upperworld (as in the case of mafia-type or terror-affiliated groups).

Despite the relative complexity of the structures and tasks capable of sustaining a successful ransom kidnapping business, this type of crime is seldom classified as *organized crime*. The distinction between (even highly organized) ordinary illegal firms and organized crime, which may be relevant for the criminal justice system responses to the problem (see Garoupa 2007), is still controversial among both scholars and practitioners (see the digital collection by Klaus von Lampe www.organized-crime.de/organizedcrimedefinitions.htm; Varese 2010). Clearly, neither the earlier descriptive definition by Donald Cressey (1969) of organized crime as an entity “rationally designed to maximize profits by performing illegal services and providing legally forbidden products demanded by the members of the broader society” nor the insightful characterization by Thomas Schelling (1971) of organized crime as the underworld counterpart of monopoly and not just organized business mirrors the multifaceted phenomenon of ransom kidnapping across time and places. However, reversing the perspective and looking with Schelling at the factors that make an illegal activity a prime target for organized crime (basically victimized criminals should have no ready access to the law, they must find it difficult to hide from the “protector” and to carry away their business in an attempt to escape, and their earnings must be easily monitored and flow smoothly and regularly), it may be argued that an appropriative activity like ransom kidnapping does not lend itself to be monopolized. Moreover, being based on a relatively simple technology with generally modest overhead costs, the requirements for a large-scale firm are hardly met. Alternatively, adopting Williamson’s

(1985) view that organizational variety arises primarily in the service of transaction cost economizing, the activities of organized criminal firms “will be guided primarily by the relative costs of completing illegal transactions within the market versus within a downstream firm” (Dick 1995).

While the debate on the nature of organized crime (Peltzman and Fiorentini 1995) goes on, the evidence gathered so far on ransom kidnapping and organized crime worldwide shows very different patterns within and between countries. In Italy, for instance, ransom kidnapping has been a key source of liquidity, to be invested in real estate development and in the drug business, for both the ‘Ndrangheta and the Sicilian Mafia since the early 1960s. By contrast, the rural gangs of Sardinia that dominated the kidnapping business in the country for two decades (late 1950s through the 1970s) never tied to or behaved as classic organized crime. Likewise, within the top seven Mexican trafficking organizations, only the Zetas and the Beltran Leyva indulged in kidnapping (Astorga 2012); with the crackdown on drugs, however, many former criminal members of all such cartels found it profitable to switch to organized kidnapping (a similar pattern has recently emerged in Kenya and Somalia as a result of the amplified efforts to crack down on piracy). In Nigeria, where the kidnapping business emerged in the 1990s, this type of crime is associated with wealth-oriented secret societies, untouchable secret sorcery society, and “419” fraud syndicates (Ebbe 2012). Finally, as reported in Levitt and Rubio (2000), a strong causal link between kidnapping and the scope of guerrilla activities has been found in Colombia over the 1990s.

In recent years, against the backdrop of rapid globalization and increased pervasiveness of ICT (information and communications technology), kidnap and ransom has become a billion-dollar industry in which major changes have taken place. Next to classic ransom kidnapping, new categories have flourished, with peculiar traits that deserve attention for both credible research and improved enforcement.

Insider kidnapping: criminals buy inside help from employees of target companies/resorts

to access key information for the abduction to take place.

Express kidnapping: abductions have minimum duration (typically less than a day) and are calibrated in order to extract maximum cash withdrawal from ATM, for example, using the victim's payment instruments or to get a quick ransom from his/her family or company.

Tiger kidnapping: one or more individuals are abducted to coerce another person to commit a crime – which can be anything from robbery, extraction of a ransom, to murder – for the benefit of the kidnappers.

Terrorist kidnapping: motivated by financial and political reasons, it operates in politically unstable countries exploiting both domestic and international networks to carry out the abduction and to launder the ransom; it involves mainly foreign victims – often from profiled nationalities – and although the demands may start out politically motivated, they may transition to financial benefits as the negotiations progress.

Piracy for ransom: incidents take place in maritime zones in which the right of State sovereignty do not apply and comprise both physical capture of cargo from a vessel and demanding a ransom in exchange for the vessel, crew, and cargo; it represents a multi-jurisdictional challenge whose escalating figures, concerning both frequency of attacks and average ransom payments, are posing serious threats to the financial, insurance, and shipping system.

Virtual kidnapping: the supposed victim is lured or forced into a place in which it is impossible or difficult to communicate or be contacted, at the same time a family member receives calls by someone who claim to have kidnapped their loved one and demand a ransom to freed the hostage.

The global kidnapping epidemic has brought close cooperation between jurisdictions, as well as effective implementation by the largest number of countries of international standards and preventing measures against money laundering, terrorist financing, and corruption, to the forefront. In the transnational setting, however, the differing

public good aspects of defensive and proactive measures magnify the risk of collective action failures when implementing such policies, with potential oversupply of the former and undersupply of the latter (Enders and Sandler 2005).

In order to gain a deeper understanding of the kidnapping for ransom phenomenon and its control, the rest of this entry will review three related aspects: the incidence of modern-day ransom kidnapping around the world, the theoretical framework providing insights into the strategic problems faced by the kidnapper and by the hostage's family, and the lessons from applied studies of kidnapping on two selected policy issues: assets freezing and marginal deterrence.

Modern-Day Ransom Kidnapping

Whereas the practice of kidnapping is as old as recorded history and is mentioned in ancient texts from all over the world, kidnap for ransom as the crime is understood today is a relatively recent phenomenon. According to Wright (2009), it was in the late nineteenth and early twentieth centuries that legislation defining the crime and prescribing harsh punishments for the perpetrators was formally adopted for the first time in many places. In the USA, for instance, seven states passed their first anti-kidnapping laws, and 18 others stiffened penalties, in the first two decades of the twenty-first century. The Federal Kidnapping Act was passed in 1932 after the son of the America's hero Charles Lindbergh was kidnapped and killed. By that time many European countries with different legal traditions had already implemented their laws, although new provisions and amendments have continued until recently in order to catch up with the evolving nature of the crime and the surge of hostage taking by terrorists.

Despite efforts by UNODC (United Nations Office on Drugs and Crime) to collect comparative data on recorded crime in member states, the available information is still heterogeneous in many dimensions (definition of crimes, periodicity, details, and so on) and does not provide a reliable basis for cross-country analysis. Both scholars and practitioners have to rely on a mix

of sources, like extracted information from newspaper accounts, historical reconstructions, and official statistics when ransom kidnapping as a separate offense category among the index crimes is available. It is well known that since its modern-day inception, ransom kidnapping flourished in the first half of the twentieth century in the USA and in the second half in Italy. Studying the *New York Times* files on kidnapping incidents, Alix (1978) reports 1,703 cases of kidnapping occurring almost entirely in the USA between 1874 and 1974 (about 1.7 per year), including 236 classic kidnappings for ransom. The crime reached its peak in the 1930s, disappeared during the next 40 years, and reappeared dramatically in the 1970s when it became a politically motivated symbolic act of power. Yet, compared to criminal homicides (with about 7,000 cases in each of the years 1924–1974), the contribution of ransom kidnapping to total crime was modest.

By contrast, kidnapping for ransom in Italy in the period 1960–2000, with 592 cases and an average of 14.4 abductions per year, had a dramatic impact and placed the country at the top of the worldwide kidnapping hot spots throughout the 1970s and most of the 1980s (Wright 2009). Failed attempts are excluded from this count, which is based primarily on a unique archive by the former law enforcement official Luigi Casalunga (2013). Exploring the dataset, the frequency of kidnappings spiked in the late 1970s (73 incidents in a single year), continued at remarkable levels through the mid-1980s, and then slowly declined in the following decades.

In this scenario, the island of Sardinia played a special role. As documented in Marongiu and Clarke (2004), the island – despite being an organized crime-free region – had a long-standing tradition on kidnapping dating back to more than 500 years. In the time span 1960–2013, around 27% of Italy's ransom kidnappings took place in Sardinia, resulting in 162 incidents against 118 in Lombardy and 96 in Calabria: 10.6 cases per 100,000 residents, i.e., ten times the national rate. In those years, Sardinian bandits were global players and spread the crime all over Italy. They were responsible for 39 out of 40 incidents in the 1960s, 62 over 258 in the 1970s, and 39 over

249 in the 1980s. To the kidnap epidemic contributed also gangs associated with organized crime from Calabria ('Ndrangheta) and Sicily (Mafia). While the former, having found a congenial operating habitat out of the Aspromonte, ran the business until recent times, the latter withdrew pretty soon for fear of additional police controls and negative reputational effects brought about by such heinous crime.

Nowadays, in many areas of the world, kidnapping is a flourishing criminal industry, with wide direct and indirect impacts on victims and society at large. According to the Control Risks Group, an international risk consultancy, at the turn of the century economic kidnappers globally would take home well over \$500 million each year. Countries involved include Colombia, Mexico, Brazil, Philippines, and Venezuela. In Colombia, where the fraction of kidnappings of firm managers and firm owners were just under 10%, it is calculated that this type of kidnapping has a statistically significant negative relationship with corporate investment (see Pshiva and Suarez 2010). In the last decade, however, the kidnapping activity has declined both in Colombia, where the annual reported kidnapping incidents went from about 3,500 cases in the year 2000 to “only” 282 10 years later (Fink and Pingle 2014), and in the rest of Latin America. Meanwhile, kidnap numbers have soared in Nigeria, India, Lebanon, Iraq, and Afghanistan, while Mexico confirms its primacy with the highest number of kidnaps for ransom recorded worldwide (roughly 2,000 cases in 2010). According to *Alto al Secuestro*, a foundation that assists victims in Mexico, between December 2012 and February 2014, there were 4,051 kidnapping victims nationwide, of which 2,922 had been freed, while 1,129 were still being held. As the very affluent targets are adopting effective defensive measures, kidnappers are increasingly switching to middle-class workers, small business owners, students, and mid-level professionals, which now make up about 70% of the total victims.

As a matter of fact, Latin America has lost its leadership in favor of African and Asian countries, which represent the new frontier for this typology of criminal activity. In 2004 about 55%

of the world's recorded kidnaps for ransom were in Latin America, while in 2012 the region accounted for only a quarter of the incidents, and Asia and Africa made up slightly less than 75% (with about 50% for Asia excluding the former USSR and with 22% for Africa).

Deeper international economic integration, greater political instability (e.g., in the former Soviet Republics and in the Middle East area), and growing interest by international tourists for adventure holidays have created new opportunities and brought about new potential victims for profitable kidnapping. In fact, it is no coincidence that kidnappings of foreign nationals globally have increased by 275% over the 2000s (Kassim and Mohamed 2008). Furthermore, since 2005, the Horn of Africa, particularly the Gulf of Aden (Somalia) and Arabic Sea, has risen to prominence as a new hotspot for maritime piracy, particularly kidnappings for ransom.

At present, ransom kidnapping is a thriving industry in developing regions, while it has virtually disappeared in North America and Europe, which together represent only about 1–3% of global activities. This polarization can be even more extreme in light of the under-reporting and under-recording biases of the official crime data, which are presumably more serious in the former countries.

Interactions Between the Kidnapper and the Victim's Family

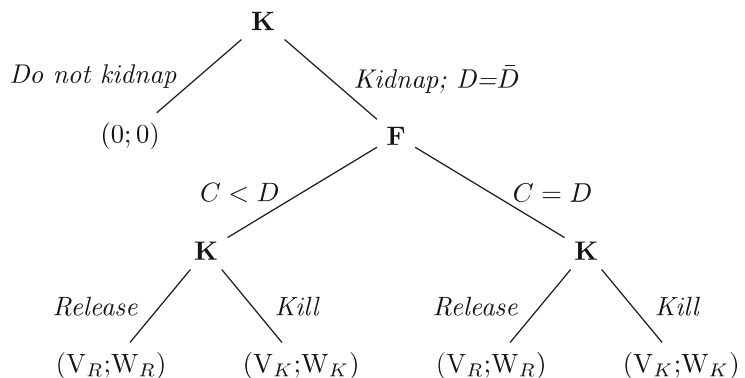
In 1976, scholars and practitioners from all horizons of game theory, physics, psychiatry,

and behavioral sciences and from private and public institutions gathered in Santa Margherita, Italy, for a seminar on the control and prevention of hostage taking organized by the Institute of Criminal Justice and Criminology of the University of Montreal in cooperation with the International Centre for Comparative Criminology of the University of Maryland (Crelisten and Laberge-Altmejd 1976). On that occasion, Reinhard Selten (1976) presented a two-person game model to gain insight into the strategic problem faced by the kidnapper (K) and the hostage's family (F). Nowadays, versions of the model feature in microeconomics textbooks to illustrate the structure of dynamic games and the notion of subgame perfection (Harrington 2014). Yet, it still represents the basic theoretical reference for the analysis of hostage-taking events. The extensive form of this game is represented in Fig. 1.

At the top of the decision tree, agent *K* makes the first move deciding whether to kidnap the victim or not. If he chooses the latter, the game ends with zero payoffs. If, instead, he chooses to kidnap the target, he announces a ransom *D*. Then, the victim's family *F* either accepts the demand from *K* or makes an alternative offer *C*. In both cases, in turn, the kidnapper will decide to release or kill the hostage.

If the family accepts the ransom requested ($C = D$), the criminal will rationally decide to release or kill the victim according to the expected payoffs V_R and V_K . If the family offers an unsatisfactory amount ($C < D$), then the kidnapper could feel offended and frustrated at the low offer and might react killing the hostage.

Ransom Kidnapping,
Fig. 1 The extensive form
of the game



It is assumed that the probability associated to this extreme reaction is equal to α , which is a function of the distance between the amount offered (C) and the amount requested (D):

$$\alpha = a(1 - C/D) \quad (1)$$

where a is a parameter ranging between 0 and 1, which measures the kidnapper's propensity to sanction noncompliant offers. As one can see, the higher the $C - D$ gap, the higher the value of α ; when $C = D$, the likelihood of this sanction equals zero.

The subgame-perfect equilibria of the game shown in Fig. 1 can be investigated by backward induction. Starting at the game's final decision nodes and analyzing player K 's choice, we see that he prefers to release the hostage if (and only if) the following relation holds:

$$V_R > V_K \quad (2)$$

where

$$V_R = (1 - q)C - qx \text{ and } V_K = -y - qz. \quad (3)$$

Briefly, V_R and V_K represent the kidnapper's expected gain in the two regimes (release or kill), where (q) is the probability of being arrested

by the police, (C) is the ransom paid, ($-x$) the disutility of being caught, ($-y$) the disutility associated with the killing of the hostage, and ($-z$) the disutility of being caught in case the victim is executed.

Since $C > 0$, $y > 0$, and $x \leq z$, the relation Eq. 2 always holds true, the kidnapper will never rationally decide to execute his threat. Moving to the upper node of the decision tree, player F has to choose the optimal offer C that maximizes his utility U taking into account the likelihood of a violent reaction by K :

$$U = (1 - \alpha)W_R + \alpha W_K \quad (4)$$

where

$$W_R = -(1 - q)C \text{ and } W_K = -w. \quad (5)$$

The parameter w captures the value of the hostage's life from the point of view of the family. Equation 4 shows that, on the one hand, the family F benefits from an offer reduction, but on the other hand, this strategy increases the probability of a sanction threat from K . So, the hostage's family will propose an offer C^* that maximizes its utility function. Equation 4 attains its maximum at C^* which can be interpreted as the best response given a , q , w , and D . Therefore, the optimal offer C^* is given by Eq. 6:

$$C^* = \begin{cases} D & \text{for } 0 < D \leq \frac{aw}{(1+a)(1-q)} \\ \frac{w}{2(1-q)} - \frac{(1-a)D}{2a} & \text{for } \frac{aw}{(1+a)(1-q)} \leq D \leq \frac{aw}{(1-a)(1-q)} \\ 0 & \text{for } D \geq \frac{aw}{(1-a)(1-q)} \end{cases} \quad (6)$$

Notably, C^* is increasing in D in the interval $0 < D \leq \frac{aw}{(1+a)(1-q)}$ and then decreasing up to

$$\frac{aw}{(1-a)(1-q)}.$$

As a matter of fact, in the upper subgame agent K sets D in order to extract the largest offer from

player F . It follows that the optimal value D^* is the following:

$$D^* = C^* = \frac{aw}{(1+a)(1-q)} \quad (7)$$

When Eq. 7 holds true, player F will accept the kidnapper's offer, D^* , and the latter will release the hostage.

Finally, looking at the first node, the kidnapper has an incentive to engage in the act of kidnapping when $V_R > 0$, which means

$$\frac{aw}{(1+a)} > qx \quad (8)$$

According to Eq. 8, kidnappings can be discouraged by increasing x or q and by decreasing a or w . Both x and q can be influenced by public policies, for instance, hardening punishment and spending more resources in detecting criminals. However, the effect of severity on x can be very tenuous in practice (Paternoster 2010), whereas the effect of increasing q can be less effective than expected due to its impact on the optimal offer from the family (which in case of capture recovers any ransom paid). This latter possibility can be mitigated by restricting the family's ability to pay. As for w , which cannot be influenced directly by public measures, it may be noted that it may affect the choice of kidnappers as to which member of the family to kidnap. The individual parameter a , instead, can partially be influenced by recommendations from the police and/or the insurance company on the way to handle the negotiations with the kidnappers.

By contemplating the possibility of an extreme reaction by the kidnapper, Selten's model bypasses situations in which one or both agents have a dominant strategy and succeeds in providing a compelling negotiation framework.

Going back to Fig. 1, if one removes the possibility of such reaction, $V_R > V_K$ and the kidnapper will always release the hostage. Hence, the optimal strategy for player F , given the dominant strategy of K , is not to pay the ransom D . The unique equilibrium is represented by the zero-payoff outcome. The choice of rational kidnappers, with no reputation to defend, does not depend on the payment but rather on the balance between the increased penalty in case of murder of the victim and the increased probability of being caught if the victim is left alive. Why then are people often willing to pay ransom?

To make sense of this puzzling result, referred to as "the mystery of kidnapping" (e.g., Gintis 2009), a modified model is required. For instance,

with Harsanyi (1967), one can hypothesize the existence of two types of kidnappers: vindictive and self-regarding. So, for some values of p_v , the share of vindictive individuals among all kidnappers, the best strategy of the family is to pay the ransom. Alternatively, the family could incur additional psychic costs if a ransom is not paid (Gintis 2009). Then, a ransom would materialize as long as such additional costs exceed the kidnappers' request (but again the victim will be released only if the cost of freeing is less than the cost of killing the victim).

Hindering Ransom Payments

In order to fight ransom kidnapping by reducing its anticipated net benefits, both defensive and proactive measures can be adopted. Relative to the second type of measures, countries like Venezuela, Colombia, and Italy have banned both ransom payments and kidnap insurance, and the freeze of assets of the victim's family is often automatic. Despite the growing international consensus on similar measures in the case of terrorist-related kidnapping, these policies are highly controversial. As noted by Block and Tinsley (2008), "one may as well enact legislation forbidding a mugger's victim from responding "life" to the threat of "your money or your life." To some extent such measures end up punishing the victim rather than the criminal, and, in any event, most people would see it as an unacceptable restriction of personal liberty. The standard justification is that ransom payment and insurance impart a negative externality on future victims, and by eliminating them the incentive to perpetrate kidnapping would disappear. Recent contributions uncover more subtle issues.

Fink and Pingle (2014), investigating within Selten's framework the impact of kidnap insurance, find that – under the assumption that kidnappers have a positive net willingness to kill – a market for kidnap insurance can benefit risk-averse agents, as long as it does not increase the risk of kidnapping too much. Individuals should fully insure in the latter case and partially insure otherwise. Since an insurance market enables the

families to rescue their loved ones from kidnap-
pers more prone to murder but it may also increase
the likelihood of kidnapping, the total effect on
the number of kidnapping deaths is indeterminate.
However, if the likelihood of kidnapping is only
marginally affected, then a law banning kidnap
insurance implies more fatalities.

Preventing ransom payments may have an
impact not only on the risk and fatalities of kid-
napping but also on the duration of the kidnapping
experience. An extended duration escalates the
harm of the crime. Detotto et al. (2014) study
this problem in a simple setup in which as time
passes the family becomes more willing to pay the
ransom and the kidnapper balances the benefits
and costs to extending the duration. The optimal
stopping period, where K is no longer willing to
take the chance of being caught to increase the
amount collected in ransom, implies that duration
grows inversely with the initial ransom offer, the
probability of apprehension during the kidnap-
ping, the probability the hostage dies or flees, the
maintenance costs, and directly with the incre-
mental increase in the ransom. Exploiting a
micro-dataset on kidnapping incidents in Sardinia
between 1960 and 2010 to estimate a semi-
parametric survival function, the authors find
that the anticipated apprehension probability
does not have a significant effect on duration,
whereas the asset-seizure policy has mixed

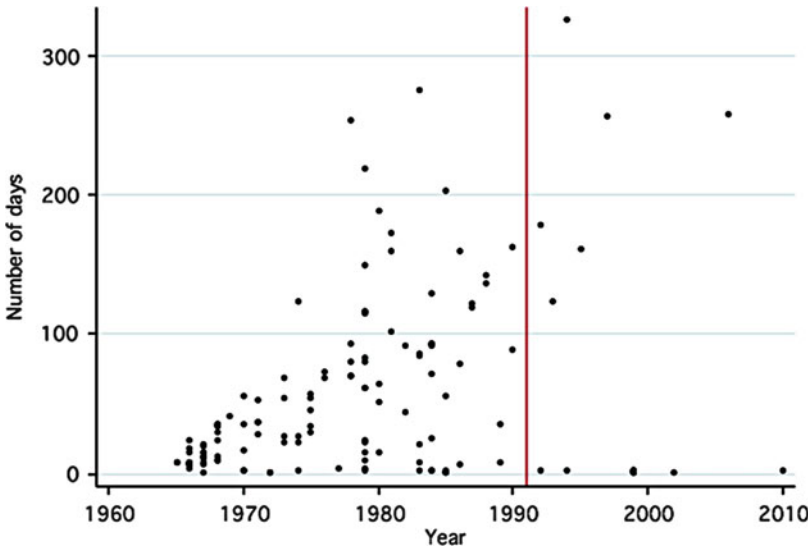
effects: it leads to significantly shorter abductions
(i.e., an express kidnapping), but if a kidnapping
is not express some evidence is found that, in fact,
duration might expand as the policy creates fric-
tions in the collection of the ransom for the family
(see Fig. 2, in which the vertical line corresponds
to the year of implementation of the asset-seizure
policy). The polarization is dramatic: to the right
of the red line, kidnappings are either very short
(less than a day on average) or very long (about
200 days on average).

Taken together with the observation that the
frequency of kidnappings dropped significantly
after the adoption of the asset-seizure policy,
these results support interventions directed to
reduce the anticipated benefit to kidnapping rather
than focus on the costs.

Increasing Penalties for Kidnappers

Deterrence is a central concept in the economic
model of crime (Becker 1968) because potential
offenders rationally (though not necessarily con-
sciously) consider and balance the costs and ben-
efits of committing a crime. In this framework,
criminals opt between legal and illegal activities
according to their utility and budget constraint. In
the real world, however, criminals not only must
choose the optimal allocation of effort between

Ransom Kidnapping,
Fig. 2 Kidnapping
duration over time



legal or illegal activities, but they may allocate their effort between several crime offenses. The idea of marginal deterrence recognizes that the setting of sanctions for one particular offense not only affects deterrence of that crime but also affects the incentives to engage in other activities. In his pioneering analysis of the subject, Shavell (1992) explicitly refers to the classic example of kidnapping/murder as an application of marginal deterrence.

Recently, Detotto et al. (2015) propose a reasonable application of the theory of marginal deterrence related to kidnapping and its complement, murder. While the death exposes the criminals to punishments, if the sanction for kidnapping is great, then the marginal sanction for homicide is reduced.

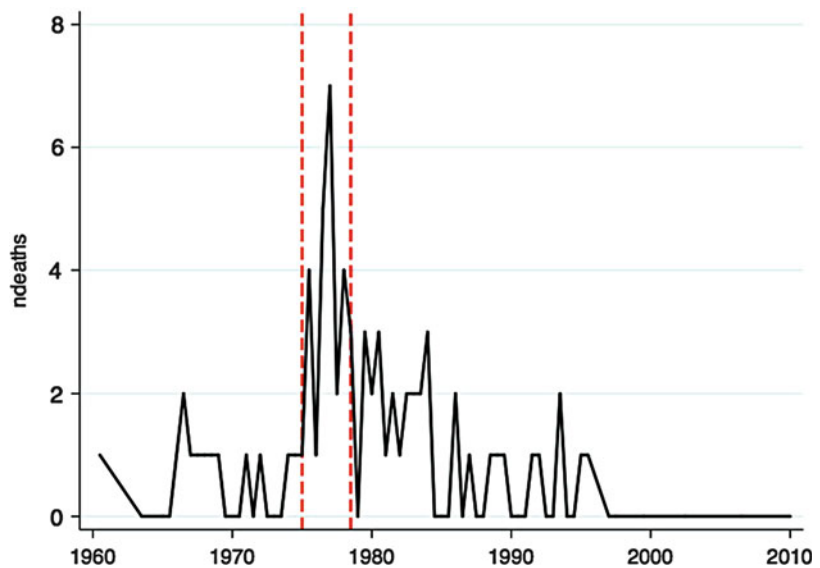
They use a unique data set of kidnappings in Italy between 1960 and 2013 to test the theory of marginal deterrence. Kidnapping was a major concern for Italy in the 1960s, and, as a consequence, in 1974 a new set of sentencing policies were set with greatly enhanced punishments for the crime. Precisely, the range penalty passed from imprisonment for 8–15 years to 10–20 years, if the ransom was not paid, and from 12–18 years to 12–25 years, if the ransom was actually paid. Such reforms, though, did not change the sanction for murder which was punishable with 21 years to life

sentences. Thus, marginal deterrence of death reduced. Later, in 1978 in response to increased homicides, the Italian government escalated sanctions for deaths resulting from kidnappings, addressing the marginal deterrence problem. The new change law enacted in May 1978 specifically increased the sanction for deaths associated with kidnapping.

Figure 3 provides illustration of kidnapping-related deaths over time. The vertical dashed lines denote the policy changes. As the theory of marginal deterrence predicts, the adoption of the enhanced sanctions for kidnapping in 1974 resulted in an increase in the prevalence of death, while the latter became minimal after the 1978 policy.

To test the hypothesis that the changes in the sanctions affected the incentive to murder the victim, a binary probit model is estimated with the dummy variable death as the dependent variable and the characteristics of the victim, time of year, and location as the controls. The results in Table 1 provide confirmation of the theory. The adoption of the enhanced sanctions for kidnapping in 1974 resulted in an increase in the prevalence of death. Precisely, the marginal effect is estimated to increase the likelihood of death occurrence by 4%. Then, according to the predictions of the theory, the escalation of sanctions

Ransom Kidnapping,
Fig. 3 Number of
kidnapping-related murders



Ransom Kidnapping, Table 1 Probit analysis (dep. var. = death)

	(I)	(II)	(III)
D1974	0.04* (0.03)	0.05* (0.03)	0.04** (0.02)
D1978	−0.05** (0.03)	−0.05** (0.02)	−0.03* (0.02)
Events	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)
Female	−0.04** (0.02)	−0.04** (0.02)	−0.04** (0.0)
Age	0.003*** (0.001)	0.003*** (0.001)	0.002*** (0.001)
Local	−0.07* (0.05)	−0.08** (0.05)	−0.01 (0.06)
Italian	−0.06** (0.02)	−0.06** (0.02)	−0.03 (0.05)
Paid	−0.12*** (0.03)	−0.12*** (0.03)	−0.10*** (0.02)
Pseudo-R2	0.159	0.157	0.173
Wald	69.49***	69.89***	924.34***
Log-likelihood	−185.84	−186.45	−174.31
N	593	593	574

***1 %, **5 %, *10 %; marginal effects and robust standard errors are reported in parentheses. A constant term is included in each specification. Controls: victim occupation, season, and region of capture (I): 1974 and 1978 are lagged 3 months, (II): 1974 and 1978 are leaded 3 months, (III): dropped all obs. in the range [−3, +3] months of both 14 October 1974 and 18 May 1978

for kidnap-murders of 1978 reduced the prevalence of death, which, in numbers, is estimated to have declined by 5%.

Conclusion

While the categorization of kidnapping for ransom as organized crime is uncertain, in many regions of the world modern-day hostage taking for ransom is carried out mostly by highly organized criminal firms on a historically unprecedented scale. Recent trends show a blurring of the boundaries between ransom kidnapping, piracy, and terrorism together with an increasing international dimension that parallels contemporary legal globalization. The contributions reviewed in this entry stress the importance of both classical instruments and more controversial measures, like banning ransom insurance and/or ransom payments, to counteract the crime. However, given the likelihood of unintended negative effects (on fatalities, duration of the kidnapping

experience, and burden falling on enterprises) in some environments, policies aimed at reducing kidnapper’s anticipated benefits through these channels must carefully balance the benefits for society at large against the higher costs imposed on specific groups.

Cross-References

- [Crime and Punishment \(Becker 1968\)](#)
- [Crime: Organized Crime and the Law](#)
- [Piracy, Modern Maritime](#)
- [Terrorism](#)

References

Alix EK (1978) Ransom kidnapping in America, 1874–1974: the creation of a capital crime. Southern Illinois University Press, Carbondale

Astorga L (2012) México: organized crime politics and insecurity. In: Siegel D, Van de Bunt H (eds) Traditional organized crime in the modern world. Springer, New York, pp 149–166

- Becker GS (1968) Crime and punishment: An economic approach. *Journal of Political Economy*: 76:169–217
- Block W, Tinsley P (2008) Should the law prohibit paying Ransom to kidnappers? *Am Rev Polit Econ* 6:40–45
- Casalunga L (2013) Sequestri di persona in Italia. *Il Maestrale*, Nuoro
- Crelisten R, Laberge-Altmejd D (1976) Report on Management Training Seminar Hostage-Taking Problems of Prevention and Control. Montreal, National Institute of Justice
- Cressey D (1969) Theft of the nation: the structure and operations of organized crime in America. Harper, New York
- Detotto C, McCannon B, Vannini M (2014) Understanding Ransom kidnappings and their duration. *BE J Econ Anal Policy* 14:849–872
- Detotto C, McCannon B, Vannini M (2015) Evidence of marginal deterrence: kidnapping and murder in Italy. *Int Rev Law Econ* 41:63–67
- Di Tella R, Edwards S, Schargrodsky E (eds) (2010) The economics of crime lessons for and from Latin America. The University of Chicago Press, Chicago
- Dick AR (1995) When does organized crime pay? A transaction cost analysis. *Int Rev Law Econ* 15:25–45
- Ebbe ONJ (2012) Organized crime in Nigeria. In: Siegel D, Van de Bunt H (eds) *Traditional organized crime in the modern world*. Springer, New York, pp 169–188
- Enders W, Sandler T (2005) The political economy of terrorism. Cambridge University Press, New York
- Fink A, Pingle M (2014) Kidnap insurance and its impact on kidnapping outcomes. *Public Choice* 160:481–499
- Garoupa N (2007) Optimal law enforcement and criminal organization. *J Econ Behav Organ* 63:461–474
- Gintis H (2009) *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction* (Second Edition). Princeton University Press, Princeton
- Harrington J (2014) *Games, Strategies, and Decision Making*. Worth Pub, New York
- Harsanyi JC (1967) Games with incomplete information played by “Bayesian” players. The Basic model. *Management Science* 14:159–182
- Kassim M, Mohamed N (2008) Kidnap for ransom in South East Asia. *Asian Criminol* 3:61–73
- Levitt S, Rubio M (2000) Understanding crime in Colombia and what can be done about it. Working papers series. Bogotá, Fedesarrollo
- Marongiu P, Clarke RV (2004) Ransom kidnapping in Sardinia: subculture theory and rational choice. In: Clarke RV, Felson M (eds) *Routine activity and rational choice*. Transaction Publishers, New Brunswick, pp 179–199
- Paternoster R (2010) How much do we really know about deterrence? *J Crim Law Criminol* 100:765–823
- Peltzman S, Fiorentini G (1995) *The economics of organised crime*. Cambridge University Press, Cambridge
- Pshiva R, Suarez GA (2010) Capital crimes: kidnappings and corporate investment in Colombia. In: Di Tella et al. (ed) *The economics of crime: lessons for and from Latin America*. Chicago, National Bureau of Economic Research
- Schelling T (1971) What is the business of organized crime? *J Public Law* 20:71–84
- Selten R (1976) A simple game model of kidnapping. Institute for Mathematical Economics at Bielefeld University, working paper no 45
- Shavell S (1992) A note on marginal deterrence. *Int Rev Law Econ* 12:345–355
- Varese F (2010) What is organized crime? In: Varese F - (ed) *Organized crime (critical concepts in criminology)*. Routledge, London, pp 1–35
- Williamson OE (1985) *The economic institutions of capitalism*. Free Press, New York
- Wright RP (2009) *Kidnap for Ransom: resolving the unthinkable*. Taylor & Francis Group, Boca Raton

Rape

► Sex Offenses

Rationality

Anne-Marie Mohammed¹, Sandra Sookram² and George Saridakis³

¹Department of Economics, Faculty of Social Sciences, The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

²Sir Arthur Lewis Institute of Social and Economic Studies, The University of The West Indies, St. Augustine, Trinidad and Tobago, West Indies

³Small Business Research Centre, Business School, Kingston University, London, UK

Synonyms

[Analytical](#); [Coherence](#); [Common sense](#); [Deduction](#); [Inference](#); [Judgment](#); [Optimality](#); [Reasonableness](#)

Definition

Rationality involves the evaluation of choices to achieve a goal or to find the optimal solution to a

problem. Simon (1972, p. 161) defined rationality as “a style of behavior that is appropriate to the achievement of given goals, within the limits imposed by given conditions and constraints.”

Introduction

The notion of rationality has become a central idea in the various disciplines within the social sciences. This entry discusses the concept of rationality, which has been a core concept in the explanation of human behavior in economics and the other social sciences. In the field of economics, it is expected that individuals behave rationally and that organizations should make rational decisions. A substantial number of economic theories are established under the assumption that when individuals act they do so in a rational manner. Simon (1957) revises this assumption by proposing the idea of bounded rationality. The concept of bounded rationality accounts for the fact that a perfectly rational decision cannot be made because of shortcomings that individuals face with regard to the inadequacy of information, the time constraints, and the limitations to their cognitive processes. This entry gives particular emphasis to subject areas within economics where rationality plays an important part in the formulation of theories and in explaining behavior, particularly the economics of crime.

The structure of the entry is as follows. We start the entry by examining the concept of rationality in economics. The entry proceeds by analyzing the influence of economic rational choice theory on the development of an economic theory of crime and its impact on crime and in the crafting of policy. The last section concludes the entry by offering some key observations and directions for future research.

Rationality and Economics

Karl Popper’s 1967 essay, “The Rationality Principle”, was one of the first studies that linked the concept of rationality to social sciences (Karl Popper 1967 essay in Miller 1985.) According to

Popper’s principle, one should analyze social processes by assuming that “agents always act in a manner appropriate to the situation in which they find themselves” (p. 361). Popper’s view of rationality embraced a notion of situational analysis, and he went on to note that it is able to explain, “. . .the unintended social repercussions of intentional human actions” (Popper 1945/1966, p. 95). This concept by Popper was strongly criticized by many as it did not seem consistent with his principle of falsification. Others, such as considered the rationality principle to be consistent when taken with the view that it is “approximately false” and can be falsified in very rare cases.

Popper went on to contend that situational analysis is an approach employed by economics; this according to other researchers can be classified as neoclassical rationality. Langlois (1997) in his paper noted that the assumptions of the basic neoclassical model can be divided into four categories:

1. *Self-interest*: According to Langlois (1997), self-interested behavior can actually be categorized as purposeful behavior, and this purposefulness as pointed out by Vanberg (1993) is one of the more appealing elements of the neoclassical model. This assumption has been heavily criticized by researchers both in and outside the field of economics. This according to Langlois (1997) is because of a tendency to misidentify self-interest with selfishness.
2. *Omniscience*: This second assumption is another way of saying “perfect competition,” which in neoclassical theory means that agents have perfect information with respect to a particular structure set out for them by the analyst. An example of this is the general equilibrium theory developed by Arrow-Debreu where economic agents are required to know all the utilities and production possibilities of all other agents. Langlois (1997) criticized this assumption based on the type of knowledge that economic agents are expected to have. He noted that the neoclassical model expects agents to have “structural knowledge” (full knowledge of the structure of the economic problem that they face) but not “parametric knowledge”

(full knowledge of all the parameters). He went on to state that in neoclassical theory it is generally the case to relax the assumption of “perfect knowledge” but this falls more often on the side of “parametric knowledge.” This is problematic because more often than not people are ignorant of the very nature of the problem situation they face.

3. *Conscious deliberation*: The third assumption of the basic neoclassical model is that agents are consciously considering their options and then choosing among them. Langlois (1997) noted that as the field of economics has become more mathematical, economic agents are now solving more complicated problems through deliberation. Friedman (1953) presented the only alternative to this assumption noting that economic agents do not actually deliberate but rather behaved “as if” they had. This alternative view has not been readily accepted by most students of economic methodology.
4. *Representative agent*: Alfred Marshall (1961) established the assumption of the “representative agent.” Marshall (1961) defined the representative firm as one that represents the typical properties of the population of firms as a whole and not just the properties of any particular firm. According to Langlois (1997) this definition was used so that some measure of “population” thinking can be accommodated in the theory of comparative statistics.

Rationality, in particular, is applied to the concept of game theory which is an essential aspect of microeconomics. Rationality is a central assumption of game theory irrespective of the different variations of the game that exist. In the context of game theory, a rational player is one who always chooses his/her most preferred outcome given the expectations of his/her opponent. Game theory according to Turocy and Stengel (2001) is defined as “the formal study of conflict and interest”. Antoine Cournot (1838) was the first to discuss this in his study of duopolies. By the 1950s and the 1960s, the analysis was broadened to deal with problems like war and politics, as well as sociology and psychology. Turocy and Stengel (2001) noted that the strength in game theory is that it

provides a structure for analyzing problems of strategic choices. Turocy and Stengel (2001, p. 5) in describing how game theory works noted that “The process of formally modelling a situation as a game requires the decision-maker to enumerate explicitly the players and their strategic options, and to consider their preferences and reactions.”

The “game” in game theory involves modeling a situation that involves several players. Games can be categorized depending on the level of details involved in the process. For instance, there is coalition/cooperative game theory and then there is noncooperative game theory. The cooperative game theory requires a high level of description, and it investigates the power among different coalitions or as Turocy and Stengel (2001, p. 6) state it, “how a successful coalition should divide its proceeds.” This type of game theory is most suited to situations in political sciences and international relations where “power” plays an important role. Noncooperative game theory is more concerned with how strategic choices are analyzed. It requires less details as it focuses more on the ordering and timing of players’ choices as this is considered to be vital in determining the outcome of the game. This game is distinguished from cooperation because the modeling is done around the fact that players are making choices in their own interest, and while some cooperation might take place, it is only when players find it in their best interest. The ultimate goal of game theory, regardless of what type of game is being played, is to predict how the game will be played by rational players or at least how best to play rational opponents.

Prisoner’s dilemma is one of the most popular examples of game theory in social sciences. The game occurs between two players where each player has two strategies, cooperate or defect. The term was coined by Albert W. Tucker in 1950 to describe a situation where two prisoners are taken into custody but the police officers only has evidence to arrest one of them (Holt and Roth 2004). The details of the game are that both criminals are questioned separately at the same time and thus they do not know what the other says – this is a simultaneous game. The prisoners

are then presented with different outcomes in an attempt to persuade them to confess to the crime. At this point both prisoners are aware of the same deal and know the consequences of their decision, which brings in the assumption of complete information and complete knowledge.

In Prisoners' dilemma the defect strategy dominates the cooperate strategy. It is assumed that a rational player will never choose to play a dominated strategy since the player will always be better off by switching to the other strategy. Thus within this game players will always choose to defect and some view this as an inefficiency of the model. One way that this can be overcome, as pointed out by Turocy and Stengel, is by playing the game repeatedly so that the cooperation strategy can be viewed as rational behavior to the players. In other words the fear of punishment in the future outweighs the benefits of defecting in the present.

Rationality, Deterrence, and Crime

Within the rational choice model used by economists, a new theory of criminal behavior was shaped to form what is known in the economics of crime literature as deterrence theory. Akers (1990) highlighted the link that binds deterrence theory to rational choice theory. He noted that both theories were based on the foundation of the utilitarian view of rational human behavior. Specifically Akers (1990, p. 654) noted that "Both theories assume that human actions are based on 'rational' decisions, that is, they are informed by the probable consequences of that action." He noted that for rational choice theory, an individual takes all his/her actions, whether lawful or criminal, into account when trying to maximize his/her payoff and minimize his/her cost. Similarly, for deterrence theory an individual considers the legal punishment of a crime against the motivation before engaging in the criminal behavior.

Becker (1968) was the first researcher to apply economic models of rational decision-making to crimes, and according to his work, "The economic analysis of crime starts with one simple assumption: Criminals are rational. A mugger is a mugger

for the same reason I am a professor-because that profession makes him/her better off, by his/her own standards, than any other alternative available to him/her. Here, as elsewhere in economics, the assumption of rationality does not imply that muggers (or economics professors) calculate the costs and benefits of available alternatives to seventeen decimal places-merely that they tend to choose the one that best achieves their objectives" (p. 43). In this seminal work, a supply offense function was derived which he defined as "relating the number of offences by any person to his/her probability of conviction, to his/her punishment if convicted, and to other variables, such as the income available to him/her in legal and other illegal activities, the frequency of nuisance arrests, and his willingness to commit an illegal act" (Becker 1968, p. 177).

Becker's model was then extended by Ehrlich (1973), who considered a time allocation model. Thus, in the basic economic crime model, individuals choose between allocating their time either to legitimate activities or illegitimate activities. Thus, crime and legal employment are viewed as substitute activities (i.e., if one chooses a legal activity, he/she will have less time for criminal activities), and the difference between legal and illegal opportunities is considered into the model (see Entorf and Spengler 2002). Researchers (see Entorf and Spengler 2002) have raised concerns, however, as to whether this assumption is helpful in explaining crimes by certain groups (e.g., juveniles, low-paid individuals), and made a number of important theoretical and empirical contributions to the economics of crime model since then (for a discussion see Witte and Witt 2002). The Becker-Ehrlich model, however, has formed the foundation for the literature on the economics of crime in which individuals are considered to be rational utility maximizers who take into account the cost and benefit of a crime before engaging in criminal activities, and thus, attitudes toward risk are essential to the model.

A Simple Deterrence Model

In the economic model of crime, criminals are assumed to be utility maximizers who seek to optimize their benefits under restrictions and risk

(Becker 1968). In the case of theft, burglaries, or property crime, the benefits are comprised of the material gains. In the case of violent crime, the payoffs are the transgressor utility derived from the infliction of an assault. Punishment is the major restrictor of crime since punishment is a cost imposed on the transgressor. The risk facing the transgressor is the probability of being caught. Mathematically this may be simply expressed as

$$\pi_i(s_i) = s_i(g - cp) \quad (1)$$

where i is the criminal individual, π_i is the profit from the criminal activity, s_i is the tendency for the criminal to commit a crime, g is the combined monetary and psychic payoff for a certain criminal action, c is the probability of getting caught, and p is the punishment when caught. As seen from Eq. 1, higher punishment reduces criminal activities.

A major limitation of Becker's analysis is, however, the need for strategic interaction. The detection of crime is not an exogenous variable. The detection rate of criminal activity is a function of the actions of police officers, public prosecutors, and lawyers (Tsebelis 1989; Rauhut and Junker 2009). Additionally, criminals and law inspectors are engaged in a "discoordination game." Criminals tend to be more inclined toward committing a crime when they believe that they will not be caught or punished. In contrast, law inspectors are more inclined toward the investigation and inspection of criminals if it is believed that they will detect the crimes of the criminals (Rauhut and Junker 2009).

When game theory is applied to crime, criminals and law inspectors are assumed to be in a similar state of discoordination. In this game, if the law inspector detects a criminal committing a crime, they receive a reward r . The cost of the inspection for the law inspector is also given by k , thus the game may be represented by Table 1.

Rationality, Table 1 The inspection game

		Law inspector	
		Inspection	No inspection
Criminal	Crime	$g-p, r-k$	$g, 0$
	No crime	$0, -k$	$0, 0$

The information in this table shows that if the criminal commits a crime and there is inspection, his/her payout will be $g-p$. Since it is assumed that the criminal is caught in this scenario, the reward for the law inspector will be $r-k$. This is the material gain from the inspection less the cost of inspection. If the criminal committed no crime, but the inspector did an inspection, then the reward for the criminal will be 0; however, the inspector will bear the cost $-k$. If no inspection was done, and no crime was committed, then the payoff is 0 for both parties. In the scenario where no inspection was done, but the criminal committed a crime, the criminal will receive his/her payout of g .

Mathematically the payout functions for the criminal against the inspector and the inspector against the criminal are, respectively,

$$\pi_i(s_i, c_j) = s_i(g - c_j p) \quad (2)$$

$$\emptyset_j(s_i, c_j) = c_j(s_i, r - k) \quad (3)$$

where $\emptyset_j$ is the payoff function for the inspector, j who plays against the criminal, r is the reward for the law inspector for detecting a crime, k is the cost of the inspection for the law inspector, s_i is the tendency of the criminal to commit a crime, and c_j is the probability of getting caught. The Nash equilibrium of the joint strategies may be derived by finding the partial derivative of the payout functions (Rauhut and Junker 2009). The criminal finds the partial derivative of the law inspectors' payoff function $\partial \emptyset_j / \partial c_j$ and sets it to 0. The criminal then expresses his/her probability of committing a crime as:

$$s_i^* = k / r \quad (4)$$

The law inspector finds the partial derivative of the criminal's payoff function $\partial \pi_i / \partial s_i$ [i.e., the partial derivative of the profit for the criminal activity ($\partial \pi_i$) divided by the partial derivative of the tendency for the criminal to commit a crime (∂s_i)] and set it to 0. The law inspector expresses his/her probability of successful inspection as:

$$c_j^* = g/p \quad (5)$$

The results of such game theory are counterintuitive. It suggests that higher punishment does not reduce crime; instead it reduces inspection behavior. Empirical evidence has been mixed on the relationship between crime and punishment. Gibbs (1968) found a negative association between the homicide rate and the severity of criminal punishment. Tittle (1969) found a weak-to-moderate negative association between the severity of criminal punishment and the homicide rate. Saridakis (2011), for example, found that costly deterrence-based policies adopted by governments may have a weak impact in deterring criminals from serious violent crime in the long run. In the following section, we discuss some of the findings documented in the literature and highlighted differences that have been observed between property crime and violent crime.

Empirical Evidence Based on Deterrence Model

According to the Federal Bureau of Investigation (FBI), four offenses – murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault – are defined as violent crimes. They went on to note that any crime that involves the force of threat of violence is deemed to be a violent crime. On the other hand, property crime is defined as theft-type offenses which involve the taking of property or money. Thus offenses such as burglary, larceny-theft, motor vehicle theft, and arson are considered to be property crimes. A wide range of research has been done on deterrence theory, which assesses the impact that deterrence measures, such as the probability of being caught and punished, has both on property and violent crime levels.

Specifically, Pauwels et al. (2011) noted that there are two categories upon which research in deterrence is categorized, and these are macro-level research and individual-level research. At the macro level researchers use official crime statistics to determine the relationship between objective punishment levels and crime, while on the individual basis, they use survey methods to assess the relationship between sanctions and self-

reported crime. The Bayesian updating method is employed at the individual level and was developed by Edwards et al. (1963) and, according to Saridakis and Sookram (2014, p. 25), "... is based on a subjective belief that individuals begin with prior information, and as individual acquires more information through actions about offending and its outcomes, a rational individual would tend to rely more on this new information and less on his or her own beliefs. ..."

Using cross-sectional data from US states during the 1940s to the 1960s, Ehrlich (1996) was the first to empirically test the crime deterrence model. The results of his model showed a strong negative correlation between crime and criminal justice variables in the 10 out of 14 crime categories that he had included in the model. A major finding of the model was that law enforcement activities were not less effective in combating violent crime relative to property crime. This meant that a deterrence force acting on an individual in his/her decision to engage in crime was not stronger on violent criminal activities as opposed to property crimes. However this result was one of the few to show no distinction in the deterrence impact on property crimes over violent crimes. Woplin (1978), who used a model covering a longer time period, from 1894 to 1967, derived different results. He found that deterrence variables had a stronger impact on property crime as opposed to violent crime. Ehrlich (1996) and Woplin (1978) were not the only ones trying to ascertain the relationship and impact of deterrence variables on crime. A few examples of similar studies are from Entorf and Spengler (2000), Cherry and List (2002), Saridakis (2004), Buonanno and Montolio (2008), Saridakis and Spengler (2012), and Han et al. (2010).

Entorf and Spengler (2000) based their model on the traditional Becker-Ehrlich model but modernized it to include demographic changes, youth unemployment, and income inequality in the urban areas in Germany. In their study they were able to separate property crimes from violent crimes and they discovered that the clear-up rates for property crime were much higher than that for violent crimes. They went on to conclude that the deterrence hypothesis appeared to be

more applicable to property crime, and as such this type of crime was more directly related to a rational offender as opposed to violent crime offenders. Specifically Entorf and Spengler (2000, p. 23) noted in their paper that “For the rational offender, low legal income opportunities increase the probability of committing a crime.”

Other researchers have carried out a number of studies that analyze the deterrence hypothesis and how it relates to both property and violent crimes. In their 2012 study, Saridakis and Spengler, for example, examined the relationship between crime, deterrence, and unemployment in Greece over the period 1991 to 1998 and received similar results to Entorf and Spengler (2000). Using a generalized method of moments (GMM) model, they found that property crimes were significantly deterred by higher clear-up rates and there was a positive relationship between property crime and unemployment. Violent crime on the other hand was not significantly impacted by unemployment and clear-up rates. In explaining this result, they noted that for violent crimes, the lack of impact of clear-up rates and unemployment may be due to the fact that these crimes are driven mostly by impulsive actions, while on the other hand property crimes are generally driven more by economic incentive and rational thinking.

In assessing the relationship between unemployment and violent crimes, Saridakis and Spengler (2012) specifically found that male unemployment had a positive and significant impact on rape levels but this relationship became insignificant when other factors were considered. In terms of female unemployment, the relationship with rape was also significant but strongly negative, and Saridakis and Spengler (2012) noted that this could be explained by the routine activity theory. (Routine activity theory states that crime occurs when three elements come together which are an accessible target, the absence of an adult guardian, and the presence of a motivational offender.)

As noted earlier, one of the reasons that deterrence policy seems to work more effectively on property crimes as opposed to violent crimes is the assumption that human beings are rational. Most of the research done in this area indicates that

perpetrators of property crime tend to be more rational when deciding on whether or not to engage in this crime. This means that these perpetrators consider the consequences of their behavior before engaging in their crimes and as such they will consider any deterrence policies before engaging in the crime. Therefore, based on the assumptions of deterrence theory and on the econometric results by various authors, it can be stated that rationality is more aligned to property crimes while violent crimes are more driven by emotionality and impulsive actions rather than economic incentives.

Given that many researchers have determined that perpetrators of property crime are more rational than violent crime perpetrators, Saridakis in his 2004 paper attempted to determine the factors that influence violent crimes. In his study he examined violent crime in the United States during the 1960–2000 period and found no long-run relationships but significant short-run relationships. Specifically, Saridakis’ (2004) results indicated that because violent crimes could be motivated by a variety of reasons, it was difficult to determine direct influencing factors for violent crimes as compared to economic factors that influence property crimes. With regard to the short run, his results showed that an increase in incarceration rates led to a reduction in the violent crime rates and as such could be viewed as a possible solution for dealing with violent crime. In terms of other factors, Saridakis (2004) found that income inequality had a positive impact only on murder. For the variable alcohol consumption, the study found a positive relationship with overall violent crime, and this results fall in line with other researchers such as Raphael and Winter-Ebmer (2001), Ensor and Godfrey (1993) and Field (1990). Generally this literature provides three reasons for the positive relationship between alcohol consumption and violent crimes:

1. Alcohol tends to influence a person’s rationality when assessing the cost-benefit analysis of engaging in criminal behavior.
2. Offenders tend to consume alcohol after they have committed a crime in an attempt to excuse their behavior.

3. Users of alcohol tend to engage in more risky behavior as well as disregard their assessment of potential dangers and as such increase their chances of victimization.

In concluding his paper Saridakis (2004) noted that generally the effects of prison population on violent crimes are much smaller than that found by earlier researchers, such as Devine et al. (1988) and Marvell and Moody (1997). Saridakis further noted that, with the exception of income inequality and its positive relationship with murder, overall, economic factors tend to have an insignificant impact on violent crimes. Alcohol consumption was found to be one of the most impactful factors and thus he noted that it should be included in any study assessing violent crimes. His main conclusion was that investigating violent crimes is much more complicated than property crimes due to the fact that its perpetrators are less rational and thus there are many more motivating factors for engaging in crime and, along with other researchers such as Hartung and Pessoa (2004), concluded that the Becker-Ehrlich theory of deterrence is better suited to dealing with property crime as the many examples above show that these policies tend to have significant impacts in reducing the incidence of property crime.

Conclusion

Rationality and the theory of rational behavior have made important conceptual advances in explaining choices and decision-making and have been widely used by economists and other social scientists in various areas of research to provide solutions to different types of problems. This entry explores the concept of rationality and its role in formulating and shaping theories and in particular the association between the rationality principle and its impact on neoclassical economics. As shown rationality is a key component of the game theory aspect of microeconomics, and an analysis of the game theory example of the prisoner's dilemma is undertaken along with how rational behavior influences choices. The entry then focuses on the economics of crime

theory, in which criminals are assumed to behave as rational utility maximizers. The entry provides a detailed evaluation of rationality and the two major categories of crime – violent and property. It was pointed out, however, that the concept of rationality is better suited to formulating policy targeted to property crime rather than violent crime.

Cross-References

- ▶ [Cost–Benefit Analysis](#)
- ▶ [Equilibrium Theory](#)
- ▶ [Good Faith and Game Theory](#)
- ▶ [Prisoner's Dilemma](#)

References

- Akers RL (1990) Rational choice, deterrence, and social learning theory in criminology: the path not taken. *J Crim L Criminol* 653
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76(2):169–217
- Buonanno P, Montolio D (2008) Identifying the socioeconomic determinants of crime across Spanish provinces. *Int Rev Law Econ* 28(2):89–97
- Cherry TL, List JA (2002) Aggregation bias in the economic model of crime. *Econ Lett* 75:81–86
- Cournot A (1838) *Mémoire Sur Les Applications Du Calcul Des Chances À La Statistique Judiciaire*. journal des mathématiques pures et appliquées 12. T. 3
- Devine JA, Sheley JF, Smith D (1988) Macroeconomic and social-control policy influences on crime rate changes, 1948–1985. *Am Soc Rev* 53(3):407–420
- Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. *Psychol Rev* 70:193–242
- Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. *J Polit Econ* 81(3):521–565
- Ehrlich I (1996) Crime, punishment and the market for offense. *J Econ Perspect* 10(1):43–67
- Ensor T, Godfrey C (1993) Modelling the interactions between alcohol, crime and the criminal justice system. *Addiction* 88(4):477–487
- Entorf H, Spengler H (2002) *Crime in Europe: causes and consequences*. Springer, New York
- Entorf H, Spengler H (2000) Socioeconomic and demographic factors of crime in Germany: evidence from panel data of the German states. *Int Rev Law Econ Rev* 20:1323–1357
- Field S (1990) Trends in crime and their interpretation: a study of recorded crime in post-war England and Wales. Home Office Research Study, vol 119. HMSO, London

- Friedman M (1953) The methodology of positive economics. In: Friedman M (ed) *Essays in positive economics*. Chicago University Press, Chicago
- Gibbs JP (1968) Crime, punishment, and deterrence. *Soc Sci Quart* 48:515–530
- Han L, Bandyopadhyay S, Bhattacharya S (2010) Determinants of Violent and property crimes in England and Wales: a panel data analysis. Department of Economics, University of Birmingham Discussion Paper 10-26R
- Hartung G, Pessoa S (2004) Demographic factors as determinants of crime rates. Website: http://www.abep.nepo.unicamp.br/SeminarioPopulacaoPobrezaDesigualdade2007/docs/SemPopPob07_1062.pdf
- Holt CA, Roth AE (2004) The Nash equilibrium: a perspective. *Proc Natl Acad Sci* 101(12):3999–4002
- Langlois R (1997) Cognition and capabilities: opportunities seized and missed in the history of the computer industry. In: Garud R, Nayyar P, Shapira Z, dir (eds) *Technological innovation: oversights and foresights*. Cambridge University Press, New York, pp 71–94
- Marvell T, Moody C (1997) The Impact of Prison Growth on Homicide. *Homicide Studies*, 1(3):205–233
- Marshall A (1961/1890) *Principles of economics*, 1st edn. Macmillan, London
- Pauwels L, Weerman F, Bruinsma G, Bernasco W (2011) Perceived sanction risk, individual propensity and adolescent offending: assessing key findings from the deterrence literature in a Dutch sample. *Eur J Criminol* 8:386–400
- Popper K (1945/1966) *The open society and its enemies*, vol 1, 5th edn. Princeton University Press, Princeton
- Popper K (1967) The rationality principle. In: Miller D - (ed) *Popper selections*. Princeton University Press, Princeton, pp 357–365
- Raphael S, Winter-Ebmer R (2001) Identifying the effect of unemployment on crime. *J Law Econ* 44(1):259–284
- Rauhut H, Junker M (2009) Punishment deters crime because humans are bounded in their strategic decision-making. *J Artif Soc Soc Simul* 12(3)
- Saridakis G (2004) Violent crime in the United States of America: a time-series analysis between 1960–2000. *Eur J Law Econ* 18:203–221
- Saridakis G (2011) Violent crime and incentives in the long-run: evidence from England and Wales. *J Appl Stat* 38(4):647–660
- Saridakis G, Spengler H (2012) Crime, deterrence and unemployment in Greece: a panel data approach. *Soc Sci J* 49:167–174
- Sardakis G, Sookram S (2014) Violent crime and perceived deterrence: an empirical approach using the offending, crime and justice survey. *Econ Issue* 19(1):23–55
- Simon HA (1957) *Models of man: social and rational*. Wiley, New York
- Simon HA (1972) Theories of bounded rationality. In: McGuire CB, Radner R (eds) *Decision and organization*. North-Holland, Amsterdam
- Tittle CR (1969) Crime rates and legal sanctions. *Soc Probl* 16:408–423
- Tsebelis G (1989) The abuse of probability in political analysis: the Robinson Crusoe Fallacy. *Am Polit Sci Rev* 83:77–91
- Turocy TL, Stengel B von (2001) *Game theory*. CDAM research report LSE-CDAM-2001-09
- Vanberg V (1993) Rational choice, rule-following and institutions— an evolutionary perspective. In: Maki U, Gustafsson B, Knudsen C (eds) *Rationality, institutions, and economic methodology*. Routledge, London
- Woplin K (1978) An economic analysis of crime and punishment in England and Wales, 1847–1967. *J Polit Econ* 86(5):815–840
- Witte AD, Witt R (2002) Crime causations: economic theories. In: Dressler J (ed) *Encyclopedia of crime and justice*, vol 1, MacMillan reference library. Free Press, New York, pp 302–306

REACH Legislation

Sylvain Béal¹, Marc Deschamps¹ and Philippe Solal²

¹CRESE EA3190, Université Bourgogne

Franche-Comté, Besançon, France

²CNRS UMR 5824 GATE, Université de Saint-Etienne, Saint-Etienne, France

Abstract

Since 13 December 2006, the European Union is in a phase of implementation of a new harmonized legislative framework in the field of chemical industry: the REACH regulation (an acronym for Registration, Evaluation, Authorization, and Restriction of Chemicals). From a law and economics perspective, a large set of questions emerge with this regulation. In our entry, we will only shed some light on those which have already been analyzed, such as competition, data sharing, innovation, and socioeconomic analysis.

Introduction

The European REACH legislation is crucial for at least two reasons: it concerns the European chemical industry, and it is one of the most complex legislations that the European Union has ever decided to adopt and implement.

Chemical products are both found in nature and man-made. In our daily lives (food, paint, glue, ink, medication, cosmetics, etc.), they take the forms of *substances*, i.e., chemical elements and their compounds; of *preparations*, i.e., mixtures or solutions composed of at least two substances; and of *articles*, i.e., objects for which the shape, surface, or pattern are more important than their chemical components. The European Commission estimates that close to 100,000 different substances are used on the European territory. The chemical industry in the European Union is responsible for approximately 21% of world sales, is the world leader of exports, has several influential multinational corporations, and directly employs 1.2 million people.

Apart from its sectorial importance, the REACH regulation is also exceptional from the legal and institutional point of view (see Bergkamp 2013). It took almost 9 years of discussions and confrontations (1998–2006), including a White Paper (COM 2001, 88 final) and an internet consultation in May 2003 which received more than 6000 contributions, to publish this legislation in the Official Journal of the European Union on 30 December 2006, coming into force on 1 June 2007.

This legislation is originally constituted by Regulation No 1907/2006, consisting of 850 pages and 141 articles, as well as by Directive 2006/121/EC. It also creates a new European body, the European Chemicals Agency (ECHA), whose purpose is to ensure its proper application. This entry provides a presentation of the REACH regulation, which applies to all members of the EEA, in four sections: the genesis and objectives of REACH (1), the structuring principles of REACH (2), ECHA (3), and the role of socioeconomic analysis in REACH (4).

Origins and Goals of REACH

At an informal meeting in Chester on 26 April 1998, the European environment ministers engaged in a debate on community policy on chemicals. They came to the conclusion that a revision of the latter was necessary. The European Commission then assessed the four main legal instruments governing chemical substances (Regulation No 793/93 and Directives

67/548/EEC, 88/379/EEC, 76/769/EEC) and diagnosed at least three very serious problems.

First, the system made a distinction between substances marketed before September 1981 and those marketed afterward, requiring only the latter to undergo testing and assessment of risks to health and the environment. This led, as explicitly stated in the White Paper EU Commission (2001, p. 6), to “a general lack of knowledge about the properties and uses of existing substances,” since the substance of the previously existing substances represented more than 99% of the total volume of substances on the European market. In addition, it was pointed out that it seemed inadequate, inefficient, and costly to entrust the assessment of dangerousness to the authorities and not to the firms. Finally, the report stressed that current legislation only required manufacturers and importers to provide information on their products, thus neglecting all data on downstream uses (industrial users and formulators).

With this in mind, the European Commission set out in the White Paper of 27 February 2001 a number of proposals for action and the draft of a concerted European policy to solve these problems, thus initiating the steps that would lead to the adoption of REACH. It should be noted that in the impact studies used by the European Commission to evaluate the results that could be expected following the adoption of REACH, it was explained that the positive effects on public health would be in the order of 50 billion euros (for a presentation of the various impact studies that preceded the adoption of REACH, see Schuseil 2013). To this could also be added the benefits of innovation if the Porter hypothesis proved to be valid (see Ambec et al. 2013; Arfaoui et al. 2014, for more details on the links between regulation and innovation).

The first paragraph of Article 1 of Regulation No 1907/2006, setting out the political objectives of the strategy proposed by the European Commission in the White Paper, makes clear that REACH has three hierarchical and cumulative objectives: (1) to ensure a high-level protection of human health and the environment, (2) to allow the free movement of substances in the internal

market, and (3) to improve competitiveness and innovation.

Finally, it should be noted that this legislation is one of the main manifestations of the EU's commitment to the plan adopted at the Johannesburg World Summit in 2002, which aimed to ensure that, from 2020, chemicals would be produced and used in a manner that minimizes health and environmental effects.

REACH's Structural Principles

The overall functioning of REACH is based on a single integrated system consisting of four phases: registration, evaluation, authorization/restriction, and controls and sanctions. We will concentrate mainly on the first phase.

To remove the main flaws of the previous system, REACH has, on the one hand, abolished the distinction between existing substances and new substances and, on the other hand, reversed the burden of proof of safety. Thus, as a matter of principle, all substances and preparations must be notified to ECHA by the companies wishing to use them, except for the exemptions referred to in Article 2 (e.g., radioactive substances, medicines, waste), those requested by the member states in respect of their national defense, and the derogations provided for in Annexes IV and V to the Regulation. It is therefore up to the manufacturers, importers, or downstream users of chemicals to provide the authorities with all the information necessary for the demonstration of the absence of effect on the health and environment of the latter.

In the absence of such registration, the manufacture, importation, and use of these products are illegal, as Article 5 states: "no data, no market." Given the particularly substantial dimensions of REACH, the Regulation provided that these obligations would be implemented gradually and with a growing regulatory requirement based on tonnage. The implementation of the REACH registration phase takes place according to the following timetable: (1) before 30 November 2010 for substances produced at more than 1,000 tons per year, (2) before 31 May 2013 for substances produced between 10 and 100 tons per

year, and (3) before 31 May 2018 for substances produced at more than 1 ton per year. Furthermore, REACH requires the existence of a chemical safety report in the registration dossier only for substances produced at more than 10 tons per year.

To our knowledge, the aspect that has received the most attention in terms of registration is information. In order to avoid duplication costs and vertebrate animal testing, REACH has imposed an obligation on manufacturers, with the exception of the derogation until 1 June 2018 (the date of the end of the implementation of REACH), to share the information they have in SIEF (Substance Information Exchange Forum). Three elements emerged that could be problematic: information content, competitive aspects, and financing. We shall now return successively to these three points.

The first was whether the information required by REACH should relate only to risks known with certainty. This issue has been clearly resolved since Article 191 of the Regulation states that the EU's environmental policy is based, *inter alia*, on the precautionary principle. As a result, manufacturers must also take into account the potential risks in their file and indicate which preventive actions they intend to take (on the economic analysis of the precautionary principle, see Gollier et al. 2000).

The second issue was whether companies could not use REACH as a support for anti-competitive strategies (Béal et al. 2011). The most common idea would be to strategically use this sharing of technical information, either to exchange commercial information and thus reach cartels or, on the contrary, to create extremely expensive data to prevent certain firms from remaining in the market by increasing their costs and thereby abusing of a dominant position. Recital 48 and Article 25 of the Regulation recall that the application of the European competition rules remains valid. We are not aware of any infringements relating to possible anticompetitive behavior these last years. This being said, it should be pointed out that the latest general report of the

European Commission (2013, p. 5) stresses that: “The cost of REACH registration has discouraged some companies from competing on certain substances markets, which in these cases have increased market concentration and prices.”

Finally, the third element was related to the financial impact of data sharing, i.e., the question of how to implement a compensation mechanism between the various stakeholders. As an indication, as of September 2007, ECHA proposed a guide which it updated last November (see ECHA (2016)). We can also note, notably thanks to the work of Dehez and Tellone (2013), Béal et al. (2010), Béal and Deschamps (2016), and Béal et al. (2016), that this issue was the subject of a thorough study leading to the conclusion that cooperative game theory offers both a structured discussion path and clear answers depending on the properties desired by the stakeholders.

Lastly, it should be noted that the EU member states operate at a double level to make the regulatory system work, since they are both responsible for providing a national free assistance service for industrialists (art. 124) and also for participating in the controls, inspections, and sanctions of offenders on their territory (art. 126), which resulted in a change in their domestic legal systems.

ECHA (<https://echa.europa.eu>)

ECHA (European Chemicals Agency) has been operational since 1 June 2007 with headquarters in Helsinki, Finland. It employs approximately 600 people and has an annual budget of around 100 million euros. Its main role is to implement, inform, and enforce the European REACH regulations as well as those relating to the classification, labeling, and packaging of chemical substances (CLP, No. 1272/2008), to biocidal products (BPR, No. 528/2012), and to prior informed consent for imports and exports of certain dangerous chemicals (PIC, No. 649/2012).

In addition to its administrative nature, ECHA has the particularity of housing within it a court:

the Board of Appeal. It decides independently on appeals against certain decisions taken by ECHA under the REACH and BPR regulations (on the link between the question of the sharing of data and the Board of Appeal, see Béal and Deschamps 2016). Its decisions, where they are not final, are subject to appeal before the European Union Tribunal and an appeal to the Court of Justice of the European Union.

Socioeconomic Analysis Inside REACH

Socioeconomic analysis (SEA) is used in the REACH regulation either in the context of a request for the use of a substance subject to authorization (Annex XIV) or in the context of a restriction proposal (Annex XVII). In each of these cases, the objective is to assess the costs and benefits to society of the use (or nonuse) of a substance. In its checklist, ECHA distinguishes among five different types of impacts to be taken into account in the context of a SEA: risks to human health, environmental risks, economic impacts, social impacts, and economic impacts. In addition to the cost-benefit analysis, the ECHA guide (2011) recommends the use of multicriteria analysis, cost-effectiveness analysis, compliance cost analysis, and macroeconomic modeling. The use of SEA can be considered to offer two valuable advantages: the transparency of the elements taken into account and a structured framework allowing the consultation of the stakeholders.

Summary

As REACH regulation will completely come into force on 1 June 2018, it is difficult to have a sufficient knowledge on it before almost 2028. Nevertheless, the early stages of implementation have already produced a new European chemical market structure and innovation behavior which have led to new developments in law and economics. We are convinced that for the next 10 years, lawyers and economists will have much to analyze both theoretically and on the application of REACH.

Cross-References

- [Competition Policy: France](#)
- [Cost–Benefit Analysis](#)
- [Innovation](#)

References

- Ambec S, Cohen MA, Elgie S, Lanoie P (2013) The Porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness? *Rev Environ Econ Policy* 7(1):2–22
- Arfaoui N, Brouillat E, Saint-Jean M (2014) Policy design and technological substitution: investigating the REACH regulation in an agent-based model. *Ecol Econ* 107:347–365
- Béal S, Deschamps M, Ravix J, Sautel O (2010) Les informations exigées par la législation REACH : une analyse du partage des coûts. *Rev Econ Polit* 120:991–1014
- Béal S, Deschamps M, Ravix J, Sautel O (2011) La mise en place de la législation européenne REACH : une analyse des effets anticoncurrentiels. *Revue de l'OFCE* 118:113–128
- Béal S, Deschamps M (2016) On compensation schemes for data sharing within the European REACH legislation. *Eur J Law Econ* 41:157–181
- Béal S, Deschamps M, Solal P (2016) Comparable axiomatizations of two allocation rules for cooperative games with transferable utility and their subclass of data games. *J Publ Econ Theory* 18: 992–1004
- Bergkamp L (ed) (2013) The European union REACH regulation for chemicals: law and practice. Oxford University Press, Oxford
- Dehez P, Tellone D (2013) Data games: sharing public goods with exclusion. *J Publ Econ Theory* 15: 654–673
- ECHA (2011) Guidance on the preparation of socio-economic analysis as part of an application for authorisation. Version 1, January, 260 p
- ECHA (2016) Guidance on data sharing. Version 3, November, 200 p
- EU Commission (2001) Strategy for a future chemicals policy-white paper. COM(2001) 88 final, 27 February, Brussels, 32 p
- EU Commission (2013) General report on REACH. COM(2013) 49 final, 5 February, Brussels, 15 p
- Gollier C, Jullien B, Treich N (2000) Scientific progress and irreversibility. An economic interpretation of the precautionary principle. *J Public Econ* 75: 229–253
- Schuseil Ph (2013) Les études d'impact du règlement REACH. Rapport pour les Cahiers de l'évaluation du Trésor, 126 p

Real Options

Peter-Jan Engelen^{1,2} and Danny Cassimon³

¹Utrecht School of Economics (USE),

Utrecht University, Utrecht, The Netherlands

²Faculty of Business and Economics,

University of Antwerp, Antwerpen, Belgium

³Institute of Development Policy and

Management (IOB), University of Antwerp, Antwerp, Belgium

Definition

The incorporation and valuation of operational flexibility and of strategic value of investments in the decision-making process under uncertainty. The option nature of the investment decision arises from the coexistence of uncertainty, irreversibility, and timing flexibility and allows reaping upside potential while insulating from downside risk.

Introduction

The decision to allocate resources to investment opportunities is traditionally evaluated using a net present value (NPV) approach for private sector projects or a cost-benefit approach for public sector projects. Typically, it sums all the incoming and outgoing cash flows over the lifetime of a project, each discounted at an appropriate risk-adjusted discount rate to derive its present value. The latter is compared with the (current) initial investment cost needed to start the project. In case the outcome of this comparison renders a positive result, one decides to invest; in the other case, one rejects the project. While the NPV approach is a very valuable and a beautiful decision tool, its inherent limitations are well documented: it supposes a now-or-never decision and assumes the decision-maker to follow a rigid path once the investment decision is taken (Feinstein and Lander 2002).

In reality, in a dynamic environment with uncertainty and change, projects will often not materialize in the same shape as the decision-maker has initially expected, and he will have to

adjust the initial plans (Cassimon et al. 2004). During the lifetime of the project, new information may arrive or particular sources of uncertainty may be resolved, making it thus valuable to modify the project (Trigeorgis 2000). For instance, market demand for a product remaining below expectations will call for scaling down the project or vice versa; when demand is above expectations, a scale up might be appropriate. However, the NPV or cost-benefit model cannot handle operational flexibilities such as delaying, scaling up, scaling down, shutting down/restarting, or abandoning a project (Guerrero 2007). Moreover, the NPV model cannot handle strategic dimensions of projects either (Kester 1984). This is the case when different investment projects are not independent of each other or when one specific project consists of different interconnected phases (Dixit and Pindyck 1994). A classic example is research and development (R&D). If a decision-maker would rigidly apply the cost-benefit rule to an R&D project, he would never implement it as its NPV *as a stand-alone project* would be negative. However, in reality many companies nevertheless invest in R&D. This illustrates the limitations of an NPV framework. Such capital budgeting decisions are better handled using a real option framework (Trigeorgis 2000).

This contribution will show that the optional nature is valuable in an investment environment characterized by the simultaneous existence of uncertainty, irreversibility of investment, and some freedom on the timing of the investment (Pindyck 1988). The added value of a real option approach lies in the fact that decision-makers focus more explicitly on proactively considering operational flexibilities or strategic aspects embedded in investment opportunities, rather than making a now-or-never decision to rigidly implement a project as initially planned (Trigeorgis 1996). Moreover, it also allows them to put a precise value on these flexibilities or strategic aspects. As such real option models are an important tool to replace soft decision-making based on vague and qualitative factors, which leave decisions more open to erroneous or manipulative outcomes.

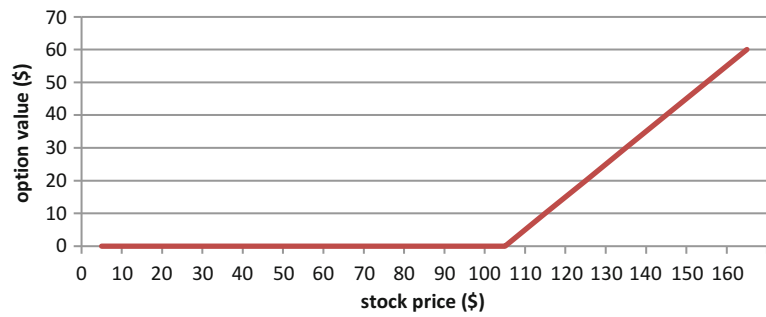
Financial Options and Analogy with Real Options

Recognizing the project itself or particular features of a project as having option characteristics is therefore the key to applying insights of real option modeling. In general, an option can be defined as the right, but not the obligation, to buy (in that case it is labeled a call option) or sell (in that case it is labeled a put option) the underlying asset at an agreed price (strike price or exercise price) during a specific period (as in the case of American options) or at a predetermined expiration date (as in the case of European options) (Hull 2011). Financial options exist on shares, on stock indices, on bonds, on currencies, and on many other financial assets.

For instance, a share of Apple Inc. is trading at \$93.50 in the market. A European call option on one stock of Apple with an exercise price of \$100 and expiration date in 3 months would cost about \$2.40 today. This call option gives the holder the right to buy one share of Apple on the expiration day at \$100. Whenever the share price at expiration is higher than the exercise price, the holder will make a profit by acquiring the share through exercising the option (at \$100) and selling it at the higher market price at that moment (e.g., \$120). In case the share price ends below the exercise price, the holder will not exercise the option and let it expire. Figure 1 shows the typical payoff profile of a European call option at expiration. In this way the holder of the option shields off negative share price evolutions while benefiting from positive share price movements. For having this luxury, the holder pays the option premium of \$2.40 at the option exchange. If the call option was of the American type instead of the European type, the holder can exercise the option earlier than the expiration date. Cassimon et al. (2007) discuss under what conditions these call options will be exercised earlier.

In contrast to financial options, real options refer to the application of the option concept to real physical investment opportunities. Any decision to go ahead with an investment project can be viewed as exercising an option, whereby the firm has the right to obtain all the underlying cash flows that are resulting from the investment

Real Options, Fig. 1 The payoff profile of a European call option at expiration (excluding option premium)



Real Options, Table 1 Analogy of the value drivers of financial and real options and the impact of option value

Symbol	Financial options	Real options
V	Underlying asset price	Present value of the expected cash flows
I	Strike price	Investment costs
σ	Volatility of underlying asset return	Volatility of underlying project return
T	Time to maturity	Window of opportunity
r	Risk-free rate	Risk-free rate
δ	Amount of dividend payments	Opportunity costs of not exercising the option

project (the so-called value of the project, with symbol V), at a particular known cost (the investment cost, with symbol I); the latter is analogous to the exercise price in financial options. When the firm decides to go along with the investment project, it executes the option. In this way, one can draw a parallel between the structure of financial options' payoffs and the payoffs a firm or investor can obtain from a real investment project. Consequently, the argument is made that fundamental value drivers of financial options are also relevant for the valuation of real investment projects. Table 1 gives an overview of the basic value drivers of financial options and the corresponding variables in real investment projects. Apart from the two determinants that drive also the NPV (as the NPV equals $V-I$), note that the other parameters explicitly account for characteristics that embed flexibility and strategic considerations: the fact that future returns are uncertain, the fact that there is some leeway in timing the investment

decision, and the fact that returns may be foregone as long as the project is not yet started.

Basic Option Models

The valuation models for real options are based on financial option models. Probably the best-known model has been developed by Black and Scholes (1973). Its popularity is derived from its closed-form solution and fast and relatively simple computation. Its main disadvantage is due to the strict assumptions underlying the model: (i) frictionless markets, implying no transaction costs or taxes, nor restrictions on short sales; (ii) continuous trading is possible; (iii) the risk-free (short-term) interest rate is constant over the life of the option; (iv) the market is arbitrage-free; and (v) the time process of the underlying asset price is stochastic and exhibits a process assuming asset prices to be log-normally distributed and returns to be normally distributed. Obviously, any violation of some of these assumptions may result in a theoretical Black-Scholes option value which deviates from the price observed at the market. The option value C according to the Black-Scholes model can be calculated as

$$C = V e^{-\delta(T-t)} N(d_1) - I e^{-r_c(T-t)} N(d_2) \quad (1)$$

$$d_1 = \frac{\ln\left(\frac{V}{I}\right) + \left(r_c - \delta + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \quad (2)$$

$$d_2 = \frac{\ln\left(\frac{V}{I}\right) + \left(r_c - \delta - \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \\ = d_1 - \sigma\sqrt{T-t}, \quad (3)$$

where V is the present value of the project's future operating cash flows, I the exercise price or the project's capital expenditure, $T - t$ the time to expiration (in years), σ the annualized standard deviation of the project return, r_c the continuous risk-free interest rate, δ the opportunity cost of waiting, and $N(d)$ the cumulative normal probability density function.

A more practical model that is often used in applications is the binomial option valuation model. This model assumes that in every time period Δt , the stock price moves either upward with factor u or downward with factor d , with $u = e^{\sigma\sqrt{\Delta t}}$ and $d = 1/u$ (Cox et al. 1979). One then has to determine the tree of future stock prices one is going to use as an approximation for future stock price evolutions. For instance, if time to expiration $T - t$ is 1 year and one uses 250 tree steps n , then the stock moves up or down every $\Delta t = \frac{T-t}{n} = \frac{1}{250}$ or approximately 1 step equals 1 trading day. In this case, the binomial model calculates the current option value by discounting at each node of the tree the upward and downward expected option values:

$$C = \{[C_u \times p] + [C_d \times (1 - p)]\}e^{-r} \quad (4)$$

where C is the current option price, C_u is the expected option value in the upward world, C_d the expected option value in the downward world, p the risk-neutral probability that the stock moves upward with $p = \frac{e^r - d}{u - d}$, and r the risk-free interest rate.

Often more complicated option models are needed to handle the specific nature of the investment opportunities and/or the nature of embedded uncertainty. Such models include jump models (Merton 1976), compound option models (Cassimon et al. 2004), or barrier models (Engelen et al. 2016). A detailed discussion of these models is beyond the confines of this contribution, but we refer the reader to the references for further details.

Different Types of Real Options

Different theoretical types of real options have been developed in the early literature: options to

delay (McDonald and Siegel 1986), growth options (Amran and Kulatilaka 1999), options to abandon (Myers and Majd 1990), or scale options (Trigeorgis and Mason 1987). This section starts by discussing the investment "timing option" (or option to delay). Next, we analyze "growth options" (single stage) or "sequential options" (multiple stages) which open up new options upon exercise. This type of option is important when investments are needed to develop new technology or open up new markets. A fourth type we distinguish is the "option to abandon" the project at salvage value. Finally, there are a series of "operational flexibility options," including the flexibility (at some additional cost) to scale up (expand), to scale down (contract), to stop and restart operations, or to switch to other inputs or outputs in response to market and cost developments; each of them can be valued correctly using a real option framework.

Option to Delay

The option to delay (or wait) focuses on the optimal investment timing. It examines whether a company should implement a project right now or better postpone the investment for x months/years before taking the investment decision. The advantage of waiting is that it allows a company to avoid being stuck in an irreversible loss-making investment if project conditions move in an unfavorable way (Ingersoll and Ross 1992). During the period of waiting, the company can learn or collect more information about sources of uncertainty to take a better informed decision at a later point in time (McDonald and Siegel 1986).

Consider an investment project with an economic life span of 15 years, and it is expected to yield every year free operating cash flows of 120 million euro. To enter into this investment opportunity, the firm has to pay an initial investment expenditure of 800 million euro. The company's cost of capital is 10%, while the risk-free interest rate amounts to 5%. The uncertainty of the market demand for these products, as measured by the standard deviation, is estimated to be 30%. Suppose, management wants to know whether to invest in this project immediately or to postpone the project with 1 year (during which

more information concerning the profitability of the project becomes available). This is an example of an option to delay. According to the traditional NPV rule, this is a valuable project which should be implemented immediately. For, the NPV amounts to $\sum_{t=1}^{15} \frac{120}{(1,10)^t} - 800$ or 112 million euro.

Does real option analysis yield a different result? If the decision is postponed for 1 year, this option to delay has a value of 133 million using a binomial option model with 250 tree steps. To come to this figure, we used the following input parameters: the present value of the cash flows from year 2 to year 16 being 829.75 (V), the investment outlay of 800 (I), the option's time to maturity of 1 year ($T - t$), volatility of 30% (σ), and the risk-free interest rate of 5% (r_F). Calculations can be checked through Derivagem (2014).

Because the real option value (measuring the value of delaying of the investment) exceeds the NPV (measuring the immediate investment value), it is better for the firm to postpone the decision with 1 year instead of investing immediately. In 1 year time, the firm can take a more informed decision. Note that delaying investments is not always possible. Also, firms will invest earlier when there is some opportunity cost of waiting (the parameter δ) such as losing market share or the loss of first-mover advantages. Any opportunity costs by waiting have to be deducted from the real option value. If in our example opportunity costs are estimated to be 30 million euro, the real option value is only

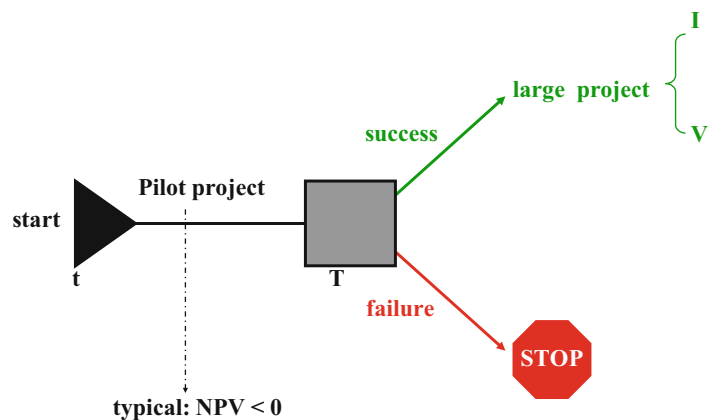
103 million, reversing the company's decision as investing immediately will yield a higher value of 112 million euro.

Growth Option

Growth options occur when projects consist of two phases or stages. The first stage is a prerequisite in order to be able to consider the following phase. Take, for instance, a project that is in fact a pilot scheme for a large-scale project. It is clear that this pilot scheme has option characteristics. The large-scale project can only be considered when the firm has indeed decided to execute the pilot phase (see Fig. 2). However, in case the pilot phase turns out to be a failure, the company does not move to the second phase and just terminates the project (Cassimon et al. 2011b). As such, the pilot phase gives the firm a call option on the large-scale project. The investment cost of the large-scale project and the additional future cash flows that result from it are the execution price, respectively, and the underlying asset of the option (Kester 1984). The value of this call option should be taken into account when calculating the value of the pilot phase, by adding this option value to the conventionally calculated NPV of the pilot phase. It might well be the case then that projects, which would be rejected on the basis of a negative NPV, are now worth executing due to their option value. If the sum of the growth option and the net present value of the pilot phase is positive, it is therefore rational for the firm to invest in the pilot project as the entire project has enough upside

Real Options,

Fig. 2 Typical profile of a growth option



potential to compensate for the (likely) initial losses in the pilot phase. In the other case, the pilot project (and thus the entire project) should be rejected. Investors care about the value drivers of the growth option as well as the pilot project itself.

An example of a growth option is a US home decoration chain exploring the possibility to enter the French market. In order to explore the market potential for their product, the company decides to launch a pilot project with a cost of 110 million euro and 100 million euro expected net operating cash flows in present value terms. It is obvious that the pilot project as a stand-alone project is loss-making (−10 million euro) and should be rejected following the NPV criterion. However, if successful, assume the company plans to launch a large-scale commercialization project opening dozens of megastores across France. The follow-up project is expected to be launched after 2 years at a cost of 200 million euro. At this moment the present value of the operating cash flows of the follow-up phase is estimated to be 180 million. Although the present value of the cash flows is still lower than the investment cost at this moment, things could change over time. The volatility of project returns in this industry is estimated to be 25% per year. The risk-free interest rate is equal to 5%. Using a binomial model with 250 tree steps, the option value of the follow-up phase amounts to 25 million euro. Compared to the value of phase one of −10 million, it makes sense to invest in the pilot project as there is enough upside potential to justify investment. The total project value comes to $-10 + 25$ or 15 million euro. Whether the company will actually invest in phase two is only decided at the end of year 2 and will depend on the market conditions at that moment. For the moment, the company only commits itself to the pilot project, with the possibility for expansion at the end of year 2.

Sequential Option

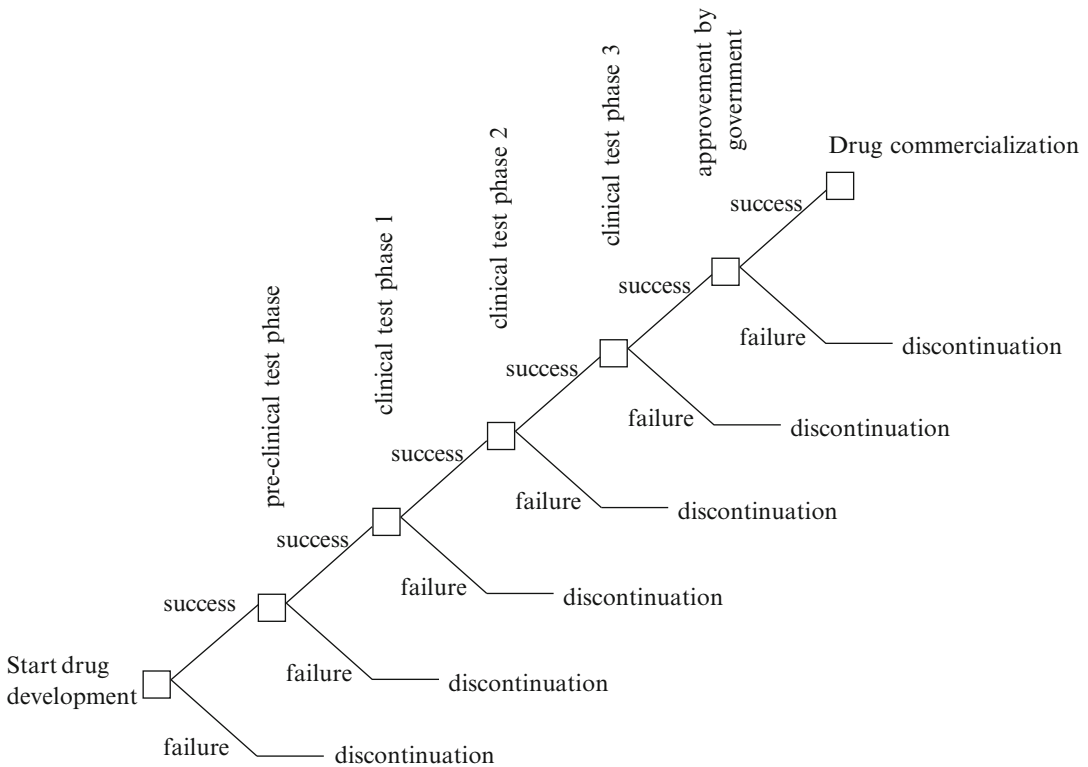
When a growth option involves more than two phases, one has to enlarge the standard option approach to so-called sequential options, which can be valued using compound option techniques (Cassimon et al. 2011a). The option value of the

second phase includes in this case also the option value of the subsequent phase. Put differently, in that case we have an option on an option or a compound option (Geske 1979).

A copybook case of a sequential option is the development of a new drug. A drug development pipeline consists of six distinct and specific phases (see Fig. 3). Developing a new drug is a chain of options starting with a preclinical test phase, followed by three clinical test phases, a governmental approval phase, and ultimately the commercialization phase. The initial R&D phase can be seen as an option on the preclinical test phase. If the initial R&D turns out to be successful, the preclinical phase is started; otherwise the research is being discontinued. The preclinical phase itself is an option on the first clinical test phase. If the preclinical tests are successful, the compound moves to the first clinical test phase; if not, the research is again abandoned. This phase is again an option on the next phase and so on, until the commercialization phase. This is a heavily regulated, standardized, and linear process because each phase is easily identified and has to be finished before the next phase can start. The initial R&D phase, being a sixfold compound option on the commercialization phase, can be valued using a generalized compound option model, such as the one by Cassimon et al. (2004), that was explicitly developed to value compound options of a higher order than two.

Option to Abandon

If market conditions deteriorate, management can also consider terminating and abandoning the project permanently. Having the option to realize a certain salvage value puts a floor to the potential losses of a project and thus adds value in comparison to the same project without this possibility. Management compares the project value as a going concern with the liquidation value the firm can realize upon termination of the project. This managerial decision can be interpreted and valued as a “put option,” the right to sell an asset at a predetermined price at or up to a predetermined point in time (Myers and Majd 1990). The value of this put option should then be added to the project’s traditional (NPV-type) value.



Real Options, Fig. 3 The development of a new drug as a sequential option

Operational Options

Operational options include options to expand, contract, suspend, and resume operations and the option to switch inputs or outputs during operation (Brennan and Schwartz 1985; Dixit 1989; Kulatilaka and Trigeorgis 1994). An example of a switch option is a chemical factory that builds in flexibility at the input or output side. At an additional cost, a company can decide to add input flexibility by building a factory that can switch from one energy source (say oil) to another energy source (electricity or gas). It allows companies to switch to cheaper energy sources during the project. This flexibility comes at a cost, which has to be compared with the value of the switch option. The company can also build in flexibility at the output side by building a factory that allows for various output products, e.g., switching from producing polyvinyl chloride (PVC) to polypropylene (PP) or polyethylene (PE). Again, balancing the value of the switch option to the

additional cost of building a flexible factory is at the core of the real option framework.

A shutdown and restart option assumes that a project does not have to be operated permanently. Depending on the net revenues and (marginal) costs, management can temporarily suspend operations and resume once net revenues again cover the (variable) cost of operation. In option terms, this means that management has the option to receive the project's net revenues minus the variable costs when the project is operated in a given year (Trigeorgis 2000). The value of the project in any given year is thus the maximum of that year's cash revenues minus that year's variable costs of operation or zero. Put differently, the project will be operated if and only if the revenue exceeds the variable costs. This limits the downside risk to the fixed costs, and therefore the project value should respond positively to increasing volatility in variable costs and in operating cash flows. Intuitively, all else equal, in more volatile environments, the

value of operational flexibility is higher (Sanders et al. 2013).

The option to scale up (expand) can be seen as a call option in a similar way as a growth option, while the option to scale down (contract) can be seen as a put option similar to an option to abandon.

Measuring Risk

One feature that all real option models have in common is that they require the input of an estimate of one or more sources of risk (proxying for uncertainty). In this contribution we focus on market risk (also labeled commercial risk) and technical risk.

Market Risk

Market risk is measured as the volatility of the project return between the current moment and the expiration date of the real option. It measures how much the project value, measured as the present value of all expected operating cash flows in the final phase (V), will vary over time. Real options only have value when things can change over time. The higher the volatility, the higher the change that tomorrow's world will materialize differently from what the decision-maker had in mind today. The estimation of the future volatility can be difficult when the amount of available data is limited. Sometimes nothing more than an educated guess is used. A "normal" range for volatility is about 30–40%. Often an educated guess is good enough as an input, especially if changes in volatility have little impact on the investment decision. Applying real option thinking compared to traditional valuation methods is often the key element; fine-tuning the volatility estimate often does not add a lot to insights. Sometimes a more precise estimate is required as the investment decision alters depending on the level of the volatility estimate. Cassimon et al. (2011a) give an overview of different proxies for volatility. A first approach is to use historical volatility derived from historical market data such as oil prices, commodity prices, real estate prices, and so on. A second approach uses volatilities of a sample of comparable firms. A third approach is the calculation of the implied volatility of options on a

comparable firm as a proxy for the future volatility (Schwartz and Moon 2001). Finally, the use of internal company data is sometimes a good strategy. It is a viable approach if the firm has a portfolio of past and current projects, considered to be representative for the risk profile of the current project.

Technical Risk

While market risk, as captured by the volatility of the project return, refers to the normal business risk any company incurs, such as the potential market size, its expected revenues, its expected cost structure, and so on, technical risk refers to a catastrophic event (technical failure) that terminates the project and is unrelated to its cost or benefits. For instance, the uncertainty about the effectiveness or about potential problematic side effects of a new drug is an example of technical risk (Cassimon et al. 2011a). Such technical failure is the type of uncertainty that presents itself as a negative shock to the project value. Technical risk can be captured by a Poisson process (in case the nature of uncertainty is best characterized by discrete jumps) or by using discrete success-failure probabilities at each stage of the project.

Applications

Applications in General

Real option models to project evaluation are applied in a wide range of sectors, such as Internet companies (Schwartz and Moon 2001), the service sector (Jensen and Warren 2001), consumer electronics (Lint and Pennings 2001), pharmaceutical R&D (Cassimon et al. 2011a), the ICT sector (Cassimon et al. 2011b), and sustainable energy solutions (Sanders et al. 2013; Engelen et al. 2016).

Applications in Law and Economics

Some law and economics studies apply real option reasoning to enrich static models on modeling agent's behavior. For instance, Engelen (2004) extends the classic Becker expected utility framework on criminal behavior to provide a more dynamic decision-making framework of criminal activity. While conventional models of crime

focus on the expected net gain of a crime as the difference between the expected profits of a crime and the expected costs (i.e., the product of the amount of punishment and its probability (Engelen et al. 2016)), such a framework ignores the simultaneous existence of uncertainty, irreversibility of the criminal act, and flexibility in the timing when to commit the crime. If those characteristics are present, a criminal can delay committing the crime until more information about the uncertain future becomes available. Waiting some time to commit the crime might imply some risks and foregone profits, but it may prevent the criminal from being trapped in an irreversible crime, which may turn out to be very costly when being caught. Engelen (2004) shows that a crime that satisfies these three characteristics is best treated analogous to holding a financial call option. For some specific time period, a criminal has the possibility, but not the obligation, to pay a certain “price” in return for an asset that has some value. When the criminal decision is made, the option is exercised, which is an irreversible decision.

Cassimon et al. (2013) apply a criminal real option framework to criminal states. They study two episodes of criminal behavior by Rwanda in the Democratic Republic of Congo (DRC): the massive killing of Hutu refugees by the Rwanda Patriotic Army (RPA) in late 1996 to early 1997 and the illegal exploitation of Congolese resources from August 1998. They show how the international community can optimally intervene proactively, by reducing the incentives for criminal states to execute their criminal options.

While the above studies focus on handling uncertainty from an agent’s point of view, other studies use real option modeling from a lawmaker’s point of view. Deffains and Obidzinski (2009) apply the real option approach to the design of legal rules. Rulemakers face the choice between rule-based regulation and standard-based regulation. Detailed rules provide clear guidance to agents but are more costly to create ex ante and risk to become obsolete over time as market conditions vary. Standards are more flexible as they allow to adapt regulation more easily to the changing environment, but they are more costly in court

interpretation ex post as interpretation is more ambiguous. A real option framework shows that this choice depends on the variability of contingencies and on the degree of innovation in the area of the law. Adding the timing dimension and uncertainty to the classic trade-off between rule and standard allows to move from a static to a more dynamic perspective which is closer to real-life decision-making. In a similar vein, Parisi et al. (2004) analyze the value of waiting in lawmaking. Waiting to enact a certain law while collecting more information can be a better choice than being stuck in an ineffective or undesirable rule.

Lee et al. (2007) analyze the impact of bankruptcy law on entrepreneurial risk taking through a real option lens. They argue that society will benefit from more entrepreneur-friendly bankruptcy laws. Speedy bankruptcy procedures, discharging bankrupt individuals from debt, and automatic stay on assets will allow inefficient firms to exit more easily and will encourage potential entrepreneurs to enter the market without having to be afraid of the damaging consequences of a possible bankruptcy. Using a real option perspective, this study clearly shows the impact of the design of legal rules on entrepreneurial activity at the societal level.

References

- Black F, Scholes M (1973) The pricing of options and corporate liabilities. *J Polit Econ* 81:637–659
- Brennan M, Schwartz E (1985) Evaluating natural resource investments. *J Bus* 58:135–157
- Cassimon D, Engelen PJ, Thomassen L, Van Wouwe M (2004) Valuing new drug applications using n-fold compound options. *Res Policy* 33:41–51
- Cassimon D, Engelen PJ, Thomassen L, Van Wouwe M (2007) Closed-form valuation of American call options with multiple dividends using n-fold compound option models. *Finan Res Lett* 4:33–48
- Cassimon D, De Backer M, Engelen PJ, Van Wouwe M, Yordanov V (2011a) Incorporating technical risk into a compound option model to value a pharma R&D licensing opportunity. *Res Policy* 40:1200–1216
- Cassimon D, Engelen PJ, Yordanov V (2011b) Compound real option valuation with phase-specific volatility: a multi-phase mobile payments case study. *Technovation* 31:240–255
- Cassimon D, Engelen PJ, Reyntjens F (2013) Rwanda’s involvement in Eastern DRC: a criminal real options approach. *Crime Law Soc Change* 59:39–62

- Cox JC, Ross SA, Rubinstein M (1979) Option pricing: a simplified approach. *J Financ Econ* 7(3): 229–263
- Deffains B, Obidzinski M (2009) Real options theory for law makers. *Recherches économiques Louvain* 75(1):93–117
- Derivagem (2014) Option calculator. <http://www-2.rotman.utoronto.ca/%7Ehull/software/index.html>
- Dixit A (1989) Entry and exit decisions under uncertainty. *J Polit Econ* 97(3):620–638
- Engelen PJ (2004) Criminal behavior: a real option approach. With an application to restricting illegal insider trading. *Eur J Law Econ* 17(3):329–352
- Engelen PJ, Kool C, Li Y (2016) A barrier options approach to modeling project failure: The case of hydrogen fuel infrastructure. *Resource and Energy Economics* 43:33–56
- Feinstein S, Lander D (2002) A better understanding of why NPV under values managerial flexibility. *Eng Econ* 47:418–435
- Geske R (1979) The valuation of compound options. *J Financ Econ* 7:63–81
- Guerrero R (2007) The case for real options made simple. *J Appl Corp Finan* 19:38–49
- Ingersoll JE Jr, Ross SA (1992) Waiting to invest: investment and uncertainty. *J Bus* 65:1–29
- Jensen K, Warren P (2001) The use of options theory to value research in the service sector. *R&D Manag* 31(2):173–180
- Kester W (1984) Today's options for tomorrow's growth. *Harv Bus Rev* 62(1):153–160
- Kulatilaka N, Trigeorgis L (1994) The general flexibility to switch: real options revisited. *Int J Finan* 6: 778–798
- Lee SH, Peng MW, Barney JB (2007) Bankruptcy law and entrepreneurship development: a real options perspective. *Acad Manag Rev* 32(1):257–272
- Lint O, Pennings E (2001) An option approach to the new product development process: a case study at Philips electronics. *R&D Manag* 31:163–172
- McDonald R, Siegel D (1986) The value of waiting. *Q J Econ* 101:707–728
- Merton RC (1976) Option pricing when underlying stock returns are discontinuous. *J Financ Econ* 3(1): 125–144
- Myers S, Majd S (1990) Abandonment value and project life. *Adv Futur Options Res* 4:1–21
- Pindyck R (1988) Irreversible investment, capacity choice, and the value of the firm. *Am Econ Rev* 79: 969–985
- Sanders M, Fuss S, Engelen PJ (2013) Mobilizing private funds for carbon capture and storage: an exploratory field study in the Netherlands. *Int J Greenh Gas Control* 19:595–605
- Schwartz E, Moon M (2001) Rational pricing of internet companies revisited. *Finan Rev* 36:7–26
- Trigeorgis L (2000) Real options and financial decision-making. *Contemp Finan Dig* 3:5–42
- Trigeorgis L, Mason S (1987) Valuing managerial flexibility. *Midl Corp Finan J* 5:14–21

Further Reading

- Amran M, Kulatilaka N (1999) Real options. Harvard Business School Press, Boston, 246 p
- Dixit A, Pindyck R (1994) Investment under uncertainty. Princeton University Press, Princeton
- Engelen PJ, Lander MW, van Essen M (2016) What determines crime rates? An empirical test of integrated economic and sociological theories of criminal behavior. *Soc Sci J* 53(2):247–262
- Hull JC (2011) Options, futures, and other derivatives, 8th edn. Pearson, Upper Saddle River
- Parisi F, Fon V, Ghei N (2004) The value of waiting in lawmaking. *Eur J Law Econ* 18(2):131–148
- Trigeorgis L (1996) Real options. MIT Press, Cambridge, 427 p

Reasonableness

► Rationality

Refugee Law

► Asylum Law

Regulatory Competition

Katrin Gödker¹ and Lars Hornuf^{2,3}

¹School of Business, Economics and Social Sciences, University of Hamburg, Hamburg, Germany

²Department of Economics, Trier University, Trier, Germany

³CESifo, Max Planck Institut for Innovation and Competition, Munich, Germany

Abstract

Regulatory competition describes the activity of private or public lawmakers who intend to produce novel or alter current legislation in response to competitive pressure from other private or public lawmakers. In this chapter, we provide an overview of the legal arbitrage tactics of various entities, which are a necessary, though not sufficient, condition for

regulatory competition. Moreover, we investigate whether and how lawmakers respond to these endeavors by adapting their national corporate, insolvency, capital markets, and environmental law as well as product standards.

Defining Regulatory Competition

Regulatory competition describes the activity of private or public lawmakers who intend to produce novel or alter current legislation in response to competitive pressure from other private or public lawmakers (Hornuf and Lindner 2016). It is often private lawmakers, nation-states, regions, communities, and supranational organizations that engage in such competitive processes.

The literature distinguishes two forms of regulatory competition. First, Tiebout (1965) outlines a theoretical framework in which lawmakers offer a package of goods and services to legal entities under their jurisdiction. These legal entities can be individuals, companies, or other organizations recognized by law. In return for this package of goods and services, legal entities make tax payments. Unlike lawmakers, legal entities are mobile and spur the competitive process by physically moving to regions where legislators offer the best bundle of goods and services. Companies usually do this by moving their headquarters or by investing in production sites in the respective region. Typical bundles offered by lawmakers may encompass a certain physical and legal infrastructure, the latter of which includes, for example, specific taxes; corporate, bankruptcy, and environmental laws; and product safety and labor market standards.

Second, regulatory competition can also arise from unbundling certain legal rules from the physical location of a legal entity (Heine and Kerber 2002; Eidenmüller 2011). A possible constellation would be when a company locates its headquarter in the United States, adopts the company law of England and Wales, finances projects with debt securities under the law of the Cayman Island, and settles disputes in a Swiss court under French law. If the choice of law is independent of the physical location of a legal entity, this

allows individuals and companies to cherry-pick the rules that are most suitable for their respective business transaction. Moreover, from an economic perspective, lawmakers that engage in this type of regulatory competition can better specialize in certain legal products and do not need to offer a complete bundle of laws to their customers.

The welfare implications of regulatory competition have been subject to a long-standing dispute in the law and economics literature. Most prominently, they were debated in the field of corporate law as unbundled competitive processes. While some scholars have persistently claimed that regulatory competition leads to the implementation of optimal legal rules, also known as the *race-to-the-top* hypothesis (see Winter 1977; Fischel 1982; Romano 1987), others have argued that regulatory competition leads to the prevalence of the lowest respective standards, also referred to as the *race-to-the-bottom* hypothesis (see Cary 1974; Bebchuk and Hamdani 2002). From a multi-level perspective, Sinn (1997) theoretically argues that when government regulation serves a purpose such as to overcome market failure and later becomes subject to competition on a higher regulatory level, regulatory competition again leads to economic inefficiencies.

Regulatory Competition in Corporate Law

In corporate law, conflict-of-law rules define whether legal arbitrage by companies in the respective jurisdictions is possible, which is a necessary, though not sufficient, condition for regulatory competition. Legal arbitrage can be understood as a legal planning technique that is carried out to avoid taxes and other legal rules to circumvent regulatory costs (Fleischer 2010). At the national level, legislators can apply either the real seat theory (*siège réel*) or the incorporation theory when dealing with foreign companies. After World War II, the real seat theory was most prevalent in Western Europe, practically prohibiting a free choice of law independent of corporate headquarters. Nevertheless, the mobility of companies has been a cornerstone of the

Single European Market, leading the European Court of Justice (ECJ) to interpret Art. 49, 54 TFEU (formerly Art. 43, 48 EC Treaty and, before that, Art. 52, 58 EEC), regarding cross-border company mobility. Starting with the *Daily Mail* (Case C-81/87, 27 September 1988) in 1988, the ECJ had largely abolished the real seat theory by the turn of the millennium, with the famous decisions on the *Centros* (Case C-212/97, 9 March 1999), *Überseering* (Case C-208/00, 5 November 2002), and *Inspire Art* (Case C-167/01, 30 September 2003) cases.

Regarding inbound cases, in which a foreign company seeks to immigrate into a certain jurisdiction, the ECJ found that companies registered under the law of their home member state had the right to transfer their head office to another member state. Unlike before, this could now be achieved without liquidating the original company and reestablishing a new legal entity. Consequently, the host member state must recognize the foreign company, which does not need to meet standards such as the minimum capital requirement for incorporation in this country. The founder of a company can thus choose a company law that best suits the business needs independent of the company's real seat. This rule largely paved the way for regulatory competition in the European Union (EU). Conversely, companies that intend to emigrate are still restricted by national legislation, as member states still have the power to prohibit such outbound cases. According to ECJ's *Daily Mail* (Case C-81/87, 27 September 1988), *Cartesio* (Case C-210/06, 16 December 2008), and *National Grid Indus* (Case C-371/10, 29 November 2011) decisions, home member states can require companies to have their registered office and head office under the national law of the respective territory. Otherwise, the company may need to be dissolved. Finally, the ECJ recently enabled existing companies to move their statutory seat to another jurisdiction in the *Cartesio* (Case C-210/06, 16 December 2008) and *VALE* (Case C-378/10, 12 September 2012) decisions of 2008 and 2012. Although such a transfer of the registered office necessarily leads to a conversion of the company into a company governed by the law of the new

member state, the company no longer must be dissolved.

How did national lawmakers react? Since 2003, not less than ten major company law reforms have been implemented in nine EU member states (Hornuf and Lindner 2016). Lawmakers in France (in 2003 and 2008), Hungary (in 2007), Germany (in 2008), Poland (in 2008), Denmark (in 2010), Sweden (in 2010), and the Netherlands (in 2012) were forced to reduce the minimum capital requirement for national legal forms, and many of these jurisdictions abolished the minimum capital requirement altogether. Seven years after the initial reform, Hungary decided to raise the minimum capital requirement back to the initial level. In some cases, lawmakers introduced a new legal form, abolished the notary requirement for setting up a company, or allowed for electronic company registrations and document filings. In almost all cases, the administrative speed of incorporation was increased. These factors were previously found to be the main drivers of legal arbitrage (Becht et al. 2008), and thus regulatory reforms might be considered a response to the competitive pressure exerted, mostly by the English Limited. Braun et al. (2013) show that the reduction of the minimum capital requirement helped countries not only improve the attractiveness of the national legal form but also increase entrepreneurial activities in general.

Unlike in Europe, where regulatory competition might also be driven by the pride of national lawmakers to offer the "regulation of choice," which fosters a prosperous private legal industry, the main objectives for US lawmakers to engage in regulatory competition and to create new company law rules is the ability to raise charter fees. Recently, approximately two-thirds of *Fortune* 500 companies and more than four-fifths of all new US initial public offerings chose to incorporate in Delaware (Bullock 2013). Although evidence shows that companies not incorporating in Delaware are subject to a home-state bias (Bebchuk and Cohen 2003), Delaware still collects approximately one-fourth of its overall tax revenues through business entity taxes and fees as well as Delaware Uniform Commercial Code fees (Bullock 2013). In Europe, such financial

incentives to engage in regulatory competition do not exist as the collection of charter fees is generally prohibited by EU legislation.

Not only does regulatory competition in corporate law take place *horizontally* between different US states or European nation-states, but it also occurs *vertically* between, for example, Canadian provinces and the federal lawmakers or national legislators and the supranational EU. In Europe, a supranational corporate law form was established with the Statute for a European Company (*Societas Europaea*, SE). Importantly, the European Company did not provide a full company law but rather refers some material matters to the national corporate law of the member states. Here, the scope for legal arbitrage emerges. While companies might adopt the European Company for various reasons, early evidence indicates that companies may choose the SE to freeze or renegotiate mandatory codetermination as well as to establish a one-tier board structure, especially if that was not possible under national corporate law (Eidenmüller et al. 2009). Fierce vertical regulatory competition in corporate law did not emerge until more recently. After reviewing the SE Statute in 2012, the European Commission noted that any benefits of a reform would not outweigh the potential challenges. Consequently, there was no intent to improve the competitive stance of the European Company. The proposal of a European private limited liability company (*Societas Privata Europaea*, SPE) faced opposition by some member states early on and so far has not been approved by the Council of the European Union.

Regulatory Competition in Insolvency Law

Insolvency laws in the EU are not harmonized; rather, individual jurisdictions keep their own national rules and proceedings. As a result, legal arbitrage activities of consumers and companies can theoretically spur regulatory competition. In the realm of personal insolvencies, consumers may shop, for example, for a shorter statutory discharge period. The discharge period can range

from under 3 years in countries such as France, Latvia, Poland, or England and Wales and increase to at least 5 years such as in Austria and previously Germany (Drometer and Oesingmann 2015). Other variations in the insolvency regime concern the right of debtors to prevent a discharge because of misconduct on the borrower side, which may exist in some jurisdictions but not in others. From an economic perspective, shorter discharge periods spur consumption and support individual risk taking. By contrast, the prospects of a discharge also increase the moral hazard problem and weaken credit discipline (Adler et al. 2000). During the past decades, reforms in consumer insolvency laws did not, however, solely result from lawmakers engaging in regulatory competition; they are also a consequence of the growing over-indebtedness of consumers in general.

As in the case of personal insolvencies, a corporate insolvency proceeding under European rules is opened under the law of the debtor's center of main interests (COMI). According to the latest Insolvency Regulation Recast, the COMI is the place where the debtor conducts the administration of its interest on a regular basis and which is ascertainable by third parties. Given that almost one-fourth of the approximately 200,000 companies that experienced insolvency in Europe between 2009 and 2011 engaged in cross-border activities, some leeway emerged for them to engage in forum shopping. That companies indeed attempt to engage in forum shopping was prominently shown by the insolvency proceedings of Deutsche Nickel and Schefenacker. The head office of both companies was historically based in Germany; however, by merging or selling the company (partly) to a legal entity under English law, both companies attempted to move their head offices to Great Britain to start an insolvency procedure under English law. Under English rules, creditors can, for example, choose an insolvency administrator of their liking and benefit from extended procedural deadlines, and creditors do not need the consent of the previous owners to swap debt to equity. To the best of our knowledge, this form of legal arbitrage did not become a mass phenomenon, partly because of

legal uncertainty about whether such a move of the company seat is legitimate. While legal reforms in insolvency laws have recently taken place, there is no clear-cut evidence that they were the result of regulatory competition.

In the United States, lawmakers established a single federal bankruptcy code. Nevertheless, some scholars argue that judges still have a sufficient leeway to develop a state-specific approach to bankruptcy law (Skeel 1998) and that courts can apply state law, for example, to determine the fiduciary duties of managers (Skeel 2000). Moreover, LoPucki (2006) claims that judges are often motivated by the glamor of dealing with insolvencies of well-known companies, which brings about a better standing for them in the legal community. Moreover, bankruptcy bars exert pressure, which resulted in business activities moving to New York and Delaware (LoPucki and Kalin 2001). As only minor differences exist in the US bankruptcy code, forum shopping is often limited to large borrowers and Chapter 11 prepackaged bankruptcies (Enriques and Gelter 2007), in which debtors offer a plan to creditors, who then vote in its favor before a bankruptcy case is filed with the court.

Regulatory Competition in Capital Markets Law

Capital markets have unique features that facilitate regulatory competition (Ringe 2016). On the one hand, capital markets are essential for many economic activities and thus put substantial pressure on regulators to provide efficient laws. On the other hand, regulatory competition in capital market law emerges not only from issuers deciding for one regulatory regime from a range of options but also from issuers' cross-listing their securities in different markets across different jurisdictions. For example, a company in Angola could issue securities on the new Angola Stock Exchange and have additional disclosure standards imposed by the New York Stock Exchange. The number of companies listing securities outside their place of incorporation as well as the number of cross-listings has grown significantly, which has

resulted in increased competition between stock exchanges in the United States and Europe (Kim and Pinnuck 2014). A frequently mentioned reason for companies to cross-list is the so-called legal bonding to more rigorous disclosure standards to improve access to capital, which in turn lowers the cost of capital and increases the value of a company (Doidge et al. 2004). Coffee (2002) provides a detailed summary of research on the bonding theory.

The main challenge for lawmakers in formulating adequate regulatory instruments in the financial sector comes from the threat of negative consequences for financial stability (Ringe 2016). Thus, a global effort was made to achieve financial stability and has resulted in widespread attempts of harmonization. Worldwide, the most prominent example is the process of harmonizing banks' capital requirements by the Basel Accords of the G20 group. Moreover, in the aftermath of the financial crisis, the EU and its member states have strengthened capital market regulation as well as supervision and prioritized regulatory convergence over regulatory competition in the financial sector. The Capital Markets Union, a recent initiative of the European Commission, aims to create a single market for capital across all EU member states by removing barriers to cross-border investment. Against the backdrop that issuers want to reach capital providers, investor protection regulation also plays a central role in harmonizing capital markets law. For example, the European Markets in Financial Instruments Directive is a strong step forward in converging regulation regarding securities trading in the EU.

As in the other domains, the question has been raised whether regulatory competition in capital market law is likely to yield a race to the top or a race to the bottom. In line with the legal bonding theory, substantial evidence shows that countries with more disclosure requirements and more stringent corporate governance rules provide a more attractive capital market environment with lower cost of capital for listed companies (Prentice 2002). This should motivate countries to raise their disclosure and governance standards to attract more company listings. However, Prentice (2005) notes that in capital markets, no true

competition emerges and that a race to the bottom is more likely because the choice of listing is not made by companies but by self-interested managers. In addition, some scholars focus on regulatory competition among security exchanges and note that this competition does not necessarily yield a race with respect to listing standards but will more likely result in an endogenous segmentation of the capital market for listings with first-tier and lower-tier market segments (Chemmanur and Fulghieri 2006).

Regulatory Competition in Environmental Law

From a political economy perspective, environmental policy making results from the interaction between lawmakers and various interest groups. These interest groups demand different environmental measures and put pressure on lawmakers to decide in their favor (Oates and Portney 2003). For example, voters who call for stricter pollution regulation can threaten governments to withdraw their support during the next election. In this context, the debate also targets the extent to which the implementation of new regulatory instruments is influenced by well-organized interest groups, such as producers and consumers.

Yet, taking international trade into account, it is argued that when states are confronted with economic competition, they have an incentive to adopt lax environmental standards to attract companies and capital, which leads to a race to the bottom of environmental standards (Engel 1996). Regulatory competition would then result in a reduction of environmental protection efforts to the level of the least stringent state. One main reason for this trend is the presence of excessive market power of the industry (Engel 1996).

However, there is no clear indication that environmental regulatory competition necessarily leads to the lowest possible standard to prevail. There might also be a weaker form of the race-to-the-bottom tendency, in which only some states lower their environmental standards under the competitive pressure, and thus not all states converge to the least stringent state (Konisky 2007).

Moreover, Vogel (1997) postulates that the existence of more stringent environmental standards in one country could even trigger a race to the top as companies in other countries want to export their goods and services to the high-standard country and thus must adopt the more stringent environmental standards (the so-called California effect). To avoid the costs of double regulation, companies might lobby their governments to regulate specific environmental issues in accordance with the high-standard countries that exports target. Indeed, an empirical investigation shows that even before the EU adopted a directive regarding a specific environmental regulation instrument, a diffusion process had begun, kicking off a race to the top among member states (Holzinger and Sommerer 2011).

Finally, more environmental standard setting is implemented and harmonized at the international and supranational level. This can be expected to further increase, for example, under the auspices of the United Nations' Paris Agreement, in which 195 nations agreed to undertake ambitious efforts to combat climate change. With accelerating internationalization of environmental regulation, the opportunities to differentiate and compete between single regulators might vanish. Heyvaert (2013) argues, however, that sufficient scope for decentralized regulators remains to implement differential environmental regulation, particularly because only few international environmental standards are truly global and there is also flexibility within transnational regimes.

Regulatory Competition in Product Standards

Consumers are directly affected by regulatory competition when it comes to product standards. Building on Akerlof's (1970) theory of adverse selection, Sinn (1997, 2003) argues that consumer markets are characterized by information asymmetry between consumers and sellers because consumers cannot fully elicit product qualities. This is especially relevant for products whose quality is not likely to be detected because they are not frequently purchased or their value is too

low to justify intensive information gathering. This information asymmetry can lead to low-quality equilibrium in which informed sellers oversupply low-quality products and undersupply high-quality ones. To reduce the negative consequences of such market failures, regulators specify minimum product standards.

Thus, governments intervene where markets have failed. Yet regulatory competition can emerge as countries engage in protectionist practices by undercutting product standards of competing countries to give their own industries a competitive advantage. According to Sinn (1997), the erosion of product standards could follow this activity and reduce consumer protection. In this vein, the recent effort in implementing international trade and investment partnerships across the world is of particular importance. Free trade agreements such as the Transatlantic Trade and Investment Partnership promise strong positive welfare implications through the repeal of non-tariff barriers and the harmonization of standards that act as barriers to trade. However, in particular the harmonization of quality standards resulted in public uprising because people feared that the new product standards could not meet previous national consumer protection preferences.

In the EU, a new era of European competition began with the *Cassis-de-Dijon* (Case C-120/78, 20 February 1979) decision (Sinn 2003). The ECJ held that a regulation applying import restrictions to a product legally produced in one of the EU member states was an unlawful restriction on the free movement of goods. The goal of this judgment was to prevent protectionism of national governments. However, if consumers are unable to distinguish different national quality standards of a product, this judgment could result in Europe settling at an equilibrium with inefficiently low product standards (Sinn 1997). Thus, considering that regulatory competition is not efficient, Sinn (2003) contends that more centralized actions in the EU should be considered. In this sense, a central supervisory authority that supervises product quality from a consumer protection perspective would be helpful – as in the case of the US Food and Drug Administration.

Summary and Outlook

Regulatory competition in the spirit of the Tiebout (1965) model today is a widespread phenomenon on the regional, national, and global level. Lawmakers permanently attempt to direct individuals, companies, and capital to their home jurisdictions by changing their applicable law and infrastructure vis-à-vis those in other countries. While this process often improves the conditions for legal entities to conduct business and other operations, in certain situations it might also lead to a race to laxity. Regulatory harmonization might then be the policy tool of choice and is, for example, argued to foster financial stability in capital markets. Yet, the so-called Brexit vote, in which the citizens of Great Britain recently decided to leave the EU, can be interpreted as an attempt to flee European harmonization and become more flexible to engage in regulatory competition again. The UK prime minister, for example, already endorsed a move by the previous conservative government to reduce the corporate tax rate, not least to attract foreign companies. However, the introduction of regulatory uncertainty in other areas of the law could backfire, especially if legal entities prefer regulatory certainty to more benevolent legal rules. Thus, future research might investigate the trade-off between the benefits of harmonization and regulatory competition, how nation-states and supranational organizations should optimally position themselves in a multi-level setting such as the EU, and which welfare effects the various actors can expect.

With regard to regulatory competition in the form of unbundling of legal rules, the debate among US scholars largely ended in the 1990s, with some scholars proclaiming that the regulatory race is a “race to nowhere in particular” (Bratton 1994, p. 401). As such, regulatory competition at the US state level has frequently led to desirable results but has also been shown to have mixed effects for shareholders such as in the case of takeover legislation (Romano 1992). In Europe, the debate is not entirely solved yet.

Although we attempted to draw a comprehensive picture of the regulatory competition literature, many areas could not be covered because of

the limited scope of this chapter. Researchers have *inter alia* also investigated regulatory competition in the domain of marriage law, the legal process and arbitration procedures (O'Hara and Ribstein 2009), as well as competition law (Gabor 2013).

References

- Adler B, Polak B, Schwartz A (2000) Regulating consumer bankruptcy: a theoretical inquiry. *J Leg Stud* 29:585–613
- Akerlof GA (1970) The market for “lemons”: quality uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Bebchuk L, Cohen A (2003) Firms' decisions where to incorporate. *J Law Econ* 46:383–425
- Bebchuk L, Hamdani A (2002) Optimal defaults for corporate law evolution. *Northwest Univ Law Rev* 96: 489–519
- Becht M, Mayer C, Wagner H (2008) Where do firms incorporate? Deregulation and the cost of entry. *J Corp Finan* 14:241–256
- Bratton W (1994) Corporate law's race to nowhere in particular: the genius of American corporate law by Roberta Romano. *Univ Tor Law J* 44:401–440
- Braun R, Eidenmüller H, Engert A, Hornuf L (2013) Does charter competition foster entrepreneurship? A difference-in-difference approach to European company law reforms. *J Common Mark Stud* 51:399–415
- Bullock J (2013) Delaware division of corporations, 2013 annual report. http://corp.delaware.gov/Corporations_2013%20Annual%20Report.pdf
- Cary W (1974) Federalism and corporate law: reflections upon Delaware. *Yale Law J* 83:663–705
- Chemmanur TJ, Fulghieri P (2006) Competition and cooperation among exchanges: a theory of cross-listing and endogenous listing standards. *J Financ Econ* 82(2):455–489
- Coffee JC Jr (2002) Racing towards the top? The impact of cross-listings and stock market competition on international corporate governance. *C Law Rev* 102: 1757–1831
- Doidge C, Karolyi GA, Stulz RM (2004) Why are foreign firms listed in the US worth more? *J Financ Econ* 71(2):205–238
- Drometer M, Oesingmann K (2015) Household debt and the importance of effective private insolvency laws. *CESifo DICE Rep* 2:63–66
- Eidenmüller H (2011) The transnational law market, regulatory competition, and transnational corporations. *Indiana J Glob Stud* 18:707–749
- Eidenmüller H, Engert A, Hornuf L (2009) Incorporating under European law: the *Societas Europaea* as a vehicle for legal arbitrage. *Eur Bus Org Law Rev* 10:1–33
- Engel KH (1996) State environmental standard-setting: is there a race and is it to the bottom? *Hastings Law J* 48:271–398
- Enriques L, Gelter M (2007) How the old world encountered the new one: regulatory competition and cooperation in European corporate and bankruptcy law. *Tulane Law Rev* 81:577–646
- Fischel D (1982) The “race to the bottom” revisited: reflections on recent developments in Delaware's corporation law. *Northwest Univ Law Rev* 76:913–945
- Fleischer V (2010) Regulatory arbitrage. *Texas Law Rev* 89:227–272
- Gabor B (2013) Regulatory competition in the internal market: comparing models for corporate law. *Securities law and competition law*. Edward Elgar, Cheltenham
- Heine K, Kerber W (2002) European corporate laws, regulatory competition and path dependence. *Eur J Law Econ* 13:47–60
- Heyvaert V (2013) Regulatory competition: accounting for the transnational dimension of environmental regulation. *J Environ Law* 25(1):1–31
- Holzinger K, Sommerer T (2011) “Race to the bottom” or “race to Brussels”? Environmental competition in Europe. *J Common Mark Stud* 49(2):315–339
- Hornuf L, Lindner J (2016) The end of regulatory competition in European corporate law? In: *Wettbewerb der Gesellschaftsrechtsordnungen in Ostmitteleuropa?*. Nomos, Baden-Baden, pp 117–138
- Kim O, Pinnuck M (2014) Competition among exchanges through simplified disclosure requirements: evidence from the American and global depository receipts. *Account Bus Res* 44(1):1–40
- Konisky DM (2007) Regulatory competition and environmental enforcement: is there a race to the bottom? *Am J Polit Sci* 51(4):853–872
- LoPucki LM (2006) Courting failure: how competition for big cases is corrupting the bankruptcy courts. University of Michigan Press, Michigan
- LoPucki LM, Kalin SD (2001) The failure of public company bankruptcies in Delaware and New York: empirical evidence of a “race to the bottom”. *Vanderbilt Law Rev* 54:231–282
- O'Hara E, Ribstein L (2009) *The law market*. Oxford University Press, Oxford
- Oates WE, Portney PR (2003) The political economy of environmental policy. *Handb Environ Econ* 1:325–354
- Prentice R (2002) Whither securities regulation? Some behavioral observations regarding proposals for its future. *Duke Law J* 51:1397–1511
- Prentice RA (2005) Regulatory competition in securities law: a dream (that should be) deferred. *Ohio State Law J* 66:1155–1230
- Ringe WG (2016) Regulatory competition in global financial markets: the case for a special resolution regime. *Annals Corp Gov* 1(3):175–247
- Romano R (1987) The state competition debate in corporate law. *Cardozo Law Rev* 8:709–757
- Romano R (1992) A guide to takeovers: theory, evidence, and regulation. *Yale Regul* 9:119–179
- Sinn HW (1997) The selection principle and market failure in systems competition. *J Public Econ* 66:247–274
- Sinn HW (2003) *The new systems competition*, Yrjö-Jahnsson lectures. Wiley-Blackwell, Oxford

- Skeel DA (1998) Bankruptcy judges and bankruptcy venue: some thoughts on Delaware. *Del Law Rev* 1:1–45
- Skeel DA (2000) Lockups and Delaware venue in corporate law and bankruptcy. *Univ Cincinnati Law Rev* 68:1243–1279
- Tiebout C (1965) A pure theory of local expenditures. *J Polit Econ* 64:416–424
- Vogel D (1997) Trading up and governing across: transnational governance and environmental protection. *J Eur Publ Policy* 4(4):556–571
- Winter R (1977) State law, shareholder protection and the theory of the corporation. *J Leg Stud* 6:251–292

Regulatory Impact Analysis Meets Economic Analysis of Law: Differences and Commonalities

Wolfgang Weigel

Faculty of Business, Economics and Statistics,
Department of Economics, Joseph von
Sonnenfels Center for the Study of Public Law
and Economics, Vienna, Austria

Definition

Regulatory Impact Assessment is a tool for law-making, which makes use of economic criteria and techniques, such as Pareto-efficiency and cost-benefit analysis. It has been developed from the mid-seventies of the twentieth century and until now is in use worldwide, mainly pushed by OECD and the European Union. RIA reflects pragmatism.

The economic analysis of law, although having early roots in Europe at the age of enlightenment, has become a worldwide renown scientific approach and a vast field of interdisciplinary research at the intersection of economics and the law. It meets highest standards of sophisticated research and at the same time serves as methodology for policy recommendations.

Wolfgang Weigel has retired.

For decades the two fields developed without explicit reference to each other. So the question emerges, what they have in common and what are the differences. This is shown and complementary use is advocated.

What's Up?

In this entry it is shown what “Regulatory Impact Assessment” has in common with the economic approach to the law or “law and economics.” Intuitively, “regulation” and “laws” have a common scope. Moreover, they can be treated with the same tool: economics. But what makes them differ? For one it is the history of origins; this is remarkable inasmuch most of the time, there are no mutual references! Secondly it is in the purpose and thirdly in the range. While laws form the molecules of the judicial framework of every society, regulation is seen to form an essential part of economic policy making. However, in the influential view of the Organization for Economic Cooperation and Development (OECD) regulatory impact assessment – RIA for short – takes a much wider stance than the law inasmuch it is a tool, which meanwhile even by scholars of the doctrine of law-making (*Gesetzgebungslehre*) is acknowledged as the most comprehensive among the tools for the accomplishment of their task (Schäffer 2007)! But how does this fit into the observation that with respect to economics RIA seems essentially to rest on cost-benefit-analysis and little else, while the economic analysis of law makes use of the full range of microeconomic theory. Still there is a bracket around these which is welfare economics and a more comprehensive approach named new institutional economics, to be outlined later.

In order to elaborate on the commonalities and differences of RIA and Law and Economics, in the entry I proceed as follows: section “[The Notion of “Regulation” and the Focus of “Regulatory Impact Assessment”](#)” deals with the notions of “regulation” as well as “regulatory impact assessment” (RIA for short), a brief history of RIA and distinct properties. It follows a refresher of the approach of the economic analysis

of law (EAL) in section “[A Bird’s Eye View of the Economic Analysis of Law](#),” the core- chapter on the comparison (section “[The Overlap](#)”), followed by a brief outlook (section “[Outlook](#)”).

The Notion of “Regulation” and the Focus of “Regulatory Impact Assessment”

Regulation

It is important to note that the term “regulation” here is not meant in the narrow sense of imposing government constraints on private undertakings alone!

In fact, regulation is understood as a particular kind of incentive mechanism, namely, a set of incentives established either by the legislature, government, or public administration that mandates or prohibits actions of citizens and enterprises. Regulations are supported by the explicit threat of punishment for noncompliance (ideas on which scholars of EAL definitely will perfectly agree).

Finally, regulation here includes the full range of legal instruments and decisions – constitutions, parliamentary laws, subordinate legislation, decrees, orders, norms, licenses, plans, codes, and often even “grey” regulations such as guidance and instructions. Note that this view of regulation contrasts to the usual and somewhat narrower understanding of regulation as, for example, in Ogus (2004).

And when President Barack Obama made Cass Sunstein of Harvard head of the Office of Information and Regulatory Affairs (OIRA), even “nudging” became part of the regulatory arsenal. Here, nudging refers to means, by which people are gently induced to make appropriate decisions or act in a specific way. No explicit constraints are used and no particular enforcement mechanisms.

The instrumental use of regulation is closely tied to “market failures.” From existing manuals by OECD and from instructions provided through several US-presidents’ executive orders, the following list emerges:

- Monopolies and natural monopolies
- Information Inadequacies

- Continuity and availability of services
- Anticompetitive behavior and predatory pricing
- Moral hazard in the presence of public goods
- Unequal bargaining power
- Scarcity and rationing
- Distributional justice and social policy
- Rationalization and coordination
- Planning
- Myopic behavior

One important observation regarding the use of regulation must be stressed here: Public policies can be classified as “resource” – intensive and as “regulation” – intensive. To illustrate: A government may intend to cut carbon-dioxide-emissions by the reduction of exhaust-fume of cars. This can readily be pursued by an order, stating that car-owners are obliged to prove installation of the most advanced catalytic converter in the car by the end of the year. This would be at the cost of the owners and just leave enforcement costs with the government. However, we could imagine that government wants to make sure that the task is accomplished with minimum delay and allocate a certain amount out of the budget for the installation of the devices. This truly would be a resource-intensive undertaking. So with respect to public policy, spending and regulation are substitutes here!

However, since spending programs by the rule of law must rest on a law, the view put forward above requires some “metatheory of regulation,” which conveys the rationale for an initial framework of some order. As will be pointed out a little later, “new institutional economics” could serve as such metatheory.

Having investigated the notion of regulation, it is now time to turn to.

Regulatory Impact Assessment

A RIA is simply a way of gathering and organizing information about the expected impacts of a law or regulation and its major feasible alternatives. (Morrall 1994, p. 3, see also this Encyclopedia, Lanneau, R., “Regulatory Impact Analysis”)

“The purpose of RIA is to improve the quality of government interventions. It operates on familiar

principles and seeks first to ensure that the impacts both intended and unintended of proposed legislation and regulations are assessed in advance, and form an input into decision-making. RIA begins by answering the questions: Will the proposed intervention actually cause welfare to increase? What are the economic effects, i.e., how do the benefits stack up against the costs?" "Next, RIA highlights the strictly redistributive impact of proposed government intervention and establishes precisely who wins, who pays and how much" (Quotations from manuals on RIA by OECD 1997a).

Basically, a RIA has two underpinnings, one being an almost exhaustive checklist for regulatory undertakings and the second a rigorous cost-benefit analysis. The gist of the checklist is the following (Council of the OECD 1995):

1. Is the problem correctly defined?
2. Is government action justified?
3. Is regulation the best form of government action?
4. Is there a legal basis for regulation?
5. What is the appropriate level (or levels) of government for this action?
6. Do the benefits of regulation justify the costs?
7. Is the distribution of effects across society transparent?
8. Is the regulation clear, consistent, comprehensible, and accessible to users?
9. Have all interested parties had the opportunity to present their views?
10. How will compliance be achieved?

This list obviously requires typical considerations rooted in economic analysis, such as looking for alternatives and their consideration, as well as cost benefit analysis (see this Encyclopedia, Torriti, J., "► [Cost-Benefit Analysis](#)"). Moreover, it stresses procedural rules (cf. point 9) as well as typical issues of compliance and enforcement, respectively (cf. point 10) (which are, almost needless to say, considerations most familiar to the scholar of the economic analysis of law).

Ultimately, each of these questions requires treatment of the following steps: An analysis of the status quo as well as the need of intervention, an analysis of alternative ways to come to grips with a problem, including consultations and the collection of information. From this the

most preferred action should emerge, depicted in adequate measures, which then allow drafting (Renda 2010).

The generality of the procedures conveys a hint as to why in the doctrine of law-making RIA has such a prominent stance.

From First Oil-Crisis to Smart Regulation: A Very Brief Look on the History

After the oil-price-shock of 1974, the Ford-administration initiated measures to stabilize the economy with primary concern about inflation and employment. To this end, all regulations had to be checked for their effectiveness and subsequently reconciled with the "office of management and budget." So in the very beginning it was a device to check macropolicies. However, it were the frictions in the implementation of the policies, which were the primary focus of said checks! The respective attempts appear to have been quite promising. Consequently the first executive order regarding the systematic use of RIA was enacted by President Reagan 1981, (Order 12291). There one could read: "Regulatory action should not be undertaken unless the potential benefits to society for the regulation outweigh the potential costs to society ... Regulatory objectives shall be chosen to maximize the net benefits to society."

While RIA enjoyed changing popularity in subsequent years, it received stimulation from a study by Hopkins (1996), who shows that the aggregate compliance cost from federal regulation borne by the private sector and territorial authorities amounted to US \$ 668 billion or approximately 10% of GDP in 1995. This was, when the struggle against frictions in the economy due to flaws in regulation started, aptly illustrated by the upswing of the "office of information and regulatory affairs," founded by President Clinton through executive order 12866, which quickly reached 40 permanent staff.

Whereas several countries outside the USA fairly early adopted RIA, the boost came from an initiative of (OECD 1997b) regarding "transition economies," culminating in the program SIGMA (Support for Improvement in Governance and Management in Central and Eastern European

Countries) and PUMA (Public Management Service). Via OECD it entered the EU, where it first was adopted under the label “less and better regulation” which has muted to “smart regulation” meanwhile, being widely used, as can be seen from an ever increasing number of studies on <http://ria-studies.net/en> (cf Renda 2010).

Where Hope Lies There Is Also Disappointment

RIA for sure corresponds to the claim of a comprehensive approach. It meets several demands to a tool of law-making, the most prominent being lucidity and traceability, flexibility, and consistency. Moreover, it meets technical standards of lawmaking such as terminology and the allocation of competences, but also standards of implementation, such as practicality, feasibility, enforceability, but also acceptance.

The most prominent features are however those of economic efficiency as well as effectiveness, where the former is enhanced by cost-benefit analysis as the economic core of the tool. In a nutshell, the basic structure of every CBA is to secure that $B(x) - C(x) \geq 0$, where B are the (aggregated social) benefits (appropriately measured) of an act such as lowering/increasing the degree of regulation to some desired level x and C are the (social) costs. In the case of a continuous variation of x the criterion requires that $\partial B/\partial x = \partial C/\partial x$, constraints and contingencies notwithstanding.

The benefits B , accruing from some project (level of regulation) x , are captured by the (consumer) surplus brought about at prevailing (imputed) price of some regulations.

The basic structure of CBA is structurally equivalent to the Kaldor–Hicks test, which means that ostensibly there is a rigorous criterion for the selection of solutions to a regulatory agenda. However, one has to be cautious here for at least two reasons.

Leaving the tool itself unquestioned, the Kaldor–Hicks test itself has a serious weakness in its distributional ambiguity: To illustrate, assume that the benefits from some environmental measures go to a densely populated area, while the (opportunity) costs are borne by taxpayers. Then

the aggregated individual benefits may be equally distributed among the beneficiaries. But assume now that the beneficiary is a single entrepreneur owning the large parcel of land, which will increase in value to the level B . Such effect makes the beneficiary in a sense privileged, although the criterion for potential compensation still holds.

But there is a still harder objection: CBA carried out *lege artis* follows a couple of strict assumptions rooted in neoclassical microeconomics. And it is held that in many cases these could – or should – not be applied, which can easily be seen from the increasing number of deviations from neoclassical results brought about by “behavioral” and “experimental” economics, respectively (Camerer et al. 2011)

It is worth mentioning here that in the EU a “better regulation package” was released in May 2015, which partly took account of the methodological weaknesses just pointed out.

With respect to the relatedness to the “economic analysis of law,” the following features of RIA should be recalled:

RIA is definitely rooted in Paretian welfare economics (see the citation from President Reagan’s executive order, *supra*)

Closely related to the latter is “methodological individualism”

RIA takes a weak “consequentialist” view. This can be inferred from the OECD-checklist (*supra*), which implicitly at least asks for ethical or moral justifications. Moreover, the questions regarding distributive issues as well as the comprehension of affected groups underline this view inasmuch they point to certain underpinnings regarding standard perceptions of a society.

Before some commonalities and differences are pointed out, a view on the “economic analysis of law” (EAL) is due.

A Bird’s Eye View of the Economic Analysis of Law

Given the broad understanding of RIA in the view of OECD (which is shared by the EU), one may ask, if in law and economics the notion of a “law” might have a different meaning here than in the

concept of RIA. Well, upon inspection of some definitions of a law, one does not get a clear idea. Here are just two tasters:

“A rule of conduct or procedure established by custom, agreement, or authority” or “A statute, ordinance, or other rule enacted by a legislature.”

Both give rise to similar implications: it is not the “nature” of the legal rule which is at stake, but the way of formation through a distinct procedure.

This is, why the law eventually is contrasted to “norms” and “standards,” as, for instance, in a well known article by Michael Adams (2002). The latter are seen as outcomes of voluntary agreements (within an industry, say) or collective action of some kind, as in the exceptional book on “social institutions” by Schotter (1981).

This notwithstanding, the legislature might have some discretion in choosing the rule, which is established, provided it is backed by the superordinate legislature.

Unfortunately, the domain of the economic analysis of law appears to be slightly broader than that (in the meantime). Recall that sometimes – and especially with respect to international law – it is referred to “soft law” (Gersen and Posner 2008).

So summarizing “law” and “regulation” do overlap to a considerable extent with respect to the domain.

What about the scope of analysis?

In contrast, look at this: “Economic analysis of law seeks to answer two basic questions about legal rules. Namely, what are the effects of legal rules on the behavior of relevant actors? And are these effects of legal rules socially desirable?” (Kaplow and Shavell 2002, p. 1661), and one must add: how can they be improved so as to meet the measures of social efficiency?

Since the law guides the legislature, the executive branch, as well as the judiciary, the domain of the approach is all-encompassing.

In pursuing its purpose, the approach follows

- Paretian welfare economics
- In particular a methodological individualism
- Most of the time assuming rationally deciding people (“homo oeconomicus”)
- And a consequentialist view.

However, the quickly emerging psychology-prone contributions appear to lead to an erosion of both, the rationality-assumption and consequentialism, since the patterns guiding judgement as well as decision are more and more coming to the fore.

The Overlap

Before looking at the commonalities, one remarkable detail in the history of RIA is worth stressing: Looking at the publications dealing with RIA and as mentioned in the introduction most of the time, there is no reference whatsoever to the economic approach to law (and vice-versa!). To assure oneself, one might consult the OECD’s website on this issue, but this would be in vain! One thus gets the impression that the approach neglects the many valuable insights from an economic analysis of law regarding incentives of people in any social role, consumers, entrepreneurs, politicians, bureaucrats, judges, mothers, thieves, whatever, despite the fact that regulation is considered an “incentive mechanism” (supra, at section “[Regulation](#)”).

However, while after first inspection, the main tool of RIA is identical with the main tool of EAL, evidently there are considerable differences which come to mind immediately. The most important of these differences seems to be that in the course of a RIA typically no analysis is undertaken, of how a single representative individual would react to a norm; that is to say, how this person would calculate expected benefits and expected cost, so as to make the inherent incentives and likely transaction costs (broadly defined) of some regulation more visible. Instead, an aggregate measure is sought right away.

Now it is beyond the scope of this note to treat the steps of a RIA in detail. However, some of the requirements, as laid down in several guidelines as well as the Clinton Directive (Executive order 12866), deserve mention:

One such requirement is the “statement of need,” which is supposed to justify government

intervention (a list of reasons for regulation has been given above).

Moreover, the possibility of “over-regulation” due to preceding rent-seeking must be stressed here, which, in terms of economic analysis of law, can be seen as a treatment of malfunctions of a law, thus fitting into the list of purposes according to Kaplow and Shavell (*supra*).

But from time to time RIA appears to be even a kind of extension of the domain of the economic analysis of law (EAL). The importance and the relevance of EAL lies in the general and robust theoretical results, e.g., regarding incentives of all kinds (e.g., regarding liability rules vs. property rights or different kinds of punishment and so on). In contrast and taking a normative stance, the aforementioned directives point out that different standards of safety and quality of commodities are legitimate according to specific demands; even the legitimacy of barriers to entry for distinct professions is stressed, one example being pilots.

Consequently, what scholars of EAL would discuss under the label of “alienability” of property rights is addressed. Property rights here are understood as “sanctioned behavioural relations among men that arise from the existence of goods and pertain to their use” (Furubotn and Pejovich 1974, p. 3).

From such considerations to the emphasis to analyze alternative structures of property rights (institutions), it is just a fairly small step. To illustrate: the Clinton directive requires to take into account lawsuits and taxation as alternatives to regulatory measures.

One advantage of EAL over RIA definitely is the explicit consideration of transaction costs (on the notion: this Encyclopedia, Vereeck, L. et al., Transaction Costs). Recall that these costs are seen as the obstacles to the internalization of externalities. Therefore, they cannot be grasped as indirect cost or stemming cost within the typical classification of a cost-benefit analysis. They are, on the contrary, instrumental to such costs: To illustrate: Where clarification of property rights will not work, transaction costs suggest clear guidelines, which then allow for bargaining in the shadow of the law.

Moreover, RIA seems to abstract entirely from the role of courts, an issue, on which EAL has a lot to say. Bearing in mind that RIA is embedded in the broader concept of regulatory reform, which in turn is summarized by the guidelines stated in section “[Regulatory Impact Assessment](#)” above. Upon inspection, one learns that in some respect, these guidelines advocate a partial approach: while it is recommended to ask, whether regulation was the appropriate form of government intervention, it is not explicitly recommended to consider alternative institutional settings (as first suggested by Ronald Coase in his seminal 1960-article).

Now it is definitely true that EAL might be less conclusive on the aggregate level than RIA: This might follow from a typical way of reasoning within the law and economics framework: Frequently, the generalized consequences for resource allocation emerging from the way in which a judge handled a particular case are at stake. The famous “Learned Hand” formula may serve as an illustrative example (forming part of the set of robust results pointed out above). Here, the problem at hand – the liability of the “cheapest cost avoider” – is worked off by induction, that is to say by the generalization of considerations in the course of handling the case *United States v. Carroll Towing Co* (159 F.2d 169 [2d Cir. 1947], excellent description of the case in Cooter and Ulen 1988, p. 360 *passim*). RIA, with its flavor of pragmatism, takes a different approach. However, given the broad definition and the subtle claim of the more recent development of RIA (see section “[Regulatory Impact Assessment](#),” *supra*), one wonders why there is so little mentioning of, e.g., incentives and the presence of obstacles to action (i.e., transaction costs).

Maybe that the evasion of such subtleties is the price one has to pay for the explicit intention of OECD to provide a tool for a standardized treatment of a very broad range of issues.

Such observations open out into the question, whether RIA and EAL are not so much two fields with a common overlap, but rather closely related branches of a common approach, which might be labeled “new institutional economics” (see this Encyclopedia, Randa, R., “► [Institutional](#)

Economics”). As an underpinning of the institutionalist view of the economy, one may recall George Stigler here. In his Presidential Address at the occasion of the 1964(!) meeting of the American Economic Association he stated: The competence of economists “...consists in understanding how an economic system works under alternative institutional frameworks,” and one is tempted to say: consequently: “The basic role of the scientist in public policy, therefore, is that of establishing the costs and benefits of alternative institutional arrangements.”

Although it is not an easy task to give a definition of such a broad term as “institution,” it ought to be outlined at least. An Institution is a system of formal and/or informal rules, which by consent guides human actions in a particular way and contains remedies for enforcement. Furubotn and Richter (1997) offer a comprehensive view on this approach. So very briefly, what is (new) institutionalism about? Economics traditionally deals with decisions over scarce resources. Thus, the constraints caused by scarcity are at the core of economic research. But there is a second set of constraints, which deserves attention: these are norms of all kinds, which govern human behavior in society. Such norms can take the form of laws, but – as has already be stated – they also comprise informal rules, customs, and traditions. They can be the result of explicit collective decision making, but they may also emerge tacitly. They are subject to change or they may be violated. The by now well established New Institutional Economics concentrates on the analysis of the emergence and change of such norms, their effect on human conduct as well as conditions for and consequences of violation. The work on both RIA and EAL may be comprised as subsets of that broader approach, which serves as the “meta-theory” here for reasoning about the emergence of norms in general.

Outlook

At first sight RIA has a strong flavor of pragmatism, much stronger than the more academically prone Economic Analysis of Law: the image

of EAL meanwhile might be to be primarily an academic approach, a playground for quibblers, whereas RIA, despite the efforts to broaden its scope and to deepen the theoretical underpinnings, has conquered the marketplace for ideas as a tool designed to be applied for the practitioner. But EAL has not been developed just as a toy for contemplation of theoretical curiosity. The contrary is true: In the USA at least, it grew out of a down-to-earth request for reform (compulsory accident insurance for car-owners). And its advocates in Europe and other parts of the world always had the aspiration of reforming or at least complementing the orthodox approach to the law by a strictly analytical approach. In some sense, RIA does much the same as EAL does in some distinct areas of application. But the approach of RIA is pushed by the EU, OECD, and various member-states and has been strongly advocated and applied earlier in the USA, whereas EAL does not enjoy such almost continuous promotion; rather it is there to be discovered.

In this contribution the commonalities and differences in the two approaches have been investigated and the conclusion is that RIA and EAL not only do meet, meaning that they have much in common. Definitely they cannot merge. Rather they can be grasped as parts of a broader approach: that of new institutional economics. There they should meet, for the benefit of our society.

Cross-References

- [Cost–Benefit Analysis](#)
- [Institutional Economics](#)
- [Transaction Costs](#)

References

- Adams M (2002) Normen, Standards, Rechte (Norms, standards, rights). In: Adams (ed) *Ökonomische Theorie des Rechts* (Economic theory of the law). Lang publishers, Frankfurt am Main, pp 25–60
- Camerer CF, Loewenstein G, Rabin M (2011) *Advances in behavioral economics*. Princeton University Press, Princeton

- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cooter R, Ulen T (1988) *Law and economics*. Harper Collins, Boston, Mass
- Furubotn EG, Pejovich S (1974) Introduction. In: *The economics of property rights*. Cambridge University Press, Cambridge, MA, pp 1–10
- Furubotn E, Richter R (1997) *Institutions and economic theory: an introduction to and an assessment of the new institutional economics*. University of Michigan Press, Ann Arbor
- Gersen JE, Posner EA (2008) *Soft law*. Stanford Law Rev. University of Chicago, public law and legal theory working paper no. 213. Available at SSRN <https://ssrn.com/abstract=1113537>
- Hopkins TD (1996) Regulatory costs in profile. Policy study no 132. Center for the Study of American Business, St. Louis
- Kaplow L, Shavell S (2002) Economic analysis of law. In: Auerbach A, Feldstein M (eds) *Handbook of public economics*, vol 3. Elsevier, North Holland, pp 1661–1784
- Morrall J (1994) Assessing costs and economic effects. In: *Improving the quality of laws and regulations: economic, legal and managerial techniques*. OECD, Paris, pp 39–43
- OECD (1995) *Budgeting for results, perspectives on public expenditure management*. OECD, Paris
- OECD (1997a) *Assessing the impacts of proposed laws and regulations*. SIGMA paper no. 13. OECD, Paris
- OECD (1997b) *Law drafting and regulatory management in Central and Eastern Europe*. SIGMA paper no. 18. OECD, Paris
- Ogus A (2004) *Regulation, legal form and economic theory*. Oxford, Hart
- Renda A (2010) The development of RIA in the European Union: an overview. 20 Sept 2010. Available at SSRN <https://ssrn.com/abstract=1679764> or <https://doi.org/10.2139/ssrn.1679764>
- Schäffer H (ed) (2007) *Evaluierung der Gesetze/Gesetzesfolgenabschätzung II (Evaluation of law/impact assessment)*. Manz, Vienna
- Schotter A (1981) *The economic theory of social institutions*. Cambridge University Press, Cambridge, MA
- Morgenstern RD (ed) (1997) *Economic analyses at EPA. Resources for the Future Inc*, Washington, DC
- Nijssen A et al (2009) *Business regulation and public policy*. Springer, New York
- Weigel W (2008) *Economics of the law*. Routledge, London/New York

Regulatory Impact Assessment

Régis Lanneau

Law School, CRDP, Université de Paris Nanterre, Nanterre, France

Abstract

If regulation is often perceived as a tool to increase welfare, it is not infrequent that it leads to unintended consequences; examples are numerous. The purpose of regulatory impact assessment is precisely to help political decision-makers to identify whether and how to regulate while also defining a conception of the public interest that could then be debated. This entry will explain the logic and the development of RIA.

Introduction

If regulation is often perceived as a tool to increase welfare, it is not infrequent that it leads to unintended consequences; examples are numerous. After all, regulating in an evolving world of inherent complexity (and interdependence between economies) and uncertainty is not an easy task; it is the reason why it is bound to be a never-ending process. Nevertheless, it is not because perfect regulations are not from this world that it is not possible to develop tools to reduce the amount of inefficient regulations, especially when these regulations could be expected to impact many domains of a national or a regional economy. Regulatory impact assessments (RIA) are precisely trying to contribute to this aim by requiring regulators to assess the probable consequences of their regulation for the economy using models,

Websites

- http://ec.europa.eu/smart-regulation/impact/key_docs/key_docs_en.htm. 17 Aug 2017
- <http://ria-studies.net/en>. 17 Aug 2017
- <http://www.oecd.org/regreform/>. 17 Aug 2017
- https://ec.europa.eu/info/law/law-making-process/better-regulation-why-and-how_en. 17 Aug 2017

Recommended Reading

- Kirkpatrick C, Parker D (eds) (2007), *Regulatory impact assessment – towards better regulation?* Edward Elgar, Cheltenham

data, and evidence. These RIA are also supposed to help political decision-makers to identify whether and how to regulate while also defining a conception of the public interest that could then be debated. It is not then surprising that the practice of RIA is highly recommended by the OECD (1997) and part of its indicators to measure regulatory performance (see Radaelli 2004, for an explanation of the diffusion of RIA).

From the point of view of law and economics, RIA is both the recognition of its practical relevance and its inherent limitation for regulators and policy makers. Moreover, it helps to understand the real value of law and economics.

The Economic Logic Behind Regulatory Impact Assessment

In traditional law and economics, regulations are required in two situations: when the regulation can reduce transaction costs (defining the basic entitlements and the way to exchange them) or when it is supposed to reduce market failures (externalities, asymmetrical information or market power) (see Baldwin et al. (2010)). Nevertheless, as public choice emphasized, unproductive profit seeking by special interest groups to secure favorable regulation is definitional of the regulatory process (Tullock 1967), so there are no reasons why rational policy makers should implement only public interest regulation (Stigler 1971; Posner 1974; Peltzman 1976) without external constraints (public opinion's pressure or procedural requirement). Hence, regulations could merely be designed to redistribute wealth, and, when it does, these regulations should be considered as highly suspicious. This does not mean that, from an economic point of view, it is not possible to engage in policy choices but that these choices should be made sufficiently clear in their consequences to avoid the lurk of blindfold populism and moralism, an unavoidable risk in any democracy.

The mere need of RIA for enacting "better and smarter" regulations cannot be explained if it is possible to consider that regulations are only driven by the public interest and efficiency

requirement. RIA is thus a tool to overcome some regulatory failures. It is not a tool to make sure that regulations are efficient but a tool to ensure that they are not blatantly inefficient because it is forcing policy makers and regulators to justify their policy not only through a political rhetoric but through an explanation of the trade-offs they are engaging in and the requirement to quantify and identify the public interest based on the best data available (of course the quality of the data will be crucial, Torriti 2011).

Of course, since RIA are costly (administrative cost and delay are necessarily creating such costs), their necessity and precision are dependent on the regulation at stake (Rose-Ackerman 2016). It is only because RIA are supposed to lead to some social benefits (avoidance of bad measures, welfare gain from enacting what appeared to be efficient) that its costs could be justified. It is then to be anticipated that RIA requirement should be higher when the regulation at stake, either by its breadth or magnitude, is significantly impacting the economy. The fact that this requirement appeared in the field of environment and health should not then be a surprise (see also Arrow et al. 1996; Livermore and Revesz 2013).

Regulatory Impact Assessment as a Process

A. Three stages of RIA could be distinguished: description, identification, and quantification (see also European Commission 2009; for an overview of the methods and domains, see Dunlop and Radaelli 2016).

At first a RIA should explain what is the problem (identification and definition) the regulation is supposed to curb and why an action is required. At this stage, this is the need for an intervention that should be assessed. Using economic theory, this need should be either a market failure or a means to reduce transaction costs.

The second stage is inquiring into the regulatory options. Indeed, a same problem could be addressed using a variety of regulatory tools

from command and control to economic incentives to mandatory disclosure or nudges. It is not only the “specie” of regulatory option that should be considered but also the variety within the “specie.”

The third stage is dedicated to the quantification of the direct and indirect benefits and costs of each option (including the administrative and information regulatory burden; see also, more generally, Marneffe and Vereeck 2011; Hahn and Tetlock 2007). Of course, the benefits and costs can only be assessed through a set of hypotheses or choices regarding, more especially, the valuation methodology and baseline scenario. These hypotheses are crucial to assess the subjective value of the RIA for policy making (see, e.g., Ogus 1998; Rose-Ackerman 2011a; Hahn and Litan 2005). It will then be possible to compare options and eliminate obviously relatively inefficient options.

Being transparent about all these stages could contribute to a better accountability of regulators (since trade-offs are made clear) and to an increased consistency in regulations (since the need of a new regulation is explained). Moreover, with experience in RIA, better data and analytical methods should be expected to emerge. RIA is, from this point of view, a learning by doing a process.

B. RIA is not only a document (or a set of documents) providing information regarding the expected consequences of a regulation (an end result); it should be conceptualized much more as a process (Adler and Posner 2006). Indeed, as an “end result,” it could be considered as a mere new rhetoric to justify regulatory choices; the hypotheses would then be chosen *ex post* to justify the desired regulation which will then be submitted to some formal or informal political debate – it is quite often this instrumental use of RIA that is criticized (Renda 2006). Conceptualized as a process, RIA exhibits two interesting features. First, it makes clear that RIA is a way to inquire into a regulatory problem: neither the problem nor its options are facts; they are identified and assessed differently depending on the state of knowledge. RIA, once made, is then not an end result but a “first draft” that will evolve

with the comments and criticism of stakeholders. Second, RIA is not a stage after drafting a regulation; it should be developed in parallel of it to reduce the likelihood of an instrumental use. This also requires a complete transparency of the process.

The Rise of Regulatory Impact Assessment

The rise of RIA began in 1974, in the USA. President Ford’s executive order 11821 established procedures for preparing inflation impact statements that were designed to illuminate the economic impact of regulatory proposals, specifically their effects on productivity and competition. Nevertheless, it is at the beginning of the 1980s that it became the mainstream in the USA. Indeed, President Reagan’s executive order 12291 of 1981 states that “Regulatory action shall not be undertaken unless the potential benefits to society from the regulation outweigh the potential costs to society.” Another 20 years was required for RIA to spread among OECD countries. During the 1990s and before 1997 (year of the first OECD publication on the subject even if RIA were recommended by the council of OECD in 1995), less than 10 OECD countries had comprehensive RIA program, by the end of 2000, 14, and since 2008, more than 30. Nowadays, almost all OECD countries have such programs despite differences (among which the definition of the type of regulation and intervention submitted to RIA). For example, in France, only primary legislation prepared by the government is subjected to RIA to the exclusion of all secondary legislation.

There are very few studies analyzing the potential of RIA for developing countries (Kirkpatrick and Parker 2003); Korea and Mexico are often mentioned as good examples of RIA design. However, in most developing countries, when RIA are present, they are not systematically used, and the methodology developed remains largely incomplete (Kirkpatrick et al. 2003). Rodrigo (2005) highlights the political requirements allowing RIA to prosper.

Conclusion

Even if RIA are often targeted as promoting “conservative” regulations and “deregulation,” it is hard to deny the necessity of assessing the consequences of regulations. Once again, the purpose is not to make sure that only good regulations will be enacted or to avoid all negative unexpected consequences but to reduce the likelihood of bad regulations by clarifying available regulatory options. RIA should remain a tool to inform policy makers and regulators (see also Jasanoff 1990; Rose-Ackerman 2011b); it can be used adequately if and only if its limits are well identified (regarding model hypotheses and data quality and aggregation).

Cross-References

- [Conflict of Interest](#)
- [Cost–Benefit Analysis](#)
- [Efficiency](#)
- [Government](#)
- [Rent Seeking](#)

References

- Adler M, Posner E (2006) New foundations of cost-benefit analysis. Harvard University Press, Cambridge, MA
- Arrow KJ et al (1996) Benefit-cost analysis in environmental, health and safety regulation. AEI Press, Washington, DC
- Baldwin R, Cave M, Lodge M (2010) The Oxford handbook of regulation. Oxford University Press, Oxford
- Dunlop C, Radaelli C (Hrsg) (2016) Handbook of regulatory impact assessment. Edward Elgar Publishing, Cheltenham
- European Commission (2009) Impact assessment guidelines, 15 Jan 2009, SEC(2009) 92
- Hahn R, Litan R (2005) Counting regulatory benefits and costs: lessons for the US and Europe. *J Int Econ Law* 8:473–508
- Hahn R, Tetlock PC (2007) The evolving role of economic analysis in regulatory decision making. AEI Brookings, November
- Jasanoff S (1990) The fifth branch: science advisers as policymakers. Harvard University Press, Cambridge, MA
- Kirkpatrick C, Parker D (2003) Regulatory impact assessment: developing its potential for use in developing countries, working paper series, no. 56, Centre on Regulation and Competition, Manchester
- Kirkpatrick C, Parker D, Zhang Y (2003) Regulatory impact assessment in developing and transition economies: a survey of current practice and recommendations for further development. Centre on regulation and competition, Institute for Development Policy and Management, University of Manchester
- Livermore M, Revesz R (Hrsg) (2013) The globalization of cost-benefit analysis in environmental policy. Oxford University Press, Oxford
- Marneffe W, Vereeck L (2011) The meaning of regulatory costs. *Eur J Law Econ* 32(3):341–356
- OECD (1997) Regulatory impact analysis, best practices in OECD countries. OECD Publishing, Paris. Available online: <http://www.oecd.org/gov/regulatory-policy/35258828.pdf>
- Ogus A (1998) Regulatory appraisal: a neglected opportunity for law and economics. *Eur J Law Econ* 6(1): 53–68
- Peltzman S (1976) Toward a more general theory of regulation. *J Law Econ* 19(2):109–148
- Posner R (1974) The social costs of monopoly and regulation. *J Polit Econ* 83(4):807–828
- Radaelli C (2004) The diffusion of regulatory impact analysis in OECD countries: best practices or lesson-drawing? *Eur J Polit Res* 43:723–747
- Renda A (2006) Impact assessment in the EU: the state of the art and the art of the state. CEPS Paperbacks, Brussels
- Rodrigo D (2005) Regulatory impact analysis in OECD countries, challenges for developing countries. OECD working paper. Available online: <https://www.oecd.org/gov/regulatory-policy/35258511.pdf>
- Rose-Ackerman S (2011a) Putting cost-benefit analysis in its place: rethinking regulatory review. *Univ Miami Law Rev* 65(2):335–356
- Rose-Ackerman S (2011b) Étude d’impact et analyse coûts-avantages : qu’impliquent-elles pour l’élaboration des politiques publiques et les réformes législatives? [Impact assessment and cost-benefit analysis: what do they imply for policymaking and law reforms?]. *Rev Fr Adm Publique* 2011(4):787–806
- Rose-Ackerman S (2016) The limits of cost/benefit analysis when disasters loom. *Global Pol* 7(S1):56–66
- Stigler G (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 2(1):3–21
- Torriti J (2011) The unsustainable rationality of impact assessment. *Eur J Law Econ* 31(3):307–320
- Tullock G (1967) The welfare cost of tariffs, monopoly and theft. *West Econ J* 5(3):224–232

Related Issues Are Business Ethics

- [Codes of Conduct](#)

Rent Seeking

Christopher J. Coyne and Garrett Wood
Department of Economics, George Mason
University, Fairfax, VA, USA

Definition

This chapter explores the concept of rent seeking. When resources earn returns in excess of those they could earn in their next best opportunity, economists call those returns “rents.” In the market, the seeking of rents is known as profit seeking; in politics it is known as “rent seeking.” Profit seeking generates social value through the competitive market process. Rent seeking results in social losses as finite resources are used to pursue transfers of wealth rather than in the production of new wealth. This entry explores the distinction between rent seeking and profit seeking and discusses the implications.

Introduction

Economic rent accrues to owners when their resources earn returns over and above those resources’ opportunity cost. Rents can exist in private markets or in political settings. In private markets, rent seeking is referred to as “profit seeking” to denote that it generates a social surplus that is beneficial to society. In contrast, in political settings rent seeking generates social costs that are harmful to society. The next section explores the distinction between profit seeking and rent seeking in greater detail. We then discuss the problems with rent seeking for social welfare. The conclusion discusses the implications.

Profit Seeking Versus Rent Seeking

Since rents are nothing more than resources earning returns above their opportunity cost, it is important to make the distinction between circumstances in which pursuing rents is productive

versus circumstances in which it is unproductive. Let us begin with rent seeking in the context of competitive markets.

In markets, entrepreneurs seek new opportunities to earn profits by reallocating scarce resources to new and better uses. This process is guided by market prices and subjected to the profit and loss test, which provides feedback as to whether the entrepreneur has been successful in increasing social value (Kirzner 1973, 1992, 1997; Thomsen 1992). A first-moving innovator, for example, who brings a profitable product to market, will earn a rent. However, these are “quasi-rents,” a term intended to indicate that, in competitive markets, profits are temporary. The reason they are temporary is that in competitive markets, these quasi-rents will attract other entrepreneurs who will enter the market and erode the profit earned by the first mover. This process is socially beneficial because it leads to the efficient allocation of scarce resources. For this reason, rent seeking in the context of the competitive market is referred to as “profit seeking.”

In political settings, rent seeking has dramatically different welfare implications. The political production of rents is artificial in nature and depends on practices, such as the granting of monopoly rights which raise barriers to entry, which destroy social wealth. Government-granted monopoly rights transfer a portion of consumer surplus to the monopoly producer. Part of the cost are foregone trades which otherwise would have taken place, resulting in a deadweight loss. The transfer itself, however, creates a pool of resources to be won and attracts fierce competition to secure the right to earn those profits (Tullock 1967). The difference in this scenario from that of the market situation, however, is that the competition in politics is over securing permanent privileges that prevent the erosion of rents (Buchanan 1980). While profits in markets represent quasi-rents, the rents earned through political competition endure precisely because the competition is over government-created protections from market forces.

Posner (1975) posits a rent-seeking game where a monopoly right worth \$100,000 is bid for by ten risk-neutral rent seekers who cannot

recover their bids should they lose the game. Each bidder offers \$10,000 resulting in the complete dissipation of rents. The \$100,000 in bids spent in pursuit of the \$100,000 in rents constitutes a social loss to the extent that the lobbying process is spent on activities that do not increase total social wealth by creating new value. Instead, these expenditures are made to secure the transfer of existing wealth.

Tullock (1967) compared the losses from rent seeking to the losses from counterbalancing investments in tools for thievery and safeguards against theft. Resources spent on the tools of thievery and preventative measures against thieves do not increase the size of the prize being sought or protected. Neither do resources spent by competing lobbyists to secure a government privilege, resulting in social waste.

Further exacerbating these social losses is that investments in rent seeking are specific. Resources spent by one individual developing influence over the politicians creating rents are mostly lost because that influence is difficult or impossible to transfer to another individual if the bid for rents fails. The social cost of rent-seeking activities is measured by the opportunity cost of the resources used. Entrepreneurs who direct their efforts to curry favor with politicians for narrow privileges cannot simultaneously exert that effort to provide a superior good or service for private consumers.

Certain scenarios can reduce the amount of rent seeking below the level of complete dissipation modeled by Posner. Baysinger and Tollison (1980) introduce consumers and consumer groups as interest groups capable of organizing to resist the creation of rents. The efforts of these groups reduce the expected payout of rent-seeking activities and thus results in lower bids and fewer resources wasted. In their model, consumers will tend to organize where the expected success of their counter-lobbying efforts is higher than the combined expected costs of organization and rent creation. Tullock (2001) models a rent-seeking game in which players are in marginal balance, each bidding \$25 for a prize of artificial rents worth \$100. The result is inframarginal gains from winning the game. This result depends on

players knowing how to find the optimal bidding strategy and knowing that the other player will also find it, an assumption that becomes more reasonable in instances of repeated play between professional lobbyists.

The incentives of the political agent creating rents must also be taken into account. Peltzman (1976) introduces political actors who face a trade-off between the rents they create for interest groups and the votes of their constituent consumers. The politician attempts to weigh the benefits gained from lobbying by rent seekers against lost votes from those constituents who incur a sufficient cost and alter their vote accordingly. Politicians are therefore incentivized, though perhaps weakly depending on the specific context, to limit the amount of rents they create in order to preserve enough votes to get reelected.

The Problem of Rent Seeking

Rent seeking encourages the use of scarce resources in competition over pools of wealth transferred from one party to another instead of in a competitive process that produces new wealth. This is harmful to the well-being of the members of a society because scarce resources are used to secure existing wealth as compared to creating new value. While beneficial to the individuals who win the rent-seeking competition, the cumulative effect of rent seeking is the “decline of nations” as members of society become increasingly focused on securing transfers of existing wealth (Olson 1982).

The extent of the problem posed by the creation of rents was underestimated for many years (Tollison 1982). Harberger (1954) demonstrated empirically that the deadweight loss to society from monopolization in US manufacturing around 1929 was low compared to GNP, less than 1% in fact. Harberger’s calculations, however, did not account for the costs of rent seeking. Krueger (1974) found that rent-seeking activity in the Indian public sector accounted for 7.3% of GNP in 1964, while rent-seeking activity in Turkish import licensing accounted for 15% of GNP in 1968. Rama (1993) found that

rent-seeking activities in Uruguay led to temporary sectoral growth, but an overall decline in economic growth. Del Rosal (2011) offers a survey of empirical estimates of rent-seeking costs as a percentage of GNP and GDP, the amount of resources misallocated to unproductive professions, and other indicators. These estimates of the costs of rent seeking have to be added to the costs captured in the Harberger triangle to form more accurate estimates of the overall costs of monopoly.

Given that costs to society of rent seeking are often significant, it remains to be explained why the legislation that creates rents is so rarely repealed. Tollison (1982) highlights key features of the rent-seeking game to explain this phenomenon. In a world of zero transaction and information costs, no rent seeking would occur because the groups who incurred the dispersed costs of rent seeking would easily organize to protect themselves. In a world with positive transaction and information costs, however, some groups will find it easier to organize for lobbying activity. The groups that find it feasible to organize will seek transfers of wealth at the expense of groups who cannot easily organize to resist their efforts. Groups organized for lobbying often involve members of industry or labor, but rarely do they form around consumers in general due to the costs of organization. The concentrated transfers to be won from successful lobbying efforts, combined with the high cost of organizing the numerous, dispersed consumers who shoulder the burden of rent seeking, prevent society from realizing the potential gains of repealing rent-creating legislation.

Conclusion

Baumol (1990) argued that while policy decisions are limited in altering society's overall supply of entrepreneurs, they can heavily influence the allocation of entrepreneurial efforts across different types of activities. In societies where large profits can be made through competitive markets, entrepreneurs will allocate their resources to engaging in productive, value-added activities. In societies

where rents can be earned through political means, in contrast, entrepreneurs will invest in unproductive, rent-seeking activities. The overall balance between productive and unproductive entrepreneurial activities is a function of the rules of the game which influence the payoffs associated with each type of activity. This logic offers crucial insight into the process underlying the economic rise and decline of societies throughout time and across geographic space.

Cross-References

- ▶ [Coase, Ronald](#)
- ▶ [Economic Analysis of Law](#)
- ▶ [Economic Performance](#)
- ▶ [Entrepreneurship](#)
- ▶ [Externalities](#)
- ▶ [Gordon Tullock: A Maverick Scholar of Law and Economics](#)
- ▶ [Government](#)
- ▶ [Government Quality](#)
- ▶ [Law and Economics](#)
- ▶ [Market Definition](#)
- ▶ [Political Competition](#)
- ▶ [Political Corruption](#)
- ▶ [Political Economy](#)
- ▶ [Politicians](#)
- ▶ [Protective Factors](#)
- ▶ [Public Interest](#)
- ▶ [State Aids and Subsidies](#)
- ▶ [Transaction Costs](#)

References

- Baumol W (1990) Entrepreneurship: productive, unproductive, and destructive. *J Polit Econ* 98(5): 893–921
- Baysinger B, Tollison R (1980) Evaluating the social costs of monopoly and regulation. *Atl Econ J* 8(4):22–26
- Buchanan JM (1980) Rent seeking and profit seeking. In: Buchanan J, Tollison R, Tullock G (eds) *Toward a theory of the rent-seeking society*. Texas A&M University Press, College Park, pp 3–15
- Del Rosal I (2011) The empirical measurement of rent-seeking costs. *J Econ Surv* 25(2):298–325
- Harberger A (1954) Monopoly and resource allocation. *Am Econ Rev* 44(2):77–87

- Kirzner I (1973) Competition and entrepreneurship. University of Chicago Press, Chicago
- Kirzner I (1992) The meaning of the market process: essays in the development of modern Austrian economics. Routledge, New York
- Kirzner I (1997) Entrepreneurial discovery and the competitive market process: an Austrian approach. *J Econ Lit* 35(1):60–85
- Krueger A (1974) The political economy of the rent-seeking society. *Am Econ Rev* 64(3):291–303
- Olson M (1982) The rise and decline of nations: economic growth, stagflation, and social rigidities. Yale University Press, New Haven
- Peltzman S (1976) Toward a more general theory of regulation. *J Law Econ* 19(2):211–240
- Posner R (1975) The social costs of monopoly and regulation. *J Polit Econ* 83(4):807–827
- Rama M (1993) Rent seeking and economic growth: a theoretical model and some empirical evidence. *J Dev Econ* 42(1):35–50
- Tollison R (1982) Rent seeking: a survey. *Kyklos* 35(4):575–602
- Thomsen EF (1992) Price and knowledge: a market-process perspective. Routledge, New York
- Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. *West Econ J* 5(3):224–232
- Tullock G (2001) Efficient rent seeking. In: Lockhard A, Tullock G (eds) *Efficient rent-seeking: chronicle of an intellectual quagmire*. Kluwer Academic Publishers, Norwell, pp 3–16

Resale Right

► [Droit de Suite](#)

Retributivism

Mark D. White
Department of Philosophy, College of Staten
Island/CUNY, Staten Island, NY, USA

Definition

A theory of punishment that maintains that wrongdoers deserve to be punished, in proportion to their crimes, as a matter of justice or right.

Retributivism is a theory or philosophy of criminal punishment that maintains that wrongdoers deserve punishment as a matter of justice or

right. It is often contrasted with deterrence, which justifies punishment on the basis on the future harms it prevents. (On theories of punishment, see the papers in Acton (1969), Duff and Garland (1994), and Simmons et al. (1995); for concise syntheses, see Tunick (1992) and Brooks (2012).)

Drawing on the terminology of moral philosophy, retributivism is often characterized as *deontological* in nature, being based on qualitative concepts of justice and the right, while deterrence is *consequentialist*, focused on minimizing the negative outcomes from crime. This is, admittedly, an overgeneralization: for instance, Moore (1993), Cahill (2011), and Berman (2013) have suggested that retributivism can be regarded as a consequentialist or instrumentalist theory in which punishment or justice are goods to be maximized (see Dolinko (1997) for a counterargument). Nonetheless, it remains that retributivism is linked to justice rather than welfare or utility, and this leads to more specific and useful ways to distinguish it from alternative theories of punishment.

Before we can fully appreciate retributivism, we must briefly discuss deterrence. The formal basis for deterrence can be traced to classical utilitarianism, particularly the policy recommendations of Jeremy Bentham (1781) and the criminology of Cesare Beccaria (1764). As punishment by its nature is harmful, it is justified only insofar as it prevents greater harm: “All punishment in itself is evil. Upon the principle of utility, if it ought at all to be admitted, it ought only to be admitted in as far as it promises to exclude some greater evil” (Bentham 1781, p. 170). One way punishment “excludes some greater evil” is through deterring future crime by incapacitating the guilty and creating an incentive for others to abstain from criminal activity, both of which lessen the incidence of crime going forward.

Deterrence is not focused on lessening crime itself but rather on minimizing the overall costs of crime, including not only the harms done by criminals but also the costs of punishment, prosecution, and apprehension. This inclusive definition of the costs of crime implies that some crimes will be unworthy of punishment because it is “*groundless*”: where there is no mischief for it to prevent. . .

inefficacious: where it cannot act so as to prevent the mischief. . . *unprofitable*, or too *expensive*: where the mischief it would produce would be greater than what it prevented. . . [or] *needless*: where the mischief may be prevented, or cease of itself, without it: that is, at a cheaper rate” (Bentham 1781, p. 171). It is no surprise that this general framework of deterrence was adopted by Becker (1968) and other economists as they began to study crime and developed theories of efficient punishment that refined and expanded on the early insights of Bentham and Beccaria.

Retributivism is not concerned with the effects of punishment on future behavior or costs but instead solely addresses the crime for which the perpetrator is to be punished. For this reason, retributivism is often referred to as “backward-looking,” focusing on the crime which was committed in the past, while deterrence is characterized as “forward-looking,” taking the past crime as given and exacting a punishment designed to minimize the future costs of crime. While deterrence sees no inherent value in punishment itself, but instead sees it as a tool to maximize utility or minimize harm, retributivism maintains that the state has a moral, political, or legal duty to punish wrongdoers.

While embedded in discussions of justice going back to Plato and Aristotle, the modern roots of retributivism are usually traced back to Immanuel Kant, for whom retributivism stemmed from his belief in the respect owed to persons due to their inherent dignity as autonomous beings. While his position on punishment is more complex than often acknowledged (as pointed out by Murphy 1987), several statements he made exemplify the basic tenets of retributivism. For instance, Kant held that punishment must be imposed only in response to the crime committed and “can never be inflicted merely as means to promote some other good for the criminal himself or for civil society. It must always be inflicted upon him only because he has committed a crime. For a human being can never be treated merely as a means to the purposes of another” (1797, p. 331). Kant addressed the danger of punishing the innocent and the other goals of

punishment when he wrote a person “must previously have been found *punishable* before any thought can be given to drawing from his punishment something of use for himself or his fellow citizens” (1797, p. 331). Along with Hegel (1821), Morris (1968), and Murphy (1973), Kant emphasized that retributivism is grounded in respect for the individual, giving the guilty what they are owed by virtue of what they did, as a matter of impersonal justice rather than the base emotions of revenge or vengeance (a distinction explored by Nozick (1981, pp. 366–368)).

Retributivism can be classified as either negative or positive. *Negative retributivism* places limits on the subject and severity of punishment: the innocent should never be punished, and the guilty should not be punished more than is proportionate to their crimes. *Positive retributivism* goes further, endorsing the upper limits on punishment imposed by negative retribution while also maintaining that the guilty must be punished and not less than is proportionate to their crimes (thereby imposing a duty of overall proportionality). (This distinction is usually attributed to Mackie (1985) and has since become standard in the literature on punishment.)

Negative retributivism is often seen as a side constraint on the pursuit of deterrence that prevents it from considering the punishment of innocents or imposing disproportionately severe punishments on the guilty. Deterrence constrained by negative retributivism is a common *hybrid theory* of punishment as suggested most famously by Hart (1968); some, such as Byrd (1989), argue that this was Kant’s complete view as well. While the intentional punishment of the innocent may be more of a theoretical possibility than a practical threat in most modern democracies, the practice of disproportionately severe punishments is a logical implication of efficient deterrence. For instance, Becker (1968) showed that, in many cases, enforcement costs can be lowered while achieving the same level of deterrence if the probability of punishment is lowered – reducing enforcement costs – while the severity of punishment is raised, most likely above a level proportionate to the crime. (One example is littering, which in the

United States carries an extremely low chance of apprehension and posted fines much higher than any measure of the resulting harm.) While this may be efficient from a cost-minimization point of view, the negative retributivist would object that it subjects the guilty person to a greater punishment than he or she deserves in light of the crime committed.

Positive retributivism adds the imperative of punishing the guilty and ensuring that the punishment not be disproportionately mild; as Kant wrote, “woe to him who crawls through the windings of eudaemonism in order to discover something that releases the criminal from punishment or even reduces its amount by the advantage it promises” (1797, p. 331). Since negative retributivism does not mandate punishment at all, some doubt whether it counts as retributivism (Cottingham 1979). But positive retributivism is “complete,” mandating that all wrongdoers be punished and in proportion to their crimes, with the punishment neither too harsh or too light in comparison to the crime.

However, the stricter nature of positive retributivism presents a number of problems both theoretical and practical. For one, while it is easy to argue that the innocent must not be punished, it is more difficult to argue why the guilty *must* be punished (without recourse to preventing future harm). Perhaps the simplest justification – if we can even call it that – is to invoke the ancient code of the *lex talionis*, “an eye for an eye and a tooth for a tooth.” Kant cites this approvingly when he asks “what kind and what amount of punishment is it that public justice makes its principle and measure? None other than the principle of equality... Accordingly, whatever undeserved evil you inflict upon another within the people, that you inflict upon yourself. ... But only the *law of retribution (ius talionis)*. ... can specify definitely the quality and the quantity of punishment” (1797, p. 332). The *lex talionis* certainly prescribes a strict proportionality which would be impossible in most cases, and it is seen by most retributivists as simplistic, harsh, and inhumane; some doubt its relevance to retributivism at all (Davis 1986).

Most modern retributivists justify due punishment on other grounds. Intrinsic retributivists maintain that punishment of the guilty is an intrinsic good that does not derive its value from any more basic precept. (See Davis (1972) for one example, arguing against Honderich (1984), who coined the term and criticized the concept.) But others find a justification for retributivist punishment in a broader political theory, often based on a version of reciprocity that maintains that the guilty “owe a debt” to society or their fellow citizens and punishment restores this balance (famous from Plato’s *Crito* (360 BCE); for surveys of such theories, see Tunick (1992, ch. 3), and Duff (2001, ch. 1)). For example, many retributivists, such as Morris (1968), argue that while innocent citizens bear the costs of compliance with the law, the guilty “free ride” on these efforts, gaining unfair advantage that must be repaid to restore balance (Kant was sympathetic to this view as well, according to Murphy 1972). Similarly, some retributivists maintain that the purpose of punishment is, in the words of Hegel (1821, p. 69), “to annul the crime, which otherwise would have been held valid, and to restore the right.”

Finally, there are justifications for positive retributivism that are not based on reciprocity, such as denunciation or expression of condemnation (Feinberg 1965; von Hirsch 1985; Markel 2011) and moral education and communication (Hampton 1984; Duff 2001). Note that these justifications are goal oriented in nature, focused on an end result of punishment, albeit one that is more specific than deterrence or harm reduction and also tied closely to the crime committed. Nozick (1981, p. 371) calls such theories of punishment *teleological retributivism* because they focus on an end other than the suffering of the guilty; Berman (2013) argues that most modern retributivist theories are teleological or instrumental in nature.

Another problem facing positive retributivism (and even negative retributivism to a degree) is how to determine a type and degree of punishment that is proportionate to – or “fits” – a given crime. The *lex talionis* demands literal proportionality, the spirit of which survives today in jurisdictions

which prescribe the death penalty for murderers despite questionable deterrent effect (Ehrlich and Liu 2006), but is impractical or unthinkable in cases like battery, rape, or attempted murder. Even Kant, who supported the *lex talionis*, appreciated its limitations, asking “but what is to be done in the case of crimes that cannot be punished by a return for them because this would be either impossible or itself a punishable crime against *humanity* as such?” and recommending that “what is done to [the wrongdoer] in accordance with penal law is what he has perpetrated on others, if not in terms of its letter at least in terms of its spirit” (1797, p. 363).

While equivalence is not possible, it is widely agreed that, at the very least, more serious crimes should be punished more severely. But even this determination is not as easy as it may seem: while various degrees of one crime, such as murder or theft, can be ranked and punished appropriately, determining proportionate punishment for two different crimes is more difficult. Card (1975), von Hirsch (1976), and Davis (1983) are seminal works in this area, but some (such as Wertheimer 1975) remain skeptical that a consistent, proportional system of punishment is possible at all. More fundamentally, retributivists disagree on what determines the severity of a crime: whether the harm done or intended, the degree of wrongdoing, or the culpability of the criminal. (See Davis (1986) on harm versus wrong and Alexander et al. (2009) on culpability.) Given the controversy over what determines the severity of crimes themselves, the difficulty of ranking them in order to assign proportionate penalties becomes even more difficult.

A final problem with retributivist punishment on which economics is particularly well suited to comment is its resource cost. Economists and economics-minded legal scholars have paid very little attention to retributivism; as Posner (1980, p. 92) wrote, it is “widely viewed as immoral and irrational, at least as primitive and non-rational.” (For similar views, see also Kaplow and Shavell (2002) and Sunstein (2005); a notable exception to the antipathy among economists to retributivism is Wittman (1974)). At the same time, legal philosophers have not paid much

attention to the economic ramifications of retributivism, although Avio (1990, 1993), Cahill (2007), and White (2009) have started to look at this issue. The problem with implementing a retributivist system of punishment in the real world is that its perfectionist nature (“the guilty *must* be punished in proportion to their crimes”) does not easily allow for the compromises required in a world of scarcity. If governments tried to punish all wrongdoers according to their just deserts – including making every possible attempt to apprehend and prosecute them – the criminal justice would end up absorbing all of the society’s resources (White 2009; see Braithwaite and Pettit (1990) on system-wide considerations of retributivism). For this reason, some recommend a hybrid theory of punishment with negative retributivism as a limited factor (Avio 1993), a consequentialist retributivism that seeks to maximize some measure of justice (Cahill 2007, 2011), or a “pro tanto retributivism” that considers necessary compromises to result from a balancing of competing principles (White 2011a).

Retributivism remains a vital area of interest and research, with contemporary scholars attempting to refine its principles in the context of a constantly evolving political world (White 2011b) and exploring the implications of new developments in neuroscience and psychology (Nadelhoffer 2013). Although it does not sit well with the utilitarian foundations of the economic approach to law, economists should take note of the extent to which retributivism lies at the root of much of criminal law doctrine and how it steers the criminal justice system away from textbook models of efficiency, affecting resource allocation throughout the criminal justice system and society as a whole. This knowledge is also integral to the way that legal scholars and philosophers understand how retributivism works in the real world.

Cross-References

- [Crime and Punishment \(Becker 1968\)](#)
- [Criminal Sanctions and Deterrence](#)
- [Lex Talionis](#)

References

- Acton HB (ed) (1969) *The philosophy of punishment: a collection of papers*. St. Martin's Press, New York
- Alexander L, Ferzan KK, Morse S (2009) *Crime and culpability: a theory of criminal law*. Cambridge University Press, Cambridge
- Avio KL (1990) Retribution, wealth maximization, and capital punishment: a law and economics approach. *Stetson Law Rev* 19:373–409
- Avio KL (1993) Economic, retributive and contractarian conceptions of punishment. *Law Philos* 12:249–286
- Beccaria C (1764) on crimes and punishments (trans: Young D). Hackett Publishing, Indianapolis (1986 edn)
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Bentham J (1781) *The principles of morals and legislation*. Prometheus Books, Buffalo (1988 edn)
- Berman MN (2013) Two kinds of retributivism. In: Duff RA, Green SP (eds) *Philosophical foundation of criminal law*. Oxford University Press, Oxford, pp 433–457
- Braithwaite J, Pettit P (1990) *Not just deserts: a republican theory of criminal justice*. Clarendon, Oxford
- Brooks T (2012) *Punishment*. Routledge, Abingdon
- Byrd BS (1989) Kant's theory of punishment: deterrence in its threat, retribution in its execution. *Law Philos* 8:151–200
- Cahill MT (2007) Real retributivism. *Wash Univ Law Rev* 85:815–870
- Cahill MT (2011) Punishment pluralism. In: White MD (ed) *Retributivism: essays on theory and policy*. Oxford University Press, Oxford, pp 25–48
- Card C (1975) Retributive penal liability. *Am Philos Q Monogr* 7:17–35
- Cottingham J (1979) Varieties of retribution. *Philos Q* 29:238–246
- Davis LH (1972) They deserve to suffer. *Analysis* 32:136–140
- Davis M (1983) How to make the punishment fit the crime. *Ethics* 93:726–752
- Davis M (1986) Harm and retribution. *Philos Public Aff* 15:246–266
- Dolinko D (1997) Retributivism, consequentialism, and the intrinsic goodness of punishment. *Law Philos* 16:507–528
- Duff RA (2001) *Punishment, communication, and community*. Oxford University Press, Oxford
- Duff RA, Garland D (eds) (1994) *A reader on punishment*. Oxford University Press, Oxford
- Ehrlich I, Liu Z (eds) (2006) *The economics of crime, vol III*. Edward Elgar, Cheltenham
- Feinberg J (1965) The expressive function of punishment. In: *Doing and deserving*. Princeton University Press, Princeton, pp 95–118
- Hampton J (1984) The moral education theory of punishment. *Philos Public Aff* 13:208–238
- Hart HLA (1968) *Punishment and responsibility: essays in the philosophy of law*. Oxford University Press, Oxford
- Hegel GWF (1821) *The philosophy of right* (trans: Knox TM). Oxford University Press, Oxford (1952 edn)
- Honderich T (1984) *Punishment: the supposed justifications*, rev edn. Penguin Books, Harmondsworth
- Kant I (1797) *The metaphysics of morals* (trans: Gregor M). Cambridge University Press, Cambridge (1996 edn)
- Kaplow L, Shavell S (2002) *Fairness versus welfare*. Harvard University Press, Cambridge, MA
- Mackie JL (1985) *Persons and values*. Clarendon, Oxford
- Markel D (2011) What might retributive justice be? The confrontational conception of retributivism. In: White MD (ed) *Retributivism: essays on theory and policy*. Oxford University Press, Oxford, pp 49–72
- Moore M (1993) Justifying retributivism. *Isr Law Rev* 27:15–49
- Morris H (1968) Persons and punishment. *The Monist* 52:475–501
- Murphy JG (1972) Kant's theory of criminal punishment. In: Beck LW (ed) *Proceedings of the third international kant congress*. D. Reidel, Dordrecht, pp 434–441
- Murphy JG (1973) Marxism and punishment. *Philos Public Aff* 2:217–243
- Murphy JG (1987) Does Kant have a theory of punishment? *Columbia Law Rev* 87:509–532
- Nadelhoffer TA (2013) *The future of punishment*. Oxford University Press, Oxford
- Nozick R (1981) *Philosophical explanations*. Belknap, Cambridge, MA
- Plato (360 BCE) *Crito* (trans: Jowett B). Available at <http://classics.mit.edu/Plato/crito.html>
- Posner RA (1980) Retribution and related concepts of punishment. *J Leg Stud* 9:71–92
- Simmons AJ, Cohen M, Cohen J, Beitz CR (eds) (1995) *Punishment: a philosophy & public affairs reader*. Princeton University Press, Princeton
- Sunstein C (2005) On the psychology of punishment. In: Parisi F, Smith VL (eds) *The law and economics of irrational behavior*. Stanford University Press, Stanford, pp 339–357
- Tunick M (1992) *Punishment: theory and practice*. University of California Press, Berkeley
- Von Hirsch A (1976) *Doing justice: the choice of punishments*. Hill and Wang, New York
- Von Hirsch A (1985) *Past or future crimes: deservedness and dangerousness in the sentencing of criminals*. Rutgers University Press, New Brunswick
- Wertheimer A (1975) Should punishment fit the crime? *Soc Theory Pract* 3:403–423
- White MD (2009) Retributivism in a world of scarcity. In: White MD (ed) *Theoretical foundations of law and economics*. Cambridge University Press, Cambridge, pp 253–271
- White MD (2011a) *Pro Tanto* retributivism: judgment and the balance of principles in criminal justice. In: White MD (ed) *Retributivism: essays on theory and policy*. Oxford University Press, Oxford, pp 129–145
- White MD (ed) (2011b) *Retributivism: essays on theory and policy*. Oxford University Press, Oxford
- Wittman D (1974) Punishment as retribution. *Theory Decis* 4:209–237

Risk Management, Optimal

Andrew Torre

School of Accounting, Economics and Finance,
Deakin University, Melbourne, VIC, Australia

Abstract

Individuals continually confront a discrepancy between ever expanding and changing wants and the means that they have at their disposal, time, and income, to satisfy them. One of the consequences is the need to make constrained choices between alternatives that have uncertain outcomes. Risk is a different concept from uncertainty. Individual optimal risk management means reducing, eliminating, or fully bearing risk, after conducting a “cost-benefit” analysis. In practice, however, cognitive biases mean that many decisions are not economically rational, necessitating paternalistic government and judicial interventions. Systemic, or whole financial system collapse risk is, optimally managed using well-designed macroprudential regulatory tools. The source of this type of risk is the inherent dynamics of the financial system over the course of the business cycle, interacting with credit market negative externalities, often as in the case of the GFC, spawned by government regulatory failure.

Synonyms

Financial risk; Operational risk; Systemic risk

Definition

Individually, decision makers optimally or rationally manage operational and financial risk, when they reduce it, eliminate it completely, or fully bear it, after conducting a “cost-benefit” analysis. Systemic, or the risk of whole financial system collapse, which arises due to credit market externalities, is managed optimally by appropriate macroprudential regulation.

Introduction

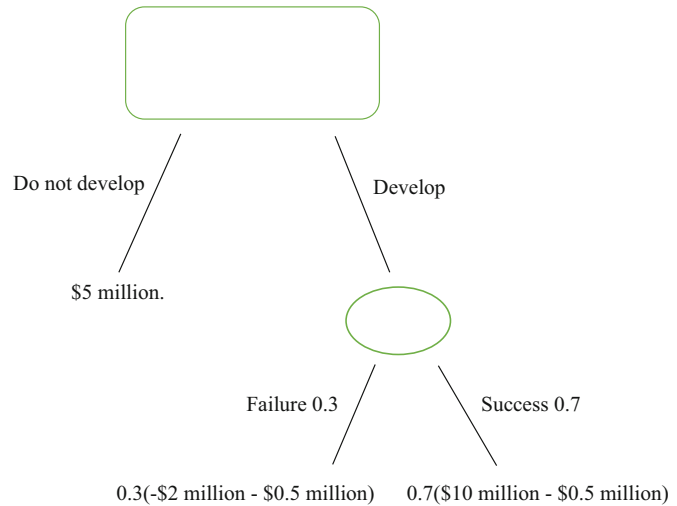
Risk and uncertainty are distinct economic concepts in economic theory; however, most contemporary microeconomic texts tend to use these terms interchangeably. In a risky situation, a rational decision maker confronts well-defined probabilities, while in an uncertain one, these are not known, or poorly defined (Knight 1921). This entry is only concerned with the management of risky choices, interpreted in an economic framework. Individuals and businesses confront and have to manage operational and financial risks. Attempts are made to minimize the cost of risk using “tangible assets or actions” in the former, and “options on financial instruments or commodities” in the latter case, respectively (Viney 2011, p. 383). In addition, governments need to optimally manage systemic or the risk of total financial system collapse. “Whole system” risk management is particularly topical given the recent global financial crisis (GFC) and its aftermath, which is still very evident in parts of Europe and to a lesser extent in America.

At the outset it should be noted that there is a large “noneconomic” literature on risk management. An example is provided by Ritchie and Reid (2013) in the context of ecotourism businesses. They provide three different definitions gleaned from different sources: first, “The Australian and New Zealand Standard,” which defines risk as “the effect of uncertainty on objectives” (p. 274); second, Glaesser who defines risk as “the product of magnitude of damage and the probability of occurrence” (ibid.); and third, Priest, who “differentiates between real and perceived risk, with perceived risk the best estimate of real risk” (ibid). The authors’ definition of risk is “any threat that will negatively impact an organization’s ability to achieve its objectives and execute its strategies successfully” (ibid.).

Rational Strategies to Deal with Risk

A simple example will enable the reader to see how rational optimizing decisions are made under risk and uncertainty. A risk-averse Amanda is

Risk Management, Optimal, Fig. 1 Simple Decision Tree illustrating Risk and Uncertainty



running a software development business and has to decide whether to develop a piece of software for a new overseas export market. If she succeeds in the export market she gains \$10 m, if not she loses \$2 m. The payoffs \$10 m and \$2 m are equal to the profit from each alternative. If the software is not developed, the business will continue to earn \$5 m a year. The estimated probabilities of success and failure are 0.7 and 0.3 respectively, and Amanda's risk premium is \$0.5 million. It is useful to set out this information in terms of a decision tree that incorporates uncertainty and, in this instance, operational risk as two distinct concepts (Fig. 1).

Operational risk is incorporated into the analysis by adjusting the decision maker's payoffs, while probabilities attached to each of the branches of the lower portion of the decision tree introduce uncertainty. When there is uncertainty, the payoffs become expected values. Since risk is a cost to a risk averter, the risk premium is subtracted from expected profit in each of the payoffs. Expected profit equals: $0.3 (-\$2 \text{ m} - \$0.5 \text{ m}) + 0.7 (\$10 \text{ m} - \$0.5 \text{ m}) = \$5.9 \text{ million}$. Since $\$5.9 \text{ m} > \5 m , the software should be developed even though Amanda is risk averse. The decision would be reversed if the cost of risk was high enough to make the expected profit from exporting $< \$5 \text{ m}$. If Amanda was a risk taker, the \$0.5 m, being a benefit, would be added to each of the payoffs. The expected profit would be

\$6.9 m, which would be $> \$5 \text{ m}$, and the software would be exported. If she was risk neutral, it would not be necessary to make any adjustment to the payoffs, since risk is neither a cost nor a benefit. The expected profit would be \$6.4 m. Again the optimal decision would be to export the software. In Amanda's case, the operational risk is rationally managed by using a decision tree that factors in the probabilities of success (failure) and the cost of risk. This is a form of "cost-benefit" analysis. If the venture fails, she absorbs the expected loss.

Legal Regulation of Operational and Financial Risk Management

The above analysis is predicated on the assumption that decision makers are economically rational; however, human cognitive biases considerably complicate risk management. Thaler and Sunstein (2008) characterize the human brain as having two separate cognitive systems, which are in conflict with each other, resulting in a tendency for irrational to dominate rational decision-making. The first or automatic system is in the words of the authors, "uncontrolled, effortless, associative, fast, unconscious and skilled" (p. 19), while they characterize the second or reflective system as "controlled, effortful, deductive, slow, self-aware and rule following" (p. 20).

As a consequence of this brain anatomy, humans are prone to exhibit a number of cognitive biases in different contexts, which make risk management along the lines of the rational “cost-benefit” model difficult to achieve in practice. There are two responses to this dilemma: government paternalistic interventions and legal regulation. An example of the former would be making the wearing of protective clothing and a mask compulsory for asbestos removal, rather than optional, based on individuals’ assessment of their own risk of subsequent harm.

In addition, judge-made or common law complements government paternalistic policies to assist individuals manage risk optimally. Both contract and torts laws are fundamentally concerned with the issue of whether an actual loss should be allowed to lie where it falls or shifted onto someone else. Torts law differs from contract law in that it regulates non-contractual interpersonal relationships. Specifically, the law pertaining to exemption clauses in standard form contracts and the tort of negligence both address risk management in settings of inequality of bargaining power and high bargaining costs, respectively.

Since the common law of contract is predicated on the principle of freedom of bargaining, there is both judicial and statutory reluctance to interfere with a voluntarily and honestly negotiated allocation of risks between the parties. An example is the specification of the contingencies that impact the promisor’s obligation to perform the contract, such as physical destruction of goods before their delivery date falls due. At common law the general rule is that “prima facie a promisor takes the risk of an event happening, which prevents him from performing his promise” (Scanlan’s *New Neon Ltd v Tooheys Ltd* (1943) 67 CLR 169 at 200 (Latham CJ), cited in Seddon and Ellinghaus (1.107, 2008). However in the case of a sale of goods, s. 12 of the 1958 Victorian Goods Act provides that “subject to contrary agreement a contract to sell specific goods is avoided if the goods perish without fault of either party” (Seddon and Ellinghaus 19.17). As the parties are free to stipulate a different allocation of risk, the common law principle of freedom of contract is preserved.

Fairness in the negotiation process, as such, is not a guiding principle in allocating risks between

the parties, rather “it is the essence of entrepreneurship that parties will sometimes act with selfishness” (Kirby P in *Biotechnology Australia Pty Ltd v Pace*, (1988) 15 NSWLR 130, 132–133, (cited in Paterson (2012), 1.40). This also applies to consumer contracts, as explained by McHugh J in *West v AGC (Advances) Ltd* (1986) 5 NSWLR 610, 621. “I do not see how that contract can be considered unjust simply because it was not in the interest of the claimant to make the contract or because she had no independent advice” (Paterson, 1.40). As long as consumers can choose between competing standard form contracts with different price and term combinations, then these contracts “represent the agreed allocation of risk by contracting parties, and interference in these agreements is likely only to produce inefficient outcomes” (Paterson, 1.60). Rational bargaining implies that expected contingencies are allocated on the basis of the lower-cost bearer or avoider of the risk. In practice, however, higher-cost bearers may be observed subsuming risks due to relatively weak bargaining positions and/or cognitive biases, infecting the negotiating process. A very likely outcome of this incongruity is terms that are unfair and/or inefficient. This issue has been addressed by both the common law and more recently the Australian legislature.

(a) **Exemption clauses in standard form contracts at common law**

Exclusion or limitation clauses are a classic risk allocation and management device found in standard form contracts; examples of these purporting to shift risk include “sold with all faults,” “no warranty given,” “all implied terms excluded,” and “at own risk” (Seddon and Ellinghaus 2008, 10.62). These are efficient (inefficient) if they shift risk onto the party whose precaution plus risk costs are lower (higher). Adopting the Kahneman (2011) formulation, they are unfair if they can be construed as an “exploitation of market power by the promisor to impose a loss on the promisee” (p. 306). In considering exemption clauses in standard-form contracts, the courts have distinguished negotiations between businesses from those between a business and consumer, on the basis of differences

in equality of bargaining power. An example of this judicial dichotomy is found in *Photo Production Ltd v Securicor Ltd* [1980] AC 827, where Lord Diplock wrote at 851: “in commercial contracts negotiated between business-men capable of looking after their own interests, and of deciding how risks inherent in the performance of various kinds of contract can be most economically borne (generally by insurance), it is, in my view, wrong to place a strained construction upon words in an exclusion clause, which are clear and fairly susceptible of one meaning only, even after due allowance has been made for the presumption in favour of the implied primary and secondary obligations” (Clarke et al. 2008, p. 252).

In relation to consumers, while the High Court of Australia (HCA) in *Sydney City Council v West* (1965) 114 CLR 481 has rejected the doctrine that an exclusion clause cannot provide protection against liability for a “fundamental breach,” regardless of how widely it is drafted, the courts have nevertheless “prevented exemption clauses from being used unconscionably against consumers by a rigorous application of the ordinary rules of construction” (Clarke et al., p. 251). Consequently, exemption clauses cannot be used to excuse a “radical breach of the promisor’s obligations under the contract” (*Windeyer J in TNT (Melbourne) Pty Ltd v May & Baker (Australia) Pty Ltd* (1966), 115 CLR 353. HCA (Clarke et al., pp. 256–258). Generally, at common law, the party who wishes to shift risk using an exemption clause has to show that it has been incorporated into the contract and, when correctly construed, is applicable to the situation at hand (Seddon and Ellinghaus, 10.66). An example of a rule of construction is the “four corners rule,” which states that “an exclusion clause will confer protection only in respect of conduct that occurred in the performance of the contract” (Clarke et al., p. 256). Therefore, if a council inserts an exemption clause in a contract of bailment seeking to protect it from loss of, or damage to, parked vehicles, this term will be construed “*contra proferentem*” against the bailee. As such, “it will not protect the Bailee from ‘storing the goods in a place or in a manner other than that authorised by the contract, or if the Bailee consumes or destroys them instead of storing them, or if he

sells them” (*The Council of the City of Sydney v West* (1965) 114 CLR 481, HCA, per Barwick CJ and Taylor J. (Clarke et al., p. 256).

The attack on unfair contract terms has been taken up by the Australian legislature, through its enactment of the Unfair Contract Terms Law (“UCTL”), contained in Pt 2-3 of the Australian Consumer Law (“ACL”), in Schedule 2 to the Australian Competition and Consumer Act 2010 (“CCA”), and Pt 2 Div 2 Subdiv. BA of the Australian Securities and Investments Commission Act 2001 (Cth) (“ASIC Act”) (Paterson 1.10.) Exclusion clauses are regulated under the consumer guarantees law (CGL) (Paterson 13.50). Consumer guarantees, which apply to the supply of goods and services to a consumer, replace the implied terms regime under the now repealed Australian Trade Practices Act (1974), since it was thought that these were too complex and uncertain (Paterson 11.30). The consumer guarantees are to be found in Chapter 3 Pt 3-2 Div. 1 Subdiv. A of the Australian Consumer Law (Paterson 11.80). An example is that the goods will be of acceptable quality, i.e., *inter alia*, free from defects, safe, and durable (Paterson 11.80, 11.90). While the risk of unacceptable quality is *prima facie* borne by the seller, the legislation shifts it onto the buyer in the following circumstances, in which case the purchaser cannot rely on the acceptable quality guarantee. First, defects are specifically drawn to the consumer’s attention; second, they are caused by abnormal use; or third, the consumer should reasonably have become aware of them since they were examined before being bought (Paterson 11.90). Other than in these cases, an exclusion clause inserted into the contract, which purports to exempt the seller from bearing the risk of unacceptable quality, is void under Australian Consumer Law, s 64(1). In addition, it may also be void because it is unfair under the Unfair Contract Terms Law (Paterson 11.180). The tests are different under the two regimes. In the first case, the “court will consider what the term purports to do and its effect,” while in the second, a court “will apply the test for an unfair term” (Paterson 11.180). Under the legislation, unfair terms are characterized as “imbalanced when they attempt to detract from the statutory rights held by consumers” and

therefore “lacking transparency” (Paterson **11.180**). A reasonable interpretation of the common law and reformulated statutory provisions taken as a whole, is that they are increasingly leading to outcomes that satisfy both the fairness and efficiency criteria.

(b) The tort of negligence

Fleming characterizes the function of torts law as ex post social risk management, “the law of torts then is concerned with the allocation of losses incident to man’s activities in modern society, and the task confronting the law of torts is, therefore, how best to allocate these losses in the interest of the public good” (Fleming (1998), p. 3). Torts law achieves this by substituting liability rules, strict liability and negligence, for voluntary bargains between the parties. Since in common law jurisdictions negligence is comparatively much more important, the discussion will focus on this tort. Mendelson, a torts law scholar, interprets this species of action on the case in a risk management framework. She writes, “the rationale for the tort of negligence is two-fold: (1) to enable those wrongfully injured through the negligent conduct of others to obtain compensation from the injurers; and (2) to impose standards based on a duty to avoid creating risks, which may result in an injury to another. If there is no way of avoiding a particular risk, or of predicting whether or not it will materialise, then the risk ought to be disclosed to those who may be harmed by it” (Mendelson, p. 278). “At common law, to be successful in a cause of action in negligence, the plaintiff (victim) has to establish on the balance of probabilities that: (i) the defendant (injurer) owes him or her a duty of care; (ii) the defendant is at fault by breaching the duty of care, by falling below the standard expected of a reasonable person in the defendant’s position; (iii) the defendant’s breach factually caused the plaintiff’s harm and (iv) the harm was reasonably foreseeable” (Mendelson, p. 281).

The second and third aspects of the common law of negligence (principles of breach and causation) have been recently modified by legislation in all of the Australian states. *The Civil Liability Act 1936* (SA), s 32 and 32(2); *Civil Liability Act*

2002 (NSW), s 5B and 5B(2); *Civil Liability Act 2003* (Qld), s 11 and s 9(2); *Civil Liability Act 2002* (WA), s 5B and s 5B(2); *Civil Liability Act 2002* (Tas), s 11 and s 11(2); *Civil Law (Wrongs) Act 2002* (ACT), s 43 and s 43(2); and *Wrongs Act 1958* (Vic), s 48 and s 48(2) each provides that “(1) A person is not negligent in failing to take precautions against a risk of harm unless: (a) the risk was foreseeable (that is, it is a risk of which the person knew or ought to have known), and (b) the risk was not insignificant, and (c) in the circumstances, a reasonable person in the person’s position would have taken those precautions. (2) In determining whether a reasonable person would have taken precautions against a risk of harm, the court is to consider the following (amongst other relevant things): (a) the probability that the harm would occur if care were not taken; and (b) the likely seriousness of the harm; and (c) the burden of taking precautions to avoid the risk of harm; and (d) the social utility of the activity that creates the risk of harm” (Mendelson **2010. 11.1.2, 11.2**).

In a basic sense, these provisions bear some resemblance to the “lower cost precaution taker plus bearer of risk principle.” Study of the cases on negligence, applying common law and the modifying statutory principles, is very valuable because it provides detailed guidance to the difficulties judges face in applying the criteria to complex individual risk management problems that arise in practice and how these might be resolved. The first HCA case to begin interpreting the new legislation, *Adeels Palace Pty Ltd v Moubarak* (2009) HCA 48 (2009) 239 CLR 420, is a good example. “The defendants as occupiers of a reception and restaurant business, owed a duty of care to two men, who while on the premises were shot by a gunman in December 2002. The gunman was one of the patrons, who initially came into the restaurant unarmed; however following a violent altercation on a dance floor, returned with a loaded gun and used it to shoot the plaintiffs” (Mendelson. **11.2.1.1**). “The High court of Australia was asked to decide, pursuant to s 5B of the NSW Civil Liability Act, whether the defendants had failed to take reasonable precautions against the risk posed by the gunman” (Mendelson. **11.2.1.1**). “Since the High Court found that the

plaintiffs had not established causation (see below), the issue was not decided, however the Court did provide guidance on the matter of reasonable precautions” (Mendelson, **11.2.1.1**).

First, given that harm has actually occurred, foreseeable risk has to be assessed before the event or activity occurs and not on the basis of what actually happens if the risk eventuates (Mendelson **11.2.1.1**). In this case, this was the probability of violence occurring at a reception and restaurant business, if “licensed security personnel who would act as crowd controllers or bouncers were not provided” (Mendelson **11.2.1.1**). Second, the foreseeable risk that is managed cost-effectively by taking precautions does not mean “general risks associated with certain activities” (the common law rule) but foreseeability of the specific risk “that called for, as a matter of reasonable precaution, the presence or physical authority of bouncers or crowd controllers to deal with it safely” (Mendelson **11.2.1.1**, p. 337). Third, “there is no liability unless factual causation is established by the plaintiff, and this ‘is determined by the ‘but for’ test: but for the negligent act or omission, would the harm have occurred?” (Mendelson, **12.2.2.1**). In the Adeels case, French CJ, Gummow, Hayne, Heydon, and Crennan JJ found that factual causation had not been established. This was because the evidence did not establish “that security personnel could or would have prevented re-entry by the gunman: a determined person armed with a gun and irrationally bent on revenge” (Mendelson, **12.2.2.1**). Since the court found for the defendants, they had impliedly managed risk optimally.

Credit Market Negative Externalities and Systemic Risk

Financial, or more specifically credit markets, only manage an individual’s or businesses’ financial risk optimally, if none of the cognitive biases infect their choices. However they will not manage aggregate financial risk optimally, because market forces generate financial cycles arising from financial system-induced negative externalities. This stricture also applies even if all decision makers are perfectly rational in the economic

sense. Failures in credit markets give rise to systemic risk, the risk that the entire financial system will collapse. Downturns in financial cycles are known as “balance sheet recessions” and differ from business (inventory or inflation) cycles (BIS 2014, Box 111.A, p. 45; Chapter 4, p. 65). Balance sheet-induced cycles “tend to be much longer than business cycles, and are best measured by a combination of credit aggregates and property prices” (BIS 2014, p. 65). “On average it takes about four and a half years for per capita output to rise above its pre-crisis peak. The recovery of employment is even slower” (BIS 2014 Box 111.A, p. 45).

Rising prices in asset markets (land, housing, and equities) that no longer reflect underlying true determinants of economic value such as land scarcity usually portend emerging and troublesome market bubbles. The cause of the 1930s depression was the bursting of the stock market bubble in 1929 and the widespread bank failures and deflation that followed. High economic growth in the 1920s drove the share market upward, and banks willingly financed most of the purchase price of share portfolios. The latest example of a debt-driven severe economic downturn, and accompanying bank failures caused by bursting bubbles, is the 2008 so called GFC, which began in the United States and then subsequently spread throughout many European economies. Ball (2014) has estimated the weighted average loss in potential output for 23 countries including the United States from the “Great Recession” as being 7.2% in 2013, increasing to 8.4% in 2015 (Greece, Hungary, and Ireland >30% and the United States 5.3%). In terms of 2015 US dollars, the 8.4% loss translates into \$4.3 trillion (Ball 2014, p. 5).

At the dawn of the twenty-first century, another stock market bubble burst. However this was eclipsed by a boom in real housing prices that had emerged in the mid-1990s and led *The Economist* magazine in 2005 to describe it as “the greatest bubble in history” (Garnaut and Llewellyn-Smith, p. 11). In 1999 in America, the Clinton administration wanted to increase home ownership among poorer members of the community. This was to be achieved by lowering interest rates through expansion of the money supply and giving the mortgage providers Fannie Mae

(Federal National Mortgage Association) and Freddie Mac (Federal Home Loan Mortgage Corporation) incentives to make loans to subprime borrowers, who being poor credit risks could not borrow in conventional capital markets (Garnaut and Llewellyn-Smith, p. 20). Incentives included tax breaks and the ability to securitize subprime loans. Securitization is a “practice through which an illiquid but income producing asset is converted into a security more easily traded by investors” (Garnaut and Llewellyn-Smith, p. 41). It “shifts and repackages risk but does not eliminate it” (Garnaut and Llewellyn-Smith, p. 41). “The securitization process was insured often by selling a guarantee called a credit default swap (CDS)” (Garnaut and Llewellyn-Smith, p. 51).

“Credit analysts at JPMorgan invented the CDS” (Garnaut and Llewellyn-Smith, p. 40). A CDS transfers the risk that a debtor will not repay a loan from the creditor to an insurer, in this instance, “risk was shifted from company to company; one paid the other to absorb the risk of default” (Garnaut and Llewellyn-Smith, p. 65). As a consequence, sellers of these instruments were encouraged to take even greater risks, instead of properly managing their balance sheets to ensure that they could meet their financial obligations. “Being over the counter derivatives, these transactions were hidden from regulators and the wider market of investors” (Garnaut and Llewellyn-Smith, p. 66). A CDS is a class of interconnecting derivative. Garnaut and Llewellyn-Smith, p. 66 write: “During the crash of 2008, when confidence was tested, the risk that had been shifted from thousands of individual companies suddenly became a risk to the entire system. Because nobody knew where the risk lay, everyone was suspect.” The crunch came when US Federal Reserve officials pursued a sharp contractionary monetary policy to break the real estate bubble, and the supply of credit rapidly contracted. There was a sharp rise in mortgage defaults and foreclosures. Since securitized mortgages had been bought all over the world, overseas financial institutions and investors were also impacted.

“A principal cause of the GFC was the policies of the U.S. monetary and regulatory authorities” (Stiglitz 2012). Poorly designed government policy and regulatory failure gave critical decision

makers in financial institutions perverse incentives, which contributed to severe credit market failures, and widespread externalized costs. They also fuelled the increasingly irrational speculative activity that was driving house prices upward. Three negative externalities have been identified in the literature: (i) strategic complementarities, (ii) fire sales, and (iii) interconnectedness. The first category arises due to vigorous competition between financial institutions and increasing adverse selection problems, lowering profits from loans as the upswing of the cycle gathers momentum. The outcome is that all institutions tend to engage in the same risky lending practices, thereby acquiring virtually the same “toxic” loan portfolios. Fire sale externalities come into play when the bubble has burst. Widespread asset sales at prices below fundamental economic values during times of financial distress cause asset prices to decline adversely, impacting all financial intermediaries’ balance sheets. The Bank of England describes the source of these two externality categories as “collective overexposure of the financial system in the upswing of the credit cycle, and excessive risk aversion in the downswing” (Bank of England, p. 10). Interconnectedness externalities arise from “the rapid dissemination of shocks between interconnected financial institutions” (De Nicolo' et al. 2012, pp. 4–5). “A network of large multinational based and interlinked financial institutions, while mitigating the impact of small shocks by spreading them, seems to amplify large shocks because they can reach more counterparties” (De Nicolo' et al. 2012, p. 10). According to The Bank of England, the source of this network externality is “the distribution and concentration of risk within the financial system” (Bank of England, p. 15). The role that “interconnecting derivatives” such as CDSs played in transmitting shocks was illustrated in the previous paragraph.

Managing credit market negative externalities optimally, necessitates an absence of government regulatory failure plus macroprudential regulation of the financial system. This differs from microprudential regulation, which is designed to align principal (lending managers)-agent (customers) incentives, given the difficulty agents confront in monitoring principals’ lending activities. An

example of a solution is the minimum capital requirement of Australian banks. “If a deposit taking institution sustains loan losses in excess of its profits, it must write them off against capital, not against liabilities (depositors’ funds). Consequently in order to deal with such a contingency, APRA (The Australian Prudential Regulation Authority) mandates that such an institution must hold at least 8% of its capital base (equity) in a stipulated form: shareholders’ funds and retained earnings (Tier 1 capital), and certain types of preference shares (Tier 2 capital)” (Viney, pp. 89–90). However microprudential regulation does not by itself protect the integrity of the whole financial system. Furthermore, the whole behaves differently from its constituent parts. “Stabilising one institution can destabilise the whole system” (De Nicolo' et al. 2012, p. 7).

There are two ways of classifying macroprudential policy tools. De Nicolo' et al. 2012, p. 11, describe these as: “capital requirements (surcharges), liquidity requirements; restrictions on activities, assets or liabilities, and taxation.” The Bank of England 2011, p. 17, adopts the following groupings: “those that affect: (i) the balance sheets of financial institutions; (ii) the terms and conditions of loans and other financial transactions and (iii) those that influence market structures.” Systemic financial system risk is either time varying, emanating from complementarity and fire sale externalities, or cross sectional, arising from interconnectedness externalities. “Balance sheet and loan restriction tools address dynamic risk, while market structure interventions target cross sectional risk” (Bank of England, p. 17). Generally regulation of capital buffers can potentially ameliorate all three externality categories (De Nicolo' et al. 2012, p. 14). Other examples of regulatory instruments are restrictions on bank asset allocations to limit asset growth in the upturn of the cycle, and adjustment costs in the downturn, (De Nicolo' et al. 2012, p. 12) and the imposition of maximum leverage ratios (Bank of England, p. 17).

Conclusion

In this entry, while financial has been distinguished from operational risk, they are both managed

optimally in the same way, through “cost-benefit” assessments. Doing so will result in risk being reduced, eliminated, or fully borne by the decision maker. In order to implement these “optimization rules,” rational behavior on the part of individuals, which implies an absence of cognitive biases, is necessary. This stringent requirement is rarely met in practice, and government paternalistic risk management is often observed. An example is outright prohibition of the consumption of harmful drugs. In addition, the common law principles of contractual exemption clauses and negligence both complement paternalistic policies. The first does this by targeting unfair risk-shifting contractual terms, while the second assesses, *inter alia*, the adequacy of the respective parties’ precautions, after an accident has occurred.

The optimal management of systemic risk or total financial system collapse requires an absence of government regulatory failure plus robust prudential regulation. An example would be the imposition of capital controls between countries. Unimpeded market forces generating credit market negative externalities will usually fail to ensure that the economy’s aggregate exposure to risk is optimal over the course of the business or inventory cycle. Financial system credit creation, during cyclical upswings, fuels increasing asset prices, balance sheet leverage, and the outcome is excessive exposure to sub-optimal amounts of aggregate risk. Similarly, during the downswing of the cycle, credit withdrawal, deleveraging, and asset price declines can generate disproportionate output and employment losses that are “extraordinarily long lasting” (BIS, p. 7). Furthermore, “financial cycles can go largely undetected. They are simply too slow moving for policy makers and observers whose attention is focused on shorter-term output fluctuations” (BIS, p. 7).

References

Monographs

Clarke P, Clarke J, Zhou M (2008) *Contract law: commentaries, cases and perspectives*. Oxford University Press, South Melbourne

- Fleming J (1998) *The law of torts*. LBC information services, 9th edn. LBC Information Services, North Ryde
- Kahneman D (2011) *Thinking fast and slow*. Allen and Lane, New York
- Knight F (1921) *Risk, uncertainty and profit*. Augustus M Kelley, New York
- Mendelson D (2010) *The new law of torts*, 2nd edn. Oxford University Press, South Melbourne
- Paterson J (2012) *Unfair contract terms in Australia*. Thomson Reuters (Professional) Australia, Pyrmont
- Ritchie D, Reid S (2013) *Risk management and ecotourism businesses*. In: Ballantyne R, Packer J (eds) *International handbook on ecotourism*. Edward Elgar, Cheltenham, UK; Northampton, MA
- Seddon N, Ellinghaus M (2008) *Cheshire and Fifoot's law of contract*. Ninth Australian edn. LexisNexis Butterworths, Chatswood, NSW
- Thaler R, Sunstein C (2008) *Nudge: improving decisions about health, wealth and happiness*. Yale University Press, New Haven

General References

- Bank of England (2011) *Instruments of macroprudential policy*. A discussion paper
- Bank for International Settlements (BIS) (2014) 84th annual report.
- De Nicolò G, Favara G, Ratnovski L (2012) *Externalities and macroprudential policy*. IMF discussion note. 7 June 2012. International Monetary Fund
- Garnaut R, Llewellyn-Smith D (2009) *The great crash of 2008*. Melbourne University Press, Carlton
- Stiglitz J (2012) *Macroeconomics, monetary policy and the crisis*, Chap. 4. In: Blanchard O, Romer D, Spence M, Stiglitz J (eds) *In the wake of the crisis: leading economists reassess economic policy*. The MIT Press, Cambridge, MA
- Viney C (2011) *Financial market essentials*. McGraw Hill, North Ryde

Journal Articles

- Ball L (2014) *Long-term damage from the great recession in OECD countries*. NBER working paper no 20185. Issued May 2014
- Tversky A, Kahneman D (1981) *The framing of decisions and the psychology of choice*. *Sci New Ser* 211(4481): 453–458

Röpke, Wilhelm

Stefan Kolev

Wilhelm Röpke Institute, Erfurt and University of Applied Sciences Zwickau, Zwickau, Germany

Abstract

The purpose of this entry is to delineate the political economy of Wilhelm Röpke. To reach this goal, a history of economics approach is harnessed. First, the entry concisely reconstructs Röpke's life, intellectual evolution, and heritage. Second, it presents the specificities of his perspective on economic embeddedness focused on the stability of social order and on the role of informal institutions as indispensable stabilizers of this order.

Biography

Wilhelm Röpke (1899–1966) was a key figure within a fascinating generation of European economists. Even though not as widely received today as his colleagues and friends F.A. Hayek (1899–1992) and Walter Eucken (1891–1950), Röpke and his age peers had the mixed privilege to experience two world wars, but also the intellectual privilege to be among the most renowned economists in Europe during the period of the Great Depression as well as in the postwar decades. Röpke's vibrant life, his scholarly achievements as political economist and as social philosopher, but also his seminal role in shaping economic policy at several crucial junctures more than vindicate a detailed portrayal, with the main aim to show how he complements the ordoliberalism of the Freiburg School by his specific research program within the ordoliberal paradigm.

Röpke was born in a small town in the North-west of Germany and retained a lifelong sympathy for rural, small-scale social contexts like the one of his youth. He studied a combination of economics, law, and administrative science at Tübingen, Göttingen, and Marburg. Having

Robbery

► [Piracy, Modern Maritime](#)

completed both a (historicist) dissertation and a (theoretical) habilitation at Marburg, Röpke moved to Jena to become Germany's youngest professor at the age of 24 (Gregg 2010, pp. 7–8; Hennecke 2005, pp. 49–53). After a brief stay at Graz, he received a call to Marburg in 1929 and stayed there until 1933, but had to leave almost immediately after the National Socialist seizure of power because of his perennial outspoken opposition (Nicholls 1994, pp. 56–59; Hennecke 2005, pp. 99–114). Röpke and his close associate Alexander Rüstow (1885–1963) received positions at Istanbul, together with several other German émigrés. Both were eager to come back to Central Europe, and Röpke was lucky to receive a call in 1937 to the Graduate Institute of International Studies in Geneva, where he remained for the rest of his life.

Unlike Eucken or his Viennese colleagues, Röpke was not the type of scholar to form a school of his own (Boarman 1999, pp. 69–73) – instead, he was a highly gifted networker and succeeded soon to set up an international network connecting various European countries, one to be expanded after 1945 to the Americas (Zmirak 2001, pp. 201–206). Röpke and Rüstow were among the most active figures at the Colloque Walter Lippmann in Paris 1938, the birthplace of “neoliberalism” – a term with a history reaching well into the nineteenth century, but now to be used for a reformulation of liberalism by twentieth-century social scientists (Gregg 2010, pp. 82–86; Burgin 2012, pp. 67–78). Along with Hayek, Röpke acted in the immediate postwar years as the second initiator of the 1947-founded Mont Pèlerin Society, a hub for the few remaining liberal scholars, until that point isolated at their individual locations in Europe and the United States (White 2012, pp. 233–238; Kolev et al. 2014). Simultaneously, Röpke functioned as a crucial “spin-doctor” to Ludwig Erhard and was the mastermind behind some of Erhard's strategic plans for postwar Germany under the auspices of the Social Market Economy (Commun 2004; Goldschmidt and Wohlgemuth 2008, pp. 262–264). However, Röpke's enthusiasm for the effects of the “economic miracle” faded away during the 1950s, leading to a deepening

pessimism about the prospects of liberty, in combination with voicing ever-sharper conservative positions on social issues. After the “Hunold affair” within the Mont Pèlerin Society, a fall-out with Hayek and the American fraction in the Society ensued (Plickert 2008, pp. 178–190; Burgin 2012, pp. 137–143), with further detrimental effects for Röpke's weakened health leading to his passing away in February 1966.

Even though not as prominent in today's policy debates in Germany as authors like Eucken or Keynes, Röpke is frequently present in the official addresses of Chancellor Merkel or of finance minister Wolfgang Schäuble. In 2007, the Wilhelm Röpke Institute was established in Erfurt, close to Röpke's first professorship at Jena. Also, Röpke has been more widely received and discussed than Eucken in some other European countries (especially Switzerland and Italy) as well as in conservative circles in the United States (Commun and Kolev 2018).

The Political Economist Is More Than Just an Economist

Röpke's impulses to the research program of the incipient neoliberal movement in the 1930s and 1940s are seminal – and often overlooked today. Despite their worsening relationship in the late 1950s amid the tensions in the Mont Pèlerin Society, Hayek acknowledged how Röpke had realized “early, probably earlier than most of our contemporaries, that an economist who is only an economist cannot be a good economist” (Hayek 1959, p. 26). And this is indeed characteristic for Röpke's oeuvre: in line with Eucken's concept of the “interdependence of orders,” Röpke conceived a theory of the economy as an entity deeply embedded in and interrelated with adjacent social orders (Kolev 2013, pp. 121–136, 2015, pp. 424–427). Two issues are of paramount importance for his specific take on ordoliberal political economy: first his overarching concern with the stability of the economic order as embedded in society, and second his particular attention to what one would call today “informal institutions” – the particular cultural perquisites

and preconditions which to Röpke were more important for enabling stability than the formal legal rules framing the economy (Zweynert 2013, pp. 116–120). His focus on these two domains, stability and informal institutions, provides a valuable complement to the ordoliberal research program of the Freiburg School, and these domains also offer a helpful structure for continuing this exposition.

Stability Is Not to Be Taken for Granted: On the Fragility of the Spontaneous Order

Following up on Hayek's assessment above, Röpke can often be seen as prescient – both in his conceptual apparatus and his theoretical arguments. For example, he used the term “spontaneous order” as a description for the market in 1937, years before it explicitly found its key place in Hayek's terminology (Röpke 1963, p. 4). However, Röpke was also early to express serious concerns about the fragility of this order and to found his political economy on the issues of stability. The initiation was the Great Depression when he conceptualized, within the Austrian Business Cycle Theory, a phase of “secondary depression” during a slump of the economy: here, the useful effects of the “primary depression” with its merits in purifying the economy of the malinvestments of the boom are depleted – and the dangerous aspects of the depression start spreading, as its deflation becomes omnipresent (James 1986, pp. 329–342; Kolev 2013, pp. 178–183). The main danger here according to Röpke is not an economic one, rather it is rooted in the interdependence of the economic and the political order: The popular sentiments of unstoppable downward dynamics can in his analysis delegitimize not only the market economy as an economic order but also democracy as a political order and the free order of society as a whole (Röpke 1932, 1936, pp. 129–132). It was because of this notion of interdependence that Röpke, in contrast to his Austrian friends (Haberler 2000), pleaded in the early 1930s for active policy responses to stop the disintegration of the political order due to illiberal extremisms of various kinds.

Again, Hayek later acknowledged publicly the correctness of Röpke's broad perspective at this key juncture and also confessed his own narrowness in judging the harm of deflation only on purely economic criteria (Magliulo 2016).

In the decades to follow, Röpke expanded comprehensively on this early intuition of the crucial importance of stability in a system of interdependent social orders. In his trilogy published in German during the war (Röpke 1948, 1950, 1959), he searched for a diagnosis of the multiple crises of his age, for a therapy for them in Europe, and also for a solution set for the pressing issues in the international economic relations (Sally 1998, pp. 133–147). He explored a specific normative vision of the market within society, later to be called a “humane economy” in the English title of his probably most well-known book today (Röpke 1960), a vision to be depicted below. In the course of these endeavors, he left technical economics aside and moved increasingly into the domain of social philosophy, a parallel evolution observable in many of his age peers (Blümle and Goldschmidt 2006).

Informal Institutions, Not the Legal Framework Are the Key to Long-Term Stability

While the ordoliberalism of the Freiburg School focuses on the order-generating properties of the formal rules framing the economic process, the specifically Röpkean agenda within the “order in liberty” program of ordoliberalism is differently nuanced. Without disregarding the crucial role of formal legal institutions as a necessary condition for a stable system of social orders, Röpke's political economy did not perceive them as a sufficient condition. Rather, in an implicit division of labor with his Freiburg colleagues, Röpke underscored in his writings the essentiality of what he called the “anthropological and sociological” preconditions or prerequisites of a stable social order with a high degree of social cohesion (Röpke 1950, pp. 191–194, 1960, pp. 74–89). These cultural preconditions are partially ideal (called above “anthropological”) and relate to necessary values within the intellectual heritage of Christianity, and

partially they also have a material side (called above “sociological”) and depict the necessary social structures within which a “humane economy” is possible (Röpke 1960, pp. 222–235). Here an interesting contrast to Hayek can be drawn: while Hayek is primarily concerned with the threat which the logic of the small group can inflict on the mechanisms of the extended order of society, Röpke worries most about the opposite, i.e., about the threats stemming from “enmassment” where the logic of anonymous society uproots the individuals from the contexts of their traditional small communities and their particularly reliable social cohesion (Kolev 2016, pp. 16–20). Correspondingly, retaining and conserving these small contexts both in terms of economy and of society is the central goal of his “humane economy,” thus preventing the individuals from falling prey to ideational vacuum or to unnatural social structures – and as the market tends to use up the resources of its stabilizing pillars, these are to be permanently checked and, if necessary, stabilized anew by the state and other players in civil society (Röpke 1942, pp. 67–71).

This Röpkean plea for specific informal institutions to guarantee stability has often been criticized as romantic, naively conservative, or even “retro-utopian” (Solchany 2015, pp. 484–501). While these critiques for Röpke’s therapies do not fully lack justification, Röpke’s diagnosis regarding the fragility inherent in “spontaneous orders” can claim validity until today, in the global economic order and, most recently, also in the political order of Western societies. It will be intriguing to observe to what extent the age of digitalization with platforms like social media can partially bring back to relevance the logic of small groups and the stabilizing statics which Röpke expected from these groups, in this way possibly providing an antidote to the “too much” of dynamics which discontented fractions of Western societies attest to the processes of globalization.

Cross-References

- Austrian School of Economics
- Böhm, Franz
- Constitutional Political Economy

- Eucken, Walter
- Hayek, Friedrich August von
- Mises, Ludwig von
- Ordoliberalism
- Political Economy
- Schmoller, Gustav von

References

- Blümle G, Goldschmidt N (2006) From economic stability to social order: the debate about business cycle theory in the 1920s and its relevance for the development of social order by Lowe, Hayek and Eucken. *Eur J Hist Econ Thought* 13(4):543–570
- Boarman P (1999) Apostle of a humane economy: remembering Wilhelm Röpke (1899–1966). *ORDO Jahrb Ordn Wirtsch Ges* 50:69–91
- Burgin A (2012) The great persuasion: reinventing free markets since the depression. Harvard University Press, Cambridge, MA
- Commun P (2004) Erhards Bekehrung zum Ordoliberalismus: Die grundlegende Bedeutung des wirtschaftspolitischen Diskurses in Umbruchszeiten. Discussion paper 04/04, Walter Eucken Institut, Freiburg
- Commun P, Kolev S (eds) (2018) Wilhelm Röpke (1899–1966): a liberal political economist and conservative social philosopher. Springer, Heidelberg
- Goldschmidt N, Wohlgemuth M (2008) Social market economy: origins, meanings, interpretations. *Constit Polit Econ* 19(3):261–276
- Gregg S (2010) Wilhelm Röpke’s political economy. Edward Elgar, Cheltenham
- Haberler G (2000) Between Mises and Keynes: an interview with Gottfried von Haberler (1900–1995). *Austrian Econ Newsl* 20(1.). Available at: <https://mises.org/library/between-mises-and-keynes-interview-gottfried-von-haberler>
- Hayek FA (1959) Glückwunschsadresse zum 60. Geburtstag von Wilhelm Röpke. In: Röpke W (ed) *Gegen die Brandung*. Eugen Rentsch, Zurich, pp 25–28
- Hennecke HJ (2005) Wilhelm Röpke: Ein Leben in der Brandung. Schäffer-Poeschel, Stuttgart
- James H (1986) The German slump: politics and economics 1924–1936. Clarendon, Oxford
- Kolev S (2013) Neoliberale Staatsverständnisse im Vergleich. Lucius & Lucius, Stuttgart
- Kolev S (2015) Ordoliberalism and the Austrian school. In: Boettke PJ, Coyne CJ (eds) *The Oxford handbook of Austrian economics*. Oxford University Press, Oxford, pp 419–444
- Kolev S (2016) Stresstest fürs globale Dorf. *Schweizer Monat* 1037:16–21
- Kolev S, Goldschmidt N, Hesse J-O (2014) Walter Eucken’s role in the early history of the Mont Pèlerin Society. Discussion paper 14/02, Walter Eucken Institut, Freiburg
- Magliulo A (2016) Hayek and the great depression of 1929: did he really change his mind? *Eur J Hist Econ Thought* 23(1):31–58

- Nicholls AJ (1994) Freedom with responsibility. The social market economy in Germany, 1918–1963. Clarendon, Oxford
- Plickert P (2008) Wandlungen des Neoliberalismus: Eine Studie zu Entwicklung und Ausstrahlung der “Mont Pèlerin Society”. Lucius & Lucius, Stuttgart
- Röpke W (1932) Krise und Konjunktur. Quelle & Meyer, Leipzig
- Röpke W (1936) Crises and cycles. William Hodge, London
- Röpke W (1942) International economic disintegration. William Hodge, London
- Röpke W (1948) Civitas humana: a humane order of society. William Hodge, London
- Röpke W (1950) The social crisis of our time. University of Chicago Press, Chicago
- Röpke W (1959) International order and economic integration. D. Reidel, Dordrecht
- Röpke W (1960) A humane economy: the social framework of the free market. Henry Regnery, Chicago
- Röpke W (1963) Economics of the free society. Henry Regnery, Chicago
- Sally R (1998) Classical liberalism and international economic order: studies in theory and intellectual history. Routledge, London
- Solchany J (2015) Wilhelm Röpke, l'autre Hayek : Aux origines du néolibéralisme. Publications de la Sorbonne, Paris
- White LH (2012) The clash of economic ideas. The great policy debates and experiments of the last hundred years. Cambridge University Press, Cambridge
- Zmirak J (2001) Wilhelm Röpke: Swiss localist, global economist. ISI Books, Wilmington
- Zweynert J (2013) How German is German neo-liberalism? Rev Austrian Econ 26(3):109–125

von Mises (1881–1973), he made major contributions in different areas of economics, especially in the fields of marginal utility theory, welfare economics, monetary theory, business cycle theory, and the history of economic thought. Rothbard's economic theory is comprised in his opus magnum *Man, Economy and State*, published in 1962. The addendum *Power and Market* contains an economic theory of the state. He also made notable contributions in libertarian ethics and anarchist philosophy. He launched and spearheaded the modern libertarian and anarcho-capitalistic movement in the United States and attracted many intellectual followers worldwide, such as Walter Block (born 1941), Gary North (born 1942), Hans-Hermann Hoppe (born 1949), Joseph T. Salerno (1950), and Jörg Guido Hülsmann (born 1966).

Definition

Murray N. Rothbard (1926–1995) was a libertarian political philosopher and an economist in the tradition of the Austrian School of economics.

Life, Work, and Influence of Murray N. Rothbard

Life, Family, and Personal Background

Murray Newton Rothbard was born on March 2, 1926 in New York City as the only child of David and Rae Rothbard, who were Jewish immigrants from Poland and Russia, respectively. Rothbard was born in the Bronx. He and his family later moved to the Upper West Side of Manhattan into a rent-controlled apartment. Rothbard initially went to a public school, where he caused massive trouble for his parents, colleagues, and teachers (Rothbard 1981). After the move to Manhattan, he happily entered Birch Wathen, a private school founded in the early twentieth century. His winning one of the scholarships that were awarded to poor and middle-class boys in order to increase the ratio of boys to girls enabled him to attend (Flood 2008).

Rothbard, Murray

Karl-Friedrich Israel

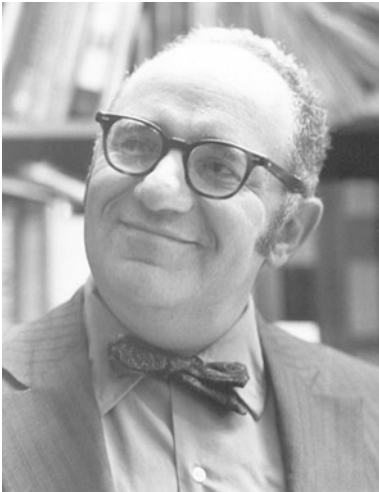
Faculty of Economics and Business

Administration, Humboldt-Universität zu Berlin, Berlin, Germany

Department of Law, Economics, and Business, University of Angers, Angers, France

Abstract

Murray Newton Rothbard (2 March 1926 in New York City–7 January 1995 in New York City) was a libertarian philosopher and one of the most influential adherents to the Austrian School of economics in the recent past. Building on the economic theory of his mentor Ludwig



Murray Rothbard, with permission from *Mises Institute* in Auburn, AL, USA

After finishing high school, he entered Columbia University in the fall of 1942 at the age of 16. As surprising it might seem in light of his aversion to mathematical economics, Rothbard was at one time in college a statistics major. He was an excellent student and apparently attracted the attention of recruiters for the Manhattan Project. Whether it actually led to a job offer or not is unknown. Rothbard was found to be unfit for military services (4-F) which exempted him from the draft during World War II (Flood 2008). He graduated in 1945 and enrolled in a graduate course in mathematical statistics with the famous statistician and economist Harold Hotelling (1895–1973), who was at the time a member of the faculty at Columbia University (1931–1946). However, disillusioned with Hotelling's lectures, he walked "out of the world of statistics, never to return" (Rothbard 2006, pp. 28–29). His main interests shifted to the social sciences. In 1946 he received the M.A. degree in economics.

According to Rothbard, growing up as a middle-class Jew in the 1930s and 1940s in New York City meant that literally everybody he knew – "friends, relatives, acquaintances, whatever — was either a Communist Party member or else was thinking about whether they should join it." He had "four aunts and uncles who were

Communist Party members and another two or three or four were pondering whether or not to join" (Rothbard 1981). His father had significant influence on him at that time as he was strongly opposed to the Communist Movement. So was Rothbard. In retrospect, he called his father "the first libertarian he knew" (Rothbard 1981). His parents met at an anarchist dance, and at some point his father probably was an active anarchist but decided to stop political action, when he found out that his friends, who were strongly opposed to World War I, had been arrested. From then on, he focused solely on his work as a chemist. One of the first books on anarchism that Rothbard read from the library of his father was Paul Eltzbacher (1908). He found it to be interesting and "exotic." However, he was not yet converted to anarchism as he was only about 10 years old (Rothbard 1981). This would happen some years later.

Rothbard was elected a member of *Phi Beta Kappa* and remained at Columbia University to pursue a Ph.D. in economics under the economic historian Joseph Dorfman (1904–1991). His thesis was entitled *The Panic of 1819* (Rothbard 1962) and it is until today the standard work on this particular economic crisis (Gordon 2007, p. 12). It took 10 years until he was awarded the doctorate degree in 1956 due to disagreements between Arthur Burns (1904–1987) and Dorfman. Burns, then professor at Columbia University and chairman of the Council of Economic Advisors (1953–1956), later chairman of the Federal Reserve (1970–1978), and ambassador to West Germany (1981–1985), lived in the same building as Rothbard. David Rothbard asked him to have a look at his son at the university, but it was Burns who brought his academic career to a virtual halt by telling Dorfman that he would "expect much more from Rothbard." JoAnn Schumacher, who became Rothbard's wife and close collaborator, found him one day sobbing at the doorsteps of his house, disheartened by Burns' obstacles (Flood 2008). Arthur Burns, Wesley Clair Mitchell (1874–1948), and John Maurice Clark (1884–1963) were among the most important professors at the faculty. They looked skeptically at economic theory and accepted the

institutionalist doctrine which conceives of economic theory as only relative to particular historical situations but not as universally true. Rothbard followed Columbia philosopher Ernest Nagel (1901–1985) and his criticism of institutionalism for its opposition to economic theory (Gordon 2007, pp. 9–11). This was an insurmountable conflict between Rothbard and Burns. It was not the only time that he was facing intellectual opposition. However, he would rarely compromise and deviate from a position that he deemed correct.

When Rothbard entered graduate school, there were a peak number of students due to the returning troops, the vast majority of which were social democrats or even communists and in any-way disapproved of the free market. At that time George Stigler (1911–1991), later Nobel Prize laureate (1982) of the Chicago school of economics, arrived at the campus (1947–1958). His first two lectures were on the evils of rent control and minimum wages, which led to some hysteria among the students (Rothbard 1981). To the intellectual relief of Rothbard, Stigler was not averse to economic theory. Through one of his pamphlets, written together with Milton Friedman (1912–2006) (Stigler and Friedman 1984), Rothbard encountered the Foundation for Economic Education. He contacted the group and made the acquaintance of founder Leonard E. Read (1898–1983); F. A. Harper (1905–1973), founder of the Institute for Humane Studies in 1961; and most importantly Ludwig von Mises (1881–1973), who immigrated to the United States in 1940. His radical defense of the free market, and in particular his book *Human Action*, published in 1949 had a profound effect on Rothbard's thinking. Rothbard participated regularly in Mises's seminar at New York University, for which he delivered a number of articles which were partly published in his later works. Mises became Rothbard's mentor, although they would not agree on everything, for example, the proper role of the state. Mises, as a classical liberal, advocated a state strictly confined to the protection of private property rights through a court and police system. Rothbard, however, went one step further and denied the necessity of the state

altogether. In this persuasion he was influenced by nineteenth-century individualist anarchists, such as Lysander Spooner (1808–1987), Benjamin Tucker (1854–1939), and Gustave de Molinari (1819–1912) (Gordon 2007, p. 13).

In terms of economic theory, Rothbard and Mises would only disagree on minor aspects but generally appreciate and admire each other's works. The William Volker Fund, which gave financial support to classical liberal scholars, commissioned Rothbard to write a textbook, which would explain Mises (1998) to college students. Something much larger would originate out of this project – his opus magnum *Man, Economy and State*, published in 1962, one of the most important contributions to Austrian economics in the past century (Hoppe 1999). Ludwig von Mises called Rothbard's treatise:

an epochal contribution to the general science of human action, praxeology, and its practically most important and up-to-now best elaborated part, economics. Henceforth all essential studies in these branches of knowledge will have to take full account of the theories and criticisms expounded by Dr. Rothbard. (Mises 1962)

Only one year later, in 1963, appeared *America's Great Depression*, in which Rothbard applied *Austrian Business Cycle Theory* as laid down by Mises (1912) to the great economic crisis of the 1930s. He worked for different institutions and foundations and published scientific and non-scientific articles, comments, and reviews in various journals and magazines. From 1966 to 1986, Rothbard taught economics at the Brooklyn Polytechnic Institute, an engineering school without a major program in economics.

Although Rothbard was an agnostic and a secularist, he became a columnist for a little right-wing magazine written for protestant ministers called *Faith and Freedom* under the pseudonym Aubrey Herbert. He was criticized by the constituents for being a communist, as he was opposing the Eisenhower administration for its “statist plans” and “attacking the idea that we should spend every drop of American blood supporting Chiang Kai-shek” (Rothbard 1981). Rothbard was against military interventions and conscription, and he favored the free market in all areas of

social cooperation. This is why he in fact considered himself to be an extreme version of the members of the “Old Right,” a movement that evolved in reaction to American entry into World War I (1917), and the New Deal (1933–1938), which Rothbard identified in his book *The Betrayal of the American Right*. The “Old Right” included authors and journalists, such as Albert Jay Nock (1870–1945), H. L. Mencken (1880–1956), Garett Garrett (1878–1954), Robert Taft (1889–1953), Frank Chodorov (1887–1966), and John T. Flynn (1882–1964) whose works played a crucial role in Rothbard’s intellectual development. It was not the only time that Rothbard was red-baited by people on the Right, since, although being an ardent critic of communism, he opposed the global military campaign against it. In his opinion, the battle should be won intellectually and not by means of violent force, and it should be fought against statism in general and not communism in particular. Rothbard even cooperated with members of the New Left in the 1960s looking for allies in his fight against war. For him, this became the most important issue of all. As he noted privately in 1956: “I am getting more and more convinced that the war-peace question is *the key* to the whole libertarian business” (as cited in Payne 2005). Rothbard appreciated the works of New Left historian William Appleman Williams (1921–1990) and in particular his foreign policy analysis and his advocacy of decentralization in domestic affairs. In retrospect, however, he thought that the alliance with the New Left was a mistake. He wrote:

We came to realize that, as Marxian groups had discovered in the past, a cadre with no organization and with no continuing program of “internal education” and reinforcement is bound to defect and melt away in the course of working with far stronger allies. The libertarian groupings would have to be rebuilt as a self-conscious movement, and its major emphasis would have to be on nourishing, maintaining, and extending the libertarian cadre itself. (Rothbard 2007, p. 202)

This is exactly what Rothbard did toward the end of his career. He published *For a New Liberty: The Libertarian Manifesto* in 1973, in which he applied libertarian philosophy to the social and economic problems of our time and thereby

founded the modern libertarian movement (Block and Rockwell 1988). His philosophical opus magnum *The Ethics of Liberty* appeared in 1982. In addition to his contributions in economic theory and philosophy, he wrote a four-volume history of colonial America, *Conceived in Liberty*, and a two-volume *Austrian Perspective on the History of Economic Thought*, tracing the history of economics from the ancient Greeks to the classical economists, such as Adam Smith (1723–1790) and David Ricardo (1772–1823), as well as the French school of classical liberalism and Marxism. He was working on a third volume on more recent developments at the time of his death. He did not finish it.

Rothbard only obtained a full professorship in 1986, as S. J. Hall Distinguished Professor of Economics at the University of Nevada, Las Vegas (UNLV), a university without a graduate program in economics, which meant that throughout his academic career Rothbard was prevented from claiming a single doctoral student as his own (Hoppe 1999). At UNLV, he closely collaborated with Hans-Hermann Hoppe, one of his most influential disciples, until his death on the 7th of January 1995. Despite his unconventional and radical beliefs, he probably could have obtained a full professorship earlier in his career. However, according to Gary North (2010), he suffered from phobia about leaving Manhattan Island, which he could only overcome toward the end of his life.

Academic Work and Influence

Murray N. Rothbard published his opus magnum in the area of economics, *Man, Economy and State*, in 1962, as a relatively young economist of only 36 years. With this volume, he occupied a very, if not the most, influential position within the tradition of the Austrian School of economics in the second half of the twentieth century – although Nobel Prize-winning economist Friedrich August von Hayek (1899–1992) is commonly considered to be the most important representative of this school of thought. However, Hayek is strictly speaking not an adherent to the rationalist mainstream of Austrian economics as espoused by Carl Menger (1840–1921), Eugen von Böhm-Bawerk

(1851–1914), Ludwig von Mises, and Murray N. Rothbard but rather its explicit opponent. He stands in the tradition of British empiricism and skepticism (Hoppe 1999). Like his intellectual forerunners, Rothbard is an “outspoken rationalist and critic of all variants of social relativism: historicism, empiricism, positivism, falsificationism, and skepticism” (Hoppe 1999, p. 223). Rothbard interpreted economic theory as universally and *a priori* true and not only hypothetically valid. In his opinion, it is not subject to constant empirical testing against data. In other words, propositions in economics concern non-hypothetical relations and assume apodictic validity. As Mises in his *Human Action*, Rothbard deduced all his economic propositions from basic axioms. He wrote in the preface to the second revised edition of *Man, Economy and State*:

The present work deduces the entire corpus of economics from a few simple and apodictically true axioms: the Fundamental Axiom of *action* – that men employ means to achieve ends, and two subsidiary postulates: that there is a *variety* of human and natural resources, and that leisure is a consumers’ good. (Rothbard 2009, p. lvi)

Rothbard interpreted economic theory as the logical deduction from the proposition that men act. It is impossible to deny this fundamental axiom without running into a performative contradiction, since arguing against it would imply an act as described above – the conscious employment of means, i.e., once vocal chords, to achieve an end, i.e., the refutation of the axiom of action. It is therefore indisputably true. The validity of economic theory, therefore, rests solely on the validity of the axiom of action and the correct exercise of logical deduction and inference (Hoppe 1999). Hence, empirical testing of economic propositions becomes unnecessary. Economic data can illustrate the propositions but neither verify nor falsify them. Furthermore, Rothbard followed the strict epistemological and methodological individualism of Menger, Böhm-Bawerk, and Mises, that is, he acknowledged the necessity to explain all economic phenomena in terms of purposeful individual action. Only individuals have desires, and “[e]very ‘holistic’ and ‘organistic’ explanation must be categorically rejected as an

unscientific pseudo-explanation” (Hoppe 1999, p. 224). Since human beings act under uncertainty, human action is always speculative, and therefore Rothbard considers mechanical explanations of social phenomena as inappropriate and unscientific. Human action is fundamentally different from the behavior of atoms and other objects considered in natural sciences.

Beyond Mises’s framework, Rothbard made several contributions to Austrian economics. One of the most important is his clarification of marginal utility theory as well as the reconstruction of welfare economics and the theory of the state that he derived from it. Rothbard explained that the word “marginal” does not refer to increments of utility, but rather to the utility of increments of a good. The former would imply measurability of utility, whereas the latter does not. Increments of a good can be characterized in physical terms and can be measured. Utility, however, is completely subjective, cannot be measured, and exhibits only an ordinal character on one-dimensional subjective preference scales. It is furthermore subject to changes over time. Consequently, interpersonal utility comparisons and the application of the rules of arithmetic to the concept of utility are impossible, such as adding utilities together to obtain a measure of “social welfare.” Rothbard took this logical conclusion from the subjective and ordinal character of utility seriously and developed a new version of welfare economics “based on the twin concepts of individual self-ownership and demonstrated preference” (Hoppe 1999, p. 228).

Self-ownership means that every person owns, or exclusively controls, his or her physical body. In every action the personal physical body serves as a means to achieve some end. Thereby, it is demonstrated that the physical body is valued as *good*, and furthermore by doing one thing rather than another, it is demonstrated what is deemed the most highly valued end. In other words, the underlying ordinal preferences are demonstrated. Trading good A against good B demonstrates that good A is ranked higher on the subjective preference scale. For the trading partner, the opposite is true. Both parties expect to benefit from the transaction; otherwise it would not take place. As long

as individuals act, trade, and cooperate voluntarily and do not harm third parties, we might speak of Pareto-superior changes that lead to increases in subjective utility and hence in “social welfare.” If transactions are brought about through coercion, at least one party is made worse off; otherwise it would have engaged in the transaction voluntarily. These changes are not Pareto superior. And it cannot be claimed that they lead to increases in social welfare, since losses and gains in utility cannot be compared. Noncontroversial examples would be criminal offenses, such as robbery and theft. However, also acts of the government, at least in part, classify as coercive. In fact, the power to coerce is the decisive characteristic of the state. Rothbard defined the state:

as that organization which possesses either or both (in actual fact, almost always both) of the following characteristics: (a) it acquires its revenue by physical coercion (taxation); and (b) it achieves a compulsory monopoly of force and of ultimate decision making power over a given territorial area. (Rothbard 2002, p. 172)

Rothbard’s conclusion is the refutation of the institution of government on welfare economic grounds. This of course made Rothbard an intellectual outlier, although his conclusion was implicitly accepted by many academics. As Hans-Hermann Hoppe points out:

Scores of political philosophers and economists, from Thomas Hobbes to James Buchanan and the modern public-choice economists, have attempted to escape from this conclusion by portraying the state as the outcome of contracts, and hence, a voluntary and welfare-enhancing institution. (Hoppe 1999, p. 231)

Rothbard, however, would agree with Joseph Schumpeter (1883–1950) that “the theory which construes taxes on the analogy of club dues or of purchase of services of, say, a doctor only proves how far removed this part of the social sciences is from scientific habits of mind” (Rothbard 1997, p. 247; as cited in Hoppe 1999, p. 231).

Rothbard’s monetary theory is also strongly influenced by the pioneering work of Ludwig von Mises. As he stated: “The Austrian theory of money virtually begins and ends with Mises’s monumental *Theory of Money and Credit*”

(Rothbard 1997, p. 297). However, Rothbard made some additions and in particular generalized the Misesian theory by adopting a broader definition of the supply of money – “money includes whatever is redeemable at par in standard money” (Gordon 2007, p. 39). Mises introduced the so-called regression theorem which states that money as a generally accepted medium of exchange originates as a commodity money, such as gold, in the course of voluntary exchange and cooperation on the market. Rothbard added a theory of the destruction or devolution of money by the government that Hoppe called a “progression theorem” (Hoppe 1999, p. 237). Rothbard (2010) is an exposition of this theory and his views on money accessible for laymen and the general public. Rothbard (2008) and *The Case against the FED* (1994) contain similar ideas.

Rothbard applied Austrian business cycle theory, as formulated by Mises (1912) and Hayek (1967), to the Great Depression of the 1930s in his 1963 book *America’s Great Depression*. Another source of influence for this volume was (Robbins 1971). In fact, Rothbard wrote in a letter to Ivan Bierly in 1959 that he considers Robbins’s book to be “one of the great economic works of our time... This is unquestionably the best work published on the Great Depression” (as cited in Gordon 2007, p. 42). Rothbard shows how the available empirical data supports the Austrian claim that artificial credit expansion leads to an inflationary boom, during which malinvestments are made, and the subsequent bust, in which the “cluster of business errors” becomes apparent (Rothbard 2000, p. 8). When central banks lower interest rates artificially, they make more investment projects look profitable than the subsistence fund, that is, the amount of real savings in the economy, can actually sustain. Hence, the structure of production is changed in an unsustainable manner as more investment projects are started than can be finished. The crisis constitutes the period, in which the necessary corrections take place, i.e., unprofitable investment projects are liquidated and, if not lost altogether, the remaining capital is reinvested into profitable lines of production. In Rothbard’s own words:

In sum, businessmen were misled by bank credit inflation to invest too much in higher-order capital goods, which could only be prosperously sustained through lower time preferences and greater savings and investment; as soon as the inflation permeates to the mass of the people, the old consumption–investment proportion is reestablished, and business investments in the higher orders are seen to have been wasteful. Businessmen were led to this error by the credit expansion and its tampering with the free-market rate of interest.

The “boom,” then, is actually a period of wasteful misinvestment. It is the time when errors are made, due to bank credit’s tampering with the free market. The “crisis” arrives when the consumers come to reestablish their desired proportions. The “depression” is actually the process by which the economy adjusts to the wastes and errors of the boom, and reestablishes efficient service of consumer desires. The adjustment process consists in rapid liquidation of the wasteful investments. Some of these will be abandoned altogether (like the Western ghost towns constructed in the boom of 1816–1818 and deserted during the Panic of 1819); others will be shifted to other uses. (Rothbard 2000, pp. 11–12)

The current financial and debt crisis exhibits the same tendencies again. The ghost towns around Madrid in Spain are only one dramatic case in point for the wasteful employment of capital in the recent past. The current crisis has led to an increased interest in the Austrian theory of the trade cycle among laymen and professional economists alike, both in order to support and to criticize it (see, e.g., Caplan 2008). It is argued that the Austrian theory has more explanatory power than mainstream Neoclassical, Monetarist, or Keynesian accounts of the current crisis, which makes it attractive for young economists, but also a challenge for economists of other schools of thought. Rothbard’s work and influence has played a major role in this development.

Rothbard was cofounder of the *Ludwig von Mises Institute* in Auburn, Alabama, together with Llewellyn H. Rockwell Jr. (born 1944) and Burton Blumert (1929–2009). Today, the institute promotes the ideas of Ludwig von Mises and Murray N. Rothbard as well as their intellectual followers more effectively than ever before. Its website, www.mises.org, makes a large number of books, journal articles, and other writings available for free. There exist a number of professional

journals and periodicals published by the institute, including *The Journal of Libertarian Studies* (1977–2008) and *The Review of Austrian Economics* (1987–1998), for both of which Rothbard served as editor, as well as *The Quarterly Journal of Austrian Economics* (since 1998), which was started after Rothbard’s death. Many of the contributors to the journal see themselves in line with Rothbardian economics. Rothbard himself, however, “by 1963 had grown discontented with economic analysis as an end in itself” (as cited in Casey 2010, p. 7) according to Gary North. Through his many radical writings in the area of libertarian ethics and philosophy as well as his political activism, for example, in support of the Libertarian Party (since 1971), Rothbard launched and spearheaded the modern libertarian and anarcho-capitalistic movement in the United States. Today, it has found followers all around the world.

References

- Block W, Rockwell LH Jr (1988) Man, economy and liberty – essays in Honor of Murray N. Rothbard, Ludwig von Mises Institute. Available online: <http://mises.org/document/3073/Man-Economy-and-Liberty-Essays-in-Honor-of-Murray-N-Rothbard>
- Caplan B (2008) What’s wrong with Austrian business cycle theory. Library of Economics and Liberty, London and New York. Available online: http://econlog.econlib.org/archives/2008/01/whats_wrong_wit_6.html#
- Casey G (2010) Murray Rothbard, volume 15 in the series major conservative and libertarian thinkers. The Continuum International Publishing Group, London and New York
- Eltzbacher P (1908) Anarchism. Benj. R Tucker, New York. Available online: <http://www.gutenberg.org/ebooks/36690>, and in the original German version: <https://archive.org/details/deranarchismus00eltzgoog>
- Flood A (2008) Murray Newton Rothbard: notes toward a biography, published on personal webpage of the author. Available online: <http://www.anthonyflood.com/murrayrothbardbio.htm>
- Gordon D (2007) The Essential Rothbard. Ludwig von Mises Institute. Available online: <http://mises.org/document/3093/The-Essential-Rothbard>
- Hoppe H-H (1999) Murray N. Rothbard: economics, science, and liberty. In: Holcombe RG (ed) 15 Great Austrian economists. Ludwig von Mises Institute. Available online: <http://mises.org/document/3797/15-Great-Austrian-Economists>

- North G (2010) Murray Rothbard as academic role model, presentation delivered at the 2010 Mises University. Available online: <http://www.youtube.com/watch?v=0tafUj6WyA>
- Payne J (2005) Rothbard's time on the left. *J Liber Stud* 19:9. Available online: https://mises.org/journals/jls/19_1/19_1_2.pdf
- Robbins L (1971) *The great depression*. Books for Libraries Press, Freeport New York. Available online: <http://mises.org/document/691/The-Great-Depression>
- Rothbard MN (1962) *The panic of 1819: reactions and policies*. Columbia University Press. Available online: <http://mises.org/document/695/The-Panic-of-1819-Reactions-and-Policies>
- Rothbard MN (1981) How Murray Rothbard became a libertarian, talk given at the libertarian party national convention in Denver, Colorado. Available online as video: <https://www.youtube.com/watch?v=M3HatEn0BjI>, and as transcript: <http://bastiat.mises.org/2014/04/transcript-how-murray-rothbard-became-a-libertarian/>
- Rothbard MN (1994) *The case against the FED*, Ludwig von Mises Institute. Available online: <http://mises.org/document/3430/The-Case-Against-the-Fed>
- Rothbard MN (1997) *The logic of action*, 2 vols. Edward Elgar, Cheltenham
- Rothbard MN (2000) *America's great depression*, 5th edn. Ludwig von Mises Institute. Available online: <http://mises.org/document/694/Americas-Great-Depression>
- Rothbard MN (2002) *The ethics of liberty*. New York University Press, New York
- Rothbard MN (2006) *Making economic sense*. Ludwig von Mises Institute. Available online: <http://mises.org/document/899/Making-Economic-Sense>
- Rothbard MN (2007) The later 1960s: the new left. In: Woods TE Jr (ed) *The betrayal of the American right*. Ludwig von Mises Institute, Auburn, AL. Available online: <http://mises.org/document/3316/The-Betrayal-of-the-American-Right>
- Rothbard MN (2008) *The mystery of banking*. Ludwig von Mises Institute. Available online: <http://mises.org/document/614/Mystery-of-Banking>
- Rothbard MN (2009) *Man, economy, and state, with power and market*. Ludwig von Mises Institute. Available online: <http://mises.org/document/1082/Man-Economy-and-State-with-Power-and-Market>
- Rothbard MN (2010) *What has government done to our money?* Ludwig von Mises Institute. Available online: <http://mises.org/document/617/What-Has-Government-Done-to-Our-Money>
- Stigler G, Friedman M (1946) *Roofs or ceilings?: the current housing problem*. Foundation for Economic Education, Irvington-on-Hudson
- von Hayek FA (1967) *Prices and production*, 2nd edn. Augustus M. Kelly, New York. Available online: <http://mises.org/document/681/Prices-and-Production>
- von Mises L (1912) *Theorie des Geldes und der Umlaufmittel*. Verlag von Duncker und Humblot, München und Leipzig. Available online: <http://mises.org/document/3298/Theorie-des-geldes-und-der-Umlau>
- [fsmittel](http://mises.org/document/194/The-Theory-of-Money-and-Credit); for the English translation see: <http://mises.org/document/194/The-Theory-of-Money-and-Credit>
- von Mises L (1962) *A new treatise on economics*. *New Individ Rev* 2(3):39–42
- von Mises L (1998) *Human action* (The Scholar's Edition). Ludwig von Mises Institute. Available online: <http://mises.org/document/3250/Human-Action>; for the German predecessor *Nationalökonomie* see: <http://mises.org/document/5371/Nationalökonomie-Theorie-des-Handelns-und-Wirtschaftens>

Rueff, Jacques

Adrien Lutz

GATE L-SE, Université Jean Monnet,
Saint-Étienne, France

Abstract

Jacques Rueff was a French economist with long experience in combining practice and theory. Deeply interested in epistemology and interdisciplinarity, Rueff was a monetary specialist and one of the leading European economists after World War II.

Biography

After studying at the École Polytechnique, **Jacques Rueff** (1896–1978) became an inspector of finance (1923–1926), special advisor to the League of Nations (1927–1930), financial attaché at the French Embassy in England in charge of the Bank of France's sterling reserve (1930–1933), deputy-director of the Mouvement général des fonds (previously the French Trésor; 1934–1939), and deputy governor of the Bank of France (1939–1941). Rueff also participated in the Walter Lippmann colloquium in 1938 (Lane 1997).

Until 1952, Rueff was deeply involved in the postwar negotiations (he was, for instance, the president of the inter-allied reparations agency). A strong advocate of European integration, Rueff was a judge of the Court of Justice of the European Communities (1952–1962). He also became an important economic advisor to

President Charles de Gaulle from 1958, and it was thanks to his plan (with Pinay) that France succeeded in balancing the budget and ensuring the convertibility of the franc. A proponent of a return to the gold standard, Rueff argued against inflation which he always considered a “false right.” He later became a member of the Economic and Social Council.

Innovative and Original Aspects

Deeply interested in epistemology and interdisciplinarity, Rueff adopted a conventionalist methodology (Frobert 2009, see also Claassen 1967), applying this to topics such as unemployment (1925b, 1926, 1931a, b) and money (1927, 1953, 1961, 1963, 1965, 1971, 1973), and being among the first to consider economics as a statistical science (1925a, [1929] 1961). Surprisingly, although his earliest work did not directly express an interest in law (1922), many legal theorists commented on his work long before he became famous for his concept of the “false right” (1945).

The English translation of Rueff’s *From the Physical to the Moral Sciences* ([1922] 1929) was widely discussed in American journals of law and philosophy. According to Rueff, economics is a moral science, all physical and moral sciences use the same tools, and all have a rational branch and an experimental branch. Although their objects, of course, differ, the aim remains always to discover the underlying laws of observed phenomena. Ironically, in this book, he never addressed the questions of law.

Practice and learning in American law had come under strong criticism in the nineteenth century, with particular emphasis on the quasi-mechanical application of its various laws and codes. Lawyers concerned with these issues turned to Rueff’s work on economics and morality for an account of how a social science could be considered a “real science” like physics, constrained by the adoption of a scientific method. Discussion arose concerning what might be the

scientific basis of the concept of law, whether social norms and the evolution of society should be taken into account in attempting to define that concept, and if so what methodology ought to be adopted.

Thus, a preface to Rueff’s work was written by H. Oliphant (1884–1939) and A. S. Hewitt (1902–1987), teachers of law at Columbia and Johns Hopkins University respectively, and fervent advocates of interdisciplinarity particularly with respect to law and economics. From this preface, it was clear that Oliphant and Hewitt accepted that law could be a science on the condition that it applied a rigorous method, allowing case studies. Specifically, Oliphant (1923, 1928) argued in favor of a rapprochement of law and economics, since both aimed to establish a method that combined logic and experience, just as outlined by Rueff. According to them, when a society evolves, this means in terms of logic that one of its first premises has changed. For instance, in several countries, the abolition of the death penalty had only recently become conceivable, and abolition had not been discussed at all in their earlier history. It was therefore necessary to modify the legal texts in such a way as to follow the social changes. These American lawyers lacked a theoretical corpus, however, and this was why Rueff was so widely discussed. His incursions concerning morality strongly interested them, particularly because, for Rueff, manners and laws are valid only for a time.

Impact and Legacy

Law and economics become combined in Rueff’s famous concept of the “false right.” Rueff’s starting point is the concept of the “property rights,” which is rooted in the 544th article of the French civil code: “ownership is the right to enjoy and dispose of things in the most absolute manner, provided they are not used in a way prohibited by statutes or regulations.” Rueff’s *Social Order* (1945) argues that economics is not the science of wealth as has generally been thought, but

rather the science of the links between desired things and men's desiring those things. Wealth does not lie in the things desired but in our ability to enjoy and dispose of things.

As recently remarked by Minart (2016), according to Rueff the lawyer gives shape to the right to property, the economist provides its content, and the police its protection. The right to property is analyzed as a recipe for value. In this respect, a real right is a right that enables a seller to "empty" the content of his right to property "onto" a specific commodity and, in return, to "fill" his right with money. The buyer and the seller agree on the price, and then the right becomes real. And, to the contrary, a false right occurs when the value of wealth contained in this right does not comply with the amount of money the seller wishes to receive – in the case, for instance, of prices being fixed by a public authority.

According to Rueff, prices determine the volume of the right. If the selling price is lower than the purchase price, the volume of the right decreases. There follows an imbalance between supply and demand, and thus the creation of false rights. Such an imbalance is the basis of inflation. The same is also true for unemployment insurance, since it interferes with the free functioning of the price mechanism.

Having successfully combined theory and practice throughout his life, Rueff has made significant contributions to the field of economic analysis through his monetary analysis and his innovative false rights theory. Jacques Rueff remains in many respects one of the major French economists of the twentieth century.

Cross-References

- [Economic Analysis of Law](#)
- [Law and Economics](#)
- [Property Rights: Limits and Enhancements](#)

References

Claassen EM (1967) Jacques Rueff comme philosophe des sciences. In: Claassen EM (ed) *Les fondements*

- philosophiques des systèmes économiques. Textes de Jacques Rueff et essais rédigés en son honneur. Payot, Paris, pp 37–62
- Fröbert L (2009) Conventionalism and liberalism in Jacques Rueff's early works (1922 to 1929). *European Journal of History of Economic Thought* 17(3):439–470
- Lane G (1997) Jacques Rueff: un libéral perdu chez les planistes. Conférence à Euro 92. In: Madelin A (ed) *Aux sources du modèle libéral français*. Perrin, Paris, pp 1–23
- Minart G (2016) Jacques Rueff. Un libéral français. Odile Jacob, Paris
- Oliphant H (1923) The problem of logical methods from the lawyers's point of view. *Proceedings of the Academy of Political Science in the City of New York* 10(3):19–20
- Oliphant H (1928) A return to stare decisis. *American Bar Association Journal* 14(2):71–158
- Rueff J (1922/1929) From the physical to the social sciences. The Johns Hopkins Press, Baltimore
- Rueff J (1925a) L'économie politique, science statistique. *Rev Metaphys Morale* October:475–487
- Rueff J (1925b) Les variations du chômage en Angleterre. *Revue Politique et Parlementaire* December, 32: 425–437
- Rueff J (1926) Chômage et salaire réel en Angleterre dans la période 1919–1925. Séance de la Société d'Économie Sociale du 14 décembre 1925. *La Réforme Sociale* February:49–73
- Rueff J (1927) Théorie des phénomènes monétaires. *Œuvres complètes, II, Théorie monétaire, 1*. Plon, Paris
- Rueff J (1929/1961) La statistique, instrument de la connaissance. *Journal de la Société Statistique de Paris* 102:220–224
- Rueff J (1931a) L'Assurance chômage cause du chômage permanent. *Revue d'Économie Politique* 45:211–251
- Rueff J (1931b) L'assurance chômage, cause du chômage permanent. In: *Académie des sciences morales et politiques, 1931, Séances et travaux de l'Académie des sciences morales et politiques: Compte rendu, 2^e semestre*. Alcan, Paris, pp 300–331
- Rueff J (1945) *L'ordre social*. Sirey, Paris
- Rueff J (1953) *La régulation monétaire et le problème institutionnel de la monnaie*. Recueil Sirey, Paris
- Rueff J (1961) *Discours sur le crédit*. Éditions du Collège libre des sciences sociales et économiques, Paris
- Rueff J (1963) *L'âge de l'inflation*. Payot, Paris
- Rueff J (1965) *Le lancinant problème des balances de paiements*. Payot, Paris
- Rueff J (1971) *Le péché monétaire de l'Occident*. Plon, Paris
- Rueff J (1973) *La réforme du système monétaire international*. Plon, Paris

Rule of Law

Rosolino A. Candela¹ and Ennio Piano²

¹Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

²Department of Economics, George Mason University, Fairfax, VA, USA

Abstract

This entry explores the role of the rule of law in a market economy. The emergence of the rule of law is a *precondition* of not only political and economic liberty but also economic growth and overall human progress. The presence, or lack thereof, of the rule of law determines whether human interaction will be positive-sum or negative-sum in nature. We explore first the fundamental attributes to the rule of law in relation to property rights and entrepreneurship. Second, we discuss in detail three attributes to the rule of law that generate this productive entrepreneurial process: (1) generality, (2) predictability, and (3) equality.

Introduction

Simply put, the rule of law refers to the absence of political or legal privilege among market actors *and* the absence of arbitrary discretion among political actors. It is a political–legal principle, whereby the governing authority of a particular society is restricted to enforcing laws applied equally to all *and* not intended to benefit one particular party at the expense of another. Any violation of the rule of law implies that political–legal privileges cannot be granted without simultaneously granting discretionary power to those political actors who are in the position to grant such privileges. Therefore, the presence, or lack thereof, of the rule of law determines whether human interaction will be positive-sum or negative-sum in nature.

The Rule of Law, Property Rights, and Entrepreneurship

According F.A. Hayek, one of the leading economists and legal scholars of the twentieth century, “the gradual transformation of a rigidly organised hierarchic system into one where men could at least attempt to shape their own life, where man gained the opportunity of knowing and choosing between different forms of life, is closely associated with the growth of commerce” (1944 [1994]: 18). Such a gradual transformation can be framed in terms of an inherent link between the rule of law, property rights, and entrepreneurship. The presence, or lack thereof, of economic development cannot be explained by an abundance or shortage of entrepreneurship in a society (see Baumol 1990; Boettke and Piano 2016); individuals are always and everywhere on the lookout for previously unnoticed profit opportunities (Kirzner 1973). However, whether or not such a pursuit of profit opportunities manifests itself as productive activities, such as trade and innovation, or unproductive activities, such as rent-seeking and theft (Tullock 1967), depends upon the allocation of entrepreneurial talent in a society (Murphy et al. 1991). Such an allocation is dependent upon the establishment of well-defined and well-enforced private property rights that can be exchanged via money prices, allowing entrepreneurs to calculate the relative scarcity of resources through the guiding signals of profit and loss. Such guiding signals communicate to entrepreneurs whether or not resources have been allocated to their most valued consumer uses (Mises 1920 [1975]). Both the productive allocation of entrepreneurship and the productive allocation of resources via exchange and innovation that follows from entrepreneurship require that property rights restrict competition only to profit-seeking that creates social wealth, not rent-seeking that destroys social wealth (see Boettke and Candela 2014b). Such a restriction, however, implies the presence of the rule of law, without which entrepreneurship will be negative-sum rather than positive-sum. Following Hayek (1960: 205–210), there are three attributes to the rule of law that generate this productive entrepreneurial process: (1) generality,

(2) predictability, and (3) equality. We discuss each of these attributes below in relation to property rights and entrepreneurship.

Generality

To understand the way in which laws are general and abstract, rather specific and concrete, it is important first to make a distinction, following James Buchanan, between the rules of the game and interaction within the rules of the game. The distinction we make between the rules of the game and social interaction within rules follows a “law and economics” approach (Marciano 2016; see also Wagner 2016). Law and economics analyzes, in a world of positive transactions costs, the degree to which different legal rules, particularly the absence or existence of the rule of law, affect economic performance, namely by ameliorating or exacerbating the costs of defining property rights and facilitating exchange. This approach is related, though distinct, to “the economic analysis of law.” Although the terms are sometimes used interchangeably, the economic analysis of law approaches legal analysis by assessing the efficiency of legal rules in a world of zero-transactions costs.

The attribute of generality implies that the rule of law operates as a “meta-legal doctrine” (Hayek 1960: 206), through which the types of laws that are filtered through the legislative process are those that are *end-independent* and *impersonal*. Laws that are consistent with the rule of law must neither command any specific purpose upon individuals nor does it assign any concrete status or outcome that differentiates individuals before the law. This does not mean that laws that recognize individuals on the basis of sex, race, or creed violate the rule of law; rather, it is only when such groups are preassigned a special status or privilege at the expense of other individuals that the rule of law is violated. This violation occurs in a twofold manner. First, resources and income are transferred by force to the politically privileged to the expense of the politically disenfranchised. For example, import tariffs on particular goods lead to the loss of consumer surplus and generate rents for

domestic producers of that good. Secondly, and more importantly, the political transfer from one party to another cannot occur without simultaneously granting discretion to the political official, who must exercise arbitrary force in making such a transfer. When the rule of law is violated, the legislator is no longer “blind,” but can in fact foresee how and to whom laws will effect particular groups of individuals in society. As a result, entrepreneurship in such a society will mutate from profit-seeking to rent-seeking, as entrepreneurs will expend resources to capture privileges through political exchange, which wastes resources that could have otherwise been used for productive innovation (Harnay and Marciano 2011).

Predictability

The rule of law does not assign marching orders to individuals but acts like traffic signals that guide the interaction of individuals in a noncoercive manner. Like traffic signals, they do not rely on legislators directing any other individual in advance, by force, or how to act explicitly. Rather, because individuals can reliably expect that laws will not be changed arbitrarily, “planning” is left to the individual at a particular place and time. Predictability, however, implies neither the complete lack of change in the law nor the absence of a judiciary. Rather, it only implies that laws will be changed on the margin by judges to facilitate coordination between contesting parties according to the circumstances of place and time. Such evolution in the law will be marginal, but adaptive for the adjudication of new disputes, yet consistent with the body of judicial precedent. Therefore, if rule of law is in place, entrepreneurs can reliably feel secure in their property and be able to deploy their resources in long-term investment projects that are productive and wealth-creating. Predictability means laws are intended for such a duration that no one, not even the legislator, can anticipate who will benefit from such laws. It is in this sense that the rule of law operates as a “fifth factor of production” (Boettke and Candela 2014a), providing the framework for

the coordination of the land, labor, and capital by entrepreneurs according to their foresight about future profit opportunities. This has important implications for entrepreneurship and economic development. First, entrepreneurs in the marketplace can reliably “predict,” or anticipate, that they will only be harmed by the threat of losses from competing entrepreneurs, without fear of being harmed by the threat of public predation in the form of outright theft or regulation that bars them from entry into the marketplace. The unintended consequence of the rule of law, then, is to discipline entrepreneurs to accrue profits only by innovating, lowering costs, and producing resources according to consumer preferences. When the rule of law is violated, the only “predictability” that entrepreneurs have in their ability to accrue profits is to bar their competition from entry into the marketplace via capture of monopoly privileges.

Equality

The link between the predictability of laws and the spontaneity of the market process is the attribute of equality before the law. Under the rule of law, laws are applied equally to all, both to the ruler and the ruled. Social interaction and the outcomes of such interactions are guided horizontally by voluntary contract between individuals within the game, not vertically by capturing legal status or manipulating political officials to “bend the rules” in an individual’s favor. Thus, under the rule of law, one’s relative wealth or status in society is not predetermined by privilege; rather, it is *discovered* by learning how best to serve their fellow man via contractual exchange and innovation. In this capacity, government only acts as a referee, or umpire, to enforce the law, setting the precondition within which not one but multiple purposes and plans can be pursued by individuals via trade and exchange. Equality before the law is crucial to entrepreneurship and economic development because it recognizes that individuals are different and unique in their talents, potentialities, and knowledge of a particular time and place. It is freedom from legal discrimination that allows the

spontaneous interaction of diverse individuals to generate entrepreneurial discoveries that are accidental and previously unforeseen. How could Malcolm McLean have applied his unique knowledge of trucking and shipping to pioneer the container ship in the 1950s if he was legally barred by shipping regulations? (see Levinson 2006) The rule of law cannot be violated without the tendency of substituting the rule of men. Under the rule of law, profits and losses in the marketplace are determined by the ability of entrepreneurs to serve consumers. Under the rule of men, profit and losses are determined by the ability of entrepreneurs to win political privileges, which come at the expense of the consumer.

Cross-References

- ▶ [Austrian Perspectives in Law and Economics](#)
- ▶ [Austrian School of Economics](#)
- ▶ [Common Law System](#)
- ▶ [Customary Law](#)
- ▶ [Efficiency](#)
- ▶ [Hayek, Friedrich August von](#)
- ▶ [Liberty](#)
- ▶ [Mises, Ludwig von](#)
- ▶ [Rule of Law and Economic Performance](#)

References

- Baumol WJ (1990) Entrepreneurship: productive, unproductive, and destructive. *J Polit Econ* 98(5): 893–921
- Boettke PJ, Candela R (2014a) Hayek, Leoni, and law as the fifth factor of production. *Atl Econ J* 42(2):123–131
- Boettke PJ, Candela RA (2014b) Development and property rights. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York
- Boettke PJ, Piano E (2016) Baumol’s productive and unproductive entrepreneurship after 25 years. *J Entrep Public Policy* 5(2):130–144
- Harnay S, Marciano A (2011) Seeking rents through class actions and legislative lobbying: a comparison. *Eur J Law Econ* 32(2):293–304
- Hayek FA (1944 [1994]) *The road to Serfdom*, 50th anniversary edition. University of Chicago Press, Chicago
- Hayek FA (1960) *The constitution of liberty*. University of Chicago Press, Chicago

- Kirzner IM (1973) *Competition & entrepreneurship*. University of Chicago Press, Chicago
- Levinson M (2006) *The box: how the shipping container made the world smaller and the world economy bigger*. Princeton University Press, Princeton
- Marciano A (2016) Economic analysis of law. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York
- Mises L (1920 [1975]) Economic calculation in the socialist commonwealth. In: Hayek FA (ed) *Collectivist economic planning*. August M. Kelley, Clifton
- Murphy KM, Shleifer A, Vishny RW (1991) The allocation of talent: implications for growth. *Q J Econ* 106(2):503–530
- Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. *West Econ J* 5(3):224–232
- Wagner RE (2016) Gordon Tullock: a Maverick scholar of law and economics. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York

Rule of Law and Economic Performance

Rosolino A. Candela¹ and Ennio Piano²

¹Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

²Department of Economics, George Mason University, Fairfax, VA, USA

Abstract

This entry explores the connection between the rule of law and economic performance. First, we analyze the role of interjurisdictional competition, and how the rule of law emerged as a by-product of this competitive process, initially in Western Europe. Second, we discuss the role of intrajurisdictional competition between interest groups, which reinforced the emergence of the rule of law within states. Finally, we discuss the mechanism by which the rule of law successfully or unsuccessfully became established in Western offshoots, particularly in Africa and the Americas, affecting the long-term economic performance of these areas.

JEL Codes

O12; P14; P16

[W]e may say that the movement of the progressive societies has hitherto been a movement *from Status to Contract*—Henry Sumner Maine (emphasis original, 1861 [1982]: 170)

Introduction

The most incredible fact in recent economic history is the rapid decline of extreme poverty throughout the developing world. According to the World Bank, the percentage of individuals living on less than two dollars per day has fallen from 37% in 1990 to less than 10% of the world's population by the end of 2015. This dramatic fall in poverty has been due to a turn towards the protection of private property, freedom of contract, and most importantly, the establishment of the rule of law during this same period, particularly in areas that abandoned central economic planning, such as China, India, and Central and Eastern Europe. This more recent phenomenon of modern economic growth is a part of a progressive transition in societies that began first in Western Europe, from one in which wealth is accumulated through political transfers of privilege to a politically connected few, to one in which wealth is created through voluntary market exchange and innovation among many anonymous strangers. This social and economic transition is fundamentally based on an *institutional transition* towards the emergence and establishment of the rule of law, or as Maine put it, “*from Status to Contract*.”

However, the importance of the rule of law is not simply due to economic development in terms of increasing growth rates (Ručinská et al. 2016). Wealth creation produced by productive entrepreneurship within the rule of law is necessary, but not sufficient for explaining human progress. Boettke and Subrick (2003) find that the rule of law also positively affects human capabilities, which refers to ability of individuals to lead the sort of lives that they value. That is, a variety of *noneconomic* fundamental aspects of the human existence, including life expectancy, child mortality, literacy rates, and immunization, are positively related with the protection the rule of law. Therefore, the rule of law is good for

development, and development is not just good for one's stomach but also one's mind and soul.

This entry explores the connection between the rule of law and economic performance. First, we analyze the role of interjurisdictional competition, and how the rule of law emerged as a by-product of this competitive process, initially in Western Europe. Second, we discuss the role of intra-jurisdictional competition between interest groups, which reinforced the emergence of the rule of law within states. Finally, we discuss the mechanism by which the rule of law successfully or unsuccessfully became established in Western offshoots, particularly in Africa and the Americas, affecting the long-term economic performance of these areas.

The Emergence of the Rule of Law Via Interjurisdictional Competition

The degree to which the rule of law will lead to economic development depends upon whether or not the discretionary hands of a ruler are tied from altering property rights arbitrarily. Otherwise, entrepreneurship will be directed into unproductive activities. According to North and Weingast (1989: 804), "A ruler can establish such commitment in two ways. One is by setting a precedent of 'responsible behavior,' appearing to be committed to a set of rules that he or she will consistently enforce. The second is by being constrained to obey a set of rules that do not permit leeway for violating commitments." Beginning in a world in which governments have not established a credible commitment to predate its subjects, the question, then, is how did the rule of the law first emerge in the West?

Weingast (1997: 245) attempts to explain "the remarkable variation among states in the rule of law." In his framework, the rule of law is understood as "a set of stable political rules and rights applied to all citizens," one among many possible equilibria in a political-economic game between the government and the different subsets of the citizenry. Governments are often tempted to act arbitrarily to their own advantage. The citizens as a collective have an interest in constraining the

arbitrary actions of the government. But any subset of the population will sometimes gain something from these violations of the rule of law.

The logic of the classic prisoners' dilemma game applies. As evidence of this, historically, most societies have been trapped in a subpar equilibrium. Members of society disregard (and sometimes encourage) the violation of the rule of law whenever it benefits them. Escaping the subpar equilibrium requires coordination among citizens. Oftentimes, this can be so costly to being unfeasible. To make the rule of law a self-enforcing constraint on governmental interference with society, the interests and beliefs of large sections of the citizenry must be aligned, at least to some degree. Thus, societies with high degrees of fractionalization (ethnic, cultural, political) are less likely to coordinate than more homogeneous ones (Weingast 1997).

These obstacles notwithstanding, coordination is possible. The rule of law, as it emerged first in Western Europe, was an emergent phenomenon of human action, though not of human design. Given the politically fragmented nature of Western Europe, violent international competition between states required European political leaders to finance the maintenance of strong military forces against the threat of conflict. In order to finance military expenditures, such political leaders were incentivized to expand their tax revenue base in a wealth-maximizing manner, the unintended result of which was for the political elite of each state to attract merchants, bankers, and technological innovators by securing property rights in their resources. The effect of this institutional change was to encourage the development of capital markets capable of mobilizing large concentrations of wealth from which to tax and borrow. Moreover, in order for the political elite to have tapped this economic potential, this required not only security from arbitrary expropriation and confiscation but also toleration of technological experimentation and new ideas that would yield savings and investment. These institutional changes facilitated the expansion of technological innovation and productive specialization under the division of labor. As an unintended consequence of such action, feudalistic privileges

were gradually eroded and economic and political liberty was reinforced until a critical threshold was met, leading to an explosion of economic growth in the early nineteenth century (Rosenberg and Birdzell 1986; Cowen 1990).

The Emergence of the Rule of Law Via Intrajurisdictional Competition

The rule of law first emerged not only via interjurisdictional competition between competing states but also through intrajurisdictional competition between competing interest groups. North et al. (2009) provide a generalized version of this framework, with a stronger focus on the role of the elites. The fundamental puzzle driving their investigation is why the elites would ever extend their rights and liberties to the public at large. Historically, the natural state of human societies has been one dominated by relatively small groups or coalitions of a few large interest groups. These interest groups have the ability of using violence to appropriate wealth from the rest of the population. But as conflict is potentially very costly even for the stronger parties, the interest groups (and the subjects as well) face a strong incentive to institutionalize wealth extraction to minimize the expected costs from conflict. The resulting coalition of interests must therefore generate enough rents for each of its members to prevent defection. Legal monopolies and other special privileges are the way in which these rents are generated (Olson 1982). The public at large suffers from the resulting welfare losses, but it is actually better off than the more likely alternative of a civil war.

To operate effectively, the rent-seeking coalition needs a set of rules to govern the interactions among its members. The resulting norms are the seeds of the modern rule of law. Thus, the government will enforce property rights and contracts by providing institutions for the resolution of disputes between coalition members. Once these institutions have been established, society is on the brink of what the authors call “open access order,” or a rule of law regime. In a pacified society, where all gains from trade within the coalition have been exhausted, there is a potential

benefit from extending these institutions to a broader subset of the population. The coalition can credibly commit not to reverse the process by decentralizing political power (Weingast 1995) or extending political enfranchisement (Acemoglu and Johnson 2005).

The classic example is post-Glorious Revolution England (North and Weingast 1989; Acemoglu and Robinson 2012). The seeds for the self-enforcement of the English constitution after the Glorious Revolution can be traced back to the reissue of the Magna Carta in 1225. Leeson and Suarez (2016: 43) emphasize how the eventual self-enforcement of the Magna Carta required three conditions: (1) common knowledge among groups of citizens about when the ruler has violated the constitution; (2) rendering it in the interest of each group of citizens to rebel in order to enforce constitutional constraints on the ruler; and (3) creating a shared expectation among those groups that the others will rebel if the ruler violates the constitution, which in turn makes it rational for each such group to itself rebel in this event. These conditions also apply to explain the success of the Glorious Revolution in England.

Following the short interlude of Cromwell’s republican regime, the English parliament had restored the Stuart dynasty to lead the country. Within the next few years, the Crown took a series of actions to undermine the interests of the reformist and constitutionalist Whig party. Finding these actions to their own advantage, the Tories (the members of the conservative and monarchist party) aligned themselves with the Crown. This lasted until the 1680s, when the Stuarts started to take similar actions against the Tories. In 1688, the two parties coordinated their efforts, dethroned James II, and establish the Prince of Orange as the new monarch under a new constitution agreed upon by both parties. The Glorious Revolution resulted in the establishment of the rule of law and had other positive economic results. By establishing a permanent role for Parliament in the management of the government, it directly checked the discretion of the Crown to call and disband Parliament unilaterally. Moreover, parliamentary veto over expenditures, the right to monitor the expenditure of funds, and

the established supremacy of common law courts assured the protection of private property rights from expropriation and discretion.

The Colonial Origins of the Rule of Law

In the previous sections, we discussed the endogenous formation of the rule of law through competition between states and competition between interest groups within states. However, the literature on the colonial origins of economic performance illustrate where and how the rule of law can be exogenously established. The empirical research on the relationship between rule of law and economic performance falls under the umbrella of development economics. Building on the theoretical insights of Smith (1776), Hayek (1960), North (1990), and others, by the early 2000s economists started paying attention to the role of institutions as determinants of economic development. This institutional paradigm emerged in response to the (at the time) influential view that geography was a major factor in determining a country long-run economic prospects (Diamond 1997; Sachs and Warner 1997; Gallup et al. 1999). The first major contribution to this literature is a famous paper by Acemoglu, Johnson, and Robinson on the “colonial origins” of economic development (Acemoglu et al. 2001). Here, the authors attempt to disentangle and identify the causal effect of geographical and institutional factors on long-run economic growth. To do so, they take advantage of a natural experiment of history: the European colonization of Africa and the Americas.

According to their story, the mortality rate of early settlers influenced the adoption of institutions by the colonial governments. Where settlers’ mortality was low, colonial governments encouraged the immigration of individuals from the homeland. To move to the colonies, European settlers demanded the adoption of a set of institutional measures associated with the rule of law (the protection of basic human rights and private property and some form of political representation). In other regions, higher settlers’ mortality prevented similar migratory patterns. Dealing

with a mostly indigenous population, colonial governments adopted political and economic institutions aimed at the extraction of resources (including indigenous labor’s services).

Relying on the sluggish nature of institutional change, and using settler’s mortality as an exogenous determinant of institutional diversity, the authors find the quality of institutions adopted by colonial governments is a strong predictor of today’s economic performance. Furthermore, the authors find that, controlling for institutional quality, geographical factors usually assumed to play a causal role in the dynamics of development such as distance from the equator, have no *independent* effect on long-run economic growth. According to Acemoglu et al. (2001), geography affects economic performance only through their effect on institutions.

In a follow-up to their original article, the same authors adopt a different empirical strategy to reach similar conclusions (Acemoglu et al. 2002). Here, urbanization and population density at the arrival of the Europeans take the place of settlers’ mortality rate as instruments for institutional quality. A higher population density at time of discovery provided colonial governments with a pool of disposable and cheap labor to be exploited, while less inhabited areas required the influx of large numbers of workers from the homeland. In the former case, the colonists adopted the indigenous extractive institutions where they were already in place and made-up their own otherwise. In the latter case, they imported private property and constitutional government. This exogenously determined institutional diversity, the authors find, accounts for the puzzling case of the “reversal of fortune” of the last five hundred years: those extra-European regions that were the richest (poorest) before colonization are today among the richest (poorest) in the world. Glaeser et al. (2004) and La Porta et al. (2008) provide an alternative (but similarly influential) empirical approach to the same question. In their strategy, the initial institutional variation across countries is due not to the decision of formal political institutions, but the prevalent legal regime in the colonists’ home country. They separate between (English) common law

on the one hand and a variety of “families” of civil law (French, German, Scandinavian, and Socialist). The authors find that, in general, common law countries outperform civil law ones on most measures of institutional quality that are strongly positively correlated with economic performance, including constraints over governmental arbitrary powers, quality of public officials, property right protection, and so forth. All these findings are consistent with the traditional understanding of the beneficial nature of the rule of law: the closer a country’s institutions to the ideal of the rule of law, the more effective these institutions and the more dynamic and productive the economy.

Conclusion

Private property rights and entrepreneurship are necessary, though not sufficient for economic development. Given the scarcity of resources, property rights and entrepreneurship are ubiquitous, but their manifestation is ultimately contingent on the presence of the rule of law. Property rights over resources can be acquired through productive entrepreneurship via market exchange (i.e. profit-seeking) or through unproductive entrepreneurship via political exchange (i.e. rent-seeking). Fundamentally, the rule of law is *the* institutional filter that restricts the exchange of private property and productive entrepreneurship to the coordinating invisible hand of the market process, rather than being guided by the discretionary visible hand of the political process (Boettke and Candela 2014).

Recognizing the link between the rule of law and economic performance has important implications not only for economic theory but also for economic policy as well. As evidenced by the American financial crisis of 2008, the European sovereign debt crisis, and the recent Brexit vote of June 2016, discretionary monetary policy, fiscal policy, public administration, and labor market policy has lead increasingly to a movement away from societies based on contractual exchange to societies based on the status of expert rule, which privileges nondemocratic, expert administration of independent regulatory agencies that violate

the rule of law. If this trend continues, the overall effect will not only lower standards of living but also erode civil liberties and increase the fractionalization of societies, as a result of legislation privileging certain groups of people at the expense of another. As we have tried to argue in this entry, political officials cannot violate the rule of law for the purpose of expediency without simultaneously stifling the principles upon which the spontaneous creative powers of a free civilization are based.

Cross-References

- ▶ [Austrian Perspectives in Law and Economics](#)
- ▶ [Austrian School of Economics](#)
- ▶ [Common Law System](#)
- ▶ [Customary Law](#)
- ▶ [Efficiency](#)
- ▶ [Hayek, Friedrich August von](#)
- ▶ [Liberty](#)
- ▶ [Mises, Ludwig von](#)
- ▶ [Rule of Law](#)

References

- Acemoglu D, Johnson S (2005) Unbundling institutions. *J Polit Econ* 113(5):949–995
- Acemoglu D, Robinson JA (2012) *Why nations fail: the origins of power, prosperity, and poverty*. Crown Publishing, New York
- Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. *Am Econ Rev* 91(5):1369–1401
- Acemoglu D, Johnson S, Robinson JA (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Q J Econ* 117(4):1231–1294
- Boettke PJ, Candela RA (2014) Development and property rights. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York
- Boettke PJ, Subrick JR (2003) Rule of law, development, and human capabilities. *Supreme Court Econ Rev* 10:109–126
- Cowen T (1990) The economic effects of a conflict-prone world order. *Public Choice* 64(2):121–134
- Diamond J (1997) *Guns, germs, and steel: the fates of human societies*. W.W. Norton, New York
- Gallup JL, Sachs JD, Mellinger AD (1999) Geography and economic development. *Int Reg Sci Rev* 22(2): 179–232

- Glaeser EL, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) Do institutions cause growth? *J Econ Growth* 9(3):271–303
- Hayek FA (1960) *The constitution of liberty*. University of Chicago Press, Chicago
- La Porta R, Lopez-de-Silanes F, Shleifer A (2008) The economic consequences of legal origins. *J Econ Lit* 46(2):285–332
- Leeson PT, Suarez PA (2016) An economic analysis of Magna Carta. *Int Rev Law Econ* 47:40–46
- Maine HS (1861 [1982]) *Ancient law*. Gryphon Editions, Birmingham
- North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, New York
- North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutional governing public choice in seventeenth-century England. *J Econ Hist* 49(4):803–832
- North DC, Wallis JJ, Weingast BR (2009) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, New York
- Olson M (1982) *The rise and decline of nations: economic growth, stagflation, and social rigidities*. Yale University Press, New Haven
- Rosenberg N, Birdzell LE Jr (1986) *How the west grew rich: the economic transformation of the industrial world*. Basic Books, New York
- Ručinská S, Müller R, Nauerth JA (2016) Economic development. In: Marciano A, Ramello GB (eds) *Encyclopedia of law and economics*. Springer, New York
- Sachs JD, Warner AM (1997) Sources of slow growth in African economies. *J Afr Econ* 6(3):335–376
- Smith A (1776 [1981]) *An inquiry into the nature and causes of the wealth of nations*. Liberty Fund Press, Indianapolis
- Weingast BR (1995) The economic role of political institutions: market-preserving federalism and economic development. *J Law Econ Organ* 11(1):1–31
- Weingast BR (1997) The political foundations of democracy and the rule of law. *Am Polit Sci Rev* 91(2):245–263

Schmoller, Gustav Von

Daniel Nientiedt

Walter Eucken Institute, Freiburg, Germany

Abstract

Gustav von Schmoller (1838–1917) was a German economist. As the leader of the younger historical school, he was a very influential academic in Imperial Germany. This entry describes Schmoller's life and career, his particular approach to economics, as well as two important methodological disputes that he participated in.

Life and Academic Career

Gustav von Schmoller was born on 24 June 1838 in Heilbronn. Starting in 1857, he studied *Kameralwissenschaft* (akin to modern public finance) at the University of Tübingen. He graduated in 1860. As part of his civil service clerkship, Schmoller worked at the Württemberg Statistical Office, where he evaluated the industrial craft census (*Gewerbebezahlung*) of 1861. This led to a publication in the *Württembergische Jahrbücher*, and, eventually, an offer of a professorship at the University of Halle.

Schmoller began teaching at Halle in 1864. His research there focused on the history of German small scale industry, resulting in the

1870 monograph *Geschichte der deutschen Kleingewerbe im 19. Jahrhundert*. In 1872, Schmoller became professor at the University of Strasbourg where he produced a book on the cloth and the weaver guilds of the city (*Die Straßburger Tucher- und Weberzunft*, published in 1879). As we shall see, the narrow focus of these earlier works is a clear indicator of Schmoller's approach to economics in general.

In 1882, Schmoller assumed the chair of political economy at the University of Berlin. During the three decades that followed, he dominated German economics, serving as editor of one of the main journals (later renamed *Schmollers Jahrbuch* in his honor) and wielding considerable influence on academic appointments. He also became a member of the Prussian parliament. Schmoller's magnum opus *Grundriss der allgemeinen Volkswirtschaftslehre* was published in two volumes in 1900 and 1904. He passed away on 27 June 1917 in Bad Harzburg.

For biographical details on Schmoller see Balabkins (1988) or Winkel (1989). Regrettably, very few of his works have been translated into English; on this aspect, see Senn (1993).

The Schmoller Program

A good description of Schmoller's research program has been provided by Joseph Schumpeter (1926). He notes that what is assumed as given in theoretical economics becomes the subject of

investigation in the *Schmollerprogramm*. These facts include “the particulars of a nation’s economic environment, the endowment with natural resources, capital and machinery, the position in international trade, the social structure, size, composition and distribution of the social product, and the economic and political constitution” (Schumpeter, quoted in Backhaus 1993, pp. 4–5).

According to Schmoller, it is the task of the researcher to collect this information, make it accessible and analyze the technical relationship of the different variables. In practice, this meant the accumulation of large amounts of historical material on very specific topics – such as the Strasbourg guilds mentioned above – so as to understand economic phenomena in their full complexity. However, Schmoller ultimately hoped to find empirical regularities that would allow for the formulation of historical laws (Ebner 2000).

From a modern perspective, a remarkable aspect of Schmoller’s approach was his commitment to interdisciplinary research that draws from disciplines such as psychology or anthropology to enrich economic analysis (Backhaus 1993). Because of his interest in the functioning of social institutions, Schmoller has been referred to as a forerunner of the field of institutional economics (Schumpeter 1926) and the new institutional economics (Richter 1996).

Schmoller’s research program must be viewed in the context of the German historical school, a nineteenth-century movement in the fields of law and economics which was rooted in German idealism. The school maintained that economic events are particular to a certain historical or cultural context. What distinguished Schmoller, the leader of the “younger” historical school, from his predecessors, was his intention to combine scientific research with the pursuit of social policies (Pribram 1983). Indeed, Schmoller has been described as being primarily motivated by the desire to address contemporary social policy issues, such as the plight of workers during industrialization (Grimmer-Solem 2003). Characteristically, the German Economic Association, which he helped to establish in 1872, was – and still is – named *Verein für Socialpolitik* (Social Policy Association). For more information on

Schmoller’s views on social reform and the role of the state see Backhaus (1997).

Two Battles over Methods

Today, Schmoller is possibly best remembered for his role in two methodological disputes or “battles over methods” (derived from the German *Methodenstreit*) that have had a significant impact on the development of the social sciences. It should be noted that, in both cases, Schmoller supported the position that was eventually abandoned by the economics profession.

The original *Methodenstreit* was fought between Schmoller and the Austrian Carl Menger. In 1883, Menger published a book defending the use of abstract theory and deductive reasoning in economics. With respect to the necessity of the theoretical method, he observed “that *history*, to be sure, has the task of making us understand *all* sides of *certain* phenomena, but that exact *theories* have the task of making us understand only *certain* sides of *all* phenomena” (Menger, quoted in Caldwell 2004, p. 68). While Schmoller could have agreed with the importance of theory, Menger also emphasized that theory cannot be deduced from empirical observations, thus attacking the very foundations of the inductive-historical method. Schmoller published a skeptical review of the book, which led to an angry exchange of letters and papers that lasted for over a decade. For more details see Hodgson (2001).

The second dispute, between Max Weber and several social reform-oriented members of the *Verein für Socialpolitik*, including Schmoller, took place in the years leading up to World War I. Since it was concerned with the problem of value-freedom in the social sciences, it has become known as the *Werturteilsstreit* (value judgment dispute). Weber (1949) famously argued that scientific assessment of facts should be strictly distinguished from value judgments. Schmoller, on the other hand, maintained that some value judgments are more objective than others because large groups of people share certain moral ideas. For a discussion of Schmoller’s contribution to this debate see Hagemann (2001).

Impact and Legacy

Schmoller's influence during his lifetime stands in marked contrast to later assessments of his work. When discussing Schmoller's legacy some 25 years ago, Jürgen Backhaus noted that he is "widely ignored by professional economists today" and that "the few who know about him generally consider his work irrelevant" (Backhaus 1993, p. 3). Still, a quarter century later, there is arguably a renewed interest in the way that Schmoller thought about economics. Following the financial crisis and recession, there exists a demand for an approach that considers economic questions within their real-world cultural, social, and political contexts. As emphasized by the current editors of *Schmollers Jahrbuch*, such an approach – sometimes referred to as "contextual economics" – is consistent with Schmoller's understanding of economics as a "practical discipline where a turn toward the object (confronting actual economic conditions), the acting agent (and her or his perceptions), as well as problems (the social question) were what really mattered" (Goldschmidt et al. 2016, p. 5).

Cross-References

- [Austrian School of Economics](#)
- [Cameratism](#)
- [Craft Guilds](#)
- [Institutional Economics](#)
- [Political Economy](#)
- [Weber, Max](#)

References

- Backhaus JG (1993) Gustav Schmoller and the problems of today. *Hist Econ Ideas* 1:3–25
- Backhaus JG (ed) (1997) *Essays on social security and taxation: Gustav von Schmoller and Adolph Wagner reconsidered*. Metropolis, Marburg
- Balabkins NW (1988) Not by theory alone... The economics of Gustav von Schmoller and its legacy to America. Duncker & Humblot, Berlin
- Caldwell B (2004) *Hayek's challenge: an intellectual biography of F. A. Hayek*. University of Chicago Press, Chicago

- Ebner A (2000) Schumpeter and the "Schmollerprogramm": integrating theory and history in the analysis of economic development. *J Evol Econ* 10:355–372
- Goldschmidt N, Grimmer-Solem E, Zweynert J (2016) On the purpose and aims of the journal of contextual economics. *Schmollers Jahrbuch J Context Econ* 136:1–14
- Grimmer-Solem E (2003) *The rise of historical economics and social reform in Germany, 1864–1894*. Clarendon Press, Oxford
- Hagemann H (2001) The Verein für Sozialpolitik from its foundation until World War I. In: Augello MM, Guidi MEL (eds) *The spread of political economy and the professionalisation of economists: economic societies in Europe, America and Japan in the nineteenth century*. Routledge, London, pp 152–175
- Hodgson GM (2001) *How economics forgot history: the problem of historical specificity in social science*. Routledge, London
- Pribram K (1983) *A history of economic reasoning*. The Johns Hopkins University Press, Baltimore
- Richter R (1996) Bridging old and new institutional economics: Gustav Schmoller, the leader of the younger German historical school, seen with neoinstitutionalists' eyes. *J Inst Theor Econ* 152:567–592
- Schumpeter JA (1926) Gustav v. Schmoller und die Probleme von heute. *Schmollers Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft* 50:337–388
- Senn PR (1993) Gustav von Schmoller in English: how has he fared? *Hist Econ Ideas* 1:267–329
- Weber M (1949) "Objectivity" in social science and social policy. In: *The methodology of the social sciences*. Free Press, New York, pp 49–112
- Winkel H (1989) Gustav von Schmoller (1838–1917). In: Starbatty J (ed) *Klassiker des ökonomischen Denkens*, vol 2. C. H. Beck, Munich, pp 97–118

Scholarly Publishing and Open Access

Matteo Migheli^{1,2,4} and Giovanni Battista Ramello^{3,4}

¹Università di Torino, Torino, Italy

²CeRP-Collegio Carlo Alberto, Torino, Italy

³DiGSPES, University of Eastern Piedmont, Alessandria, Italy

⁴IEL, Torino, Italy

Definition

Every year, millions of peer-reviewed scientific papers are written, submitted to, and evaluated

by the referees of the thousands of scholarly journals through which researchers and academics over the world try to publish to promote their ideas and at the same time enhance their reputations. In all disciplines, albeit with partly different rules, scholars take part in this collective game called “scholarly publishing” which, through a historical process of progressive refinement and consolidation, has become central to modern scientific practice.

Yet, in recent decades, technological change seems to be challenging the consolidated organization of scholarly communication, through the innovation of open access (OA) publishing. OA on the one hand promises to make knowledge universally accessible but on the other hand alters the equilibrium on which the traditional system of scientific journals was based. In fact, the change is not merely a technological substitution but rather involves the complex economic and social organization that governs the world of scholarly publishing and especially that of research. Various inertias constrain the behavior of the different stakeholders in research, with important – and also potentially distortive – effects on the adoption of new technology.

Scholarly Research and Publishing: Main Characteristics and Organizational Structure

The principal function of scientific research is to produce and circulate knowledge advances, and the scientific community has spontaneously organized itself – in more or less complex ways – to carry out this task (Origgi and Ramello 2015). To pursue this aim and govern the erratic dynamics of science which “[...] is not a system of certain, or established, statements; nor is it a system which steadily advances towards a state of finality” (Popper 1959, p. 278), and instead progresses through a widespread dialectical process of validation and critical selection of results, there spontaneously emerged a “republic of science,” i.e., a specific organizational structure based upon its own social norms and institutions (Polanyi 1967).

Within that context, certain instruments have become essential for the functioning of the community. Among these, scholarly journals can be regarded as the cornerstone of the modern international scientific system, since they effectively perform several distinct, though interrelated, functions vital to the world of research. In particular, in addition to disseminating scientific advances, publications have a central role in the selection and validation of results, as well as in producing incentives for researchers (Clemens et al. 1995; Migheli and Ramello 2018). It is no coincidence that the imperative which today governs the day-to-day work of scientists across the world is “publish or perish,” against which backdrop scholarly journals play a predominant role. The peer review system, citations, and rankings have for some decades formed the framework of research activity and furnished (though subject to much criticism) the benchmark for measuring scholars’ productivity and determining their remunerations (Leahey 2007; Davis 2009).

Today, in many countries, the number and placement of published papers determine researchers’ prospects for career, tenure, and promotion, as well as their economic rewards – whether direct (e.g., higher salary, research prizes, etc.) or indirect (e.g., fame and reputation which can be monetized in alternative markets such as consulting) – and their ability to secure funds for future work from the public system (through research assessments) or from private sponsors (Origgi and Ramello 2015; Migheli and Ramello 2018).

In studying the dynamics of citations, Merton (1988, p. 620) likens them to a coinage system which makes it possible to capitalize “pellets of peer recognition that aggregate into reputational wealth” for individual scholars. Taking up this metaphor, we can say that, in the modern scientific system, the stock of reputational wealth accumulated by researchers consists of their publications, just as the financial wealth of individuals consists of their investment portfolios. Continuing the analogy, the journals in which articles are published can be likened to the issuers of financial instruments, characterized by a certain level of quality and reputational yield. Yet, whereas

financial markets have an inverse quality/performance ratio, that of academic publication is direct. In other words, whereas financial markets expect low-quality issuers (products with low ratings) to provide higher yields than high-quality issuers, in scientific publishing, high-end journals provide authors with greater prestige and a higher remuneration.

On the whole, peer-reviewed journals provide an immediate, even if imperfect, signal that can be converted into reputation, prestige, and all the attendant benefits for individual researchers. The metrics of impact factor and journal rankings depend more or less directly on the total number of citations, irrespective of the relevance of each individual contribution. This means there is a standard “yield” for papers published in a given journal, thus making citations fundamental for defining researchers’ value in the academic job market.

Naturally, like any financial instrument, journals are subject to fluctuations in value, and a careful selection on the part of scholars generally tends to favor “safe investments,” by virtue of their history or because they are regarded as such by the social norms which govern scientific communities (Clemens et al. 1995; Park and Qin 2007; Migheli and Ramello 2014). This gives rise to a self-perpetuating rigidity and inertia in the prestige of journals, and so also in the publication choices of researchers, which can only be broken by some external intervention, random shock, or purpose-designed policy (i.e., a form of “financial policy”) that has the effect of altering perceptions and hence the decisions of individuals (Migheli and Ramello 2013).

In the case of scientific journals, this outcome can be achieved in two ways: indirectly, by using a sort of “brand extension” to transfer to a new journal prestige established elsewhere (e.g., endorsements by scientific societies, links with other journals or prestigious universities) or, directly, by altering the existing system of incentives (Migheli and Ramello 2018). In the latter case, for example, an institution that funds research could impose ex ante constraints relating to publication choice, or a research department could create a reward system, either

direct (a bonus) or indirect (higher consideration of certain publications when evaluating research, moral suasion), that makes it rational to deviate from the previous conformist route of choosing well-established and “safe” journals.

A step in this direction, for example, is the widely discussed National Institutes of Health Public Access Policy, known as the “NIH mandate.” To ensure public access to research results, the NIH requires that final, accepted peer-reviewed journal manuscripts that are the product of NIH-funded research must be made freely accessible in the PubMed Central digital repository (<http://publicaccess.nih.gov/>). Among the operations of moral suasion on the part of institutions, we can cite, in chronological order, the *Faculty Advisory Council Memorandum on Journal Pricing* of Harvard University, eloquently subtitled “Major Periodical Subscriptions Cannot Be Sustained” and which calls for authors, editorial board members, and scientific associations to take specific actions in favor of OA publication (see points 2, 3, and 4 of the memorandum; <http://isites.harvard.edu/icb/icb.do?keyword=k77982&tabgroupid=icb.tabgroup143448>).

Naturally, for scholars to play the “publish or perish” game, a necessary (though not sufficient) condition is the ability to access existing knowledge, i.e., already published papers. This poses a problem because, up until very recently, the business model on which journals relied was exclusion of readers through price – known as the *subscriber-pays* model (and hereinafter also denoted Restricted Access, or RA) – as commonly practiced in markets for a wide array of goods and services. The subscriber-pays solution involves the transformation of a public good – scientific information – into a private information good, thus creating a trade-off between rationing as a way to give publishers an economic incentive and the exclusion of follow-on researchers who cannot pay the cost of access (Origgi and Ramello 2015).

Various compromises have been arrived at over time to minimize this exclusion while at the same time producing sufficient incentives. Generally, the approach taken has been to make journals

a club good, that is to say a sort of local public good for a vast population of individuals (Cornes and Sandler 1996). For example, scientific associations have often included their own journals in their membership fees, thereby making them accessible to all members. On their part, libraries have likewise played a fundamental part in making journals accessible to users located in the geographical vicinity of the library. This practice in effect solved the financing problem through an unprecedented form of price discrimination, which imposed on the subject best able to bear the financial burdens – i.e., the libraries – the high costs of purchasing journals, while providing access at a much lower price (or sometimes for free) to the individual users of the library (Liebowitz 1986). The system worked so well that, from the point of view of academics, the online accessibility of paid-for journals was often perceived as a form of OA.

However, in recent times the above mechanism started to falter. Commercial publishers, also thanks to a process of concentration, increased their market share. This, combined with an essentially rigid demand, led to steep increases in price, further favored by the bundling practices made possible by selling large catalogues of journals in electronic format, known as the Big Deal (McCabe 2002). Compounding this was an unconnected – but not for this less important – contraction in the budgets of universities, and consequently of libraries, which began to force some dramatic choices. Initially, the rigid demand for journals worked in their favor, prompting libraries to assign journals a larger share of their budgets, at the expense, for example, of books and monographs. However, a more significant consequence, for the purposes of this analysis, is that eventually libraries became less and less able to maintain their existing catalogues, culminating in what is now referred to as the “serials crisis” in the literature (Panitch and Michalak 2005). At first, the problem affected the smaller universities and those of less wealthy countries, but it has now come to plague even the most prestigious institutions, as testified by the aforementioned case of Harvard University.

These tensions, which made it harder to access journals, prompted a wide-ranging debate on the need to alter the status quo through policies designed to favor a shift toward an open system (Bergstrom 2001) and in some cases prompted more drastic reactions, such as the abandonment of certain journals on the part of editorial boards or scientific societies – as happened, for example, to the *European Economic Review* (Cavaleri et al. 2009) – or the boycotting of certain publishers by declining the various possible forms of collaboration.

New Trends in Scholarly Publishing

Against this backdrop, there emerged the innovation of free electronic journals, introduced around the 1990s (and only later labeled open access, or OA), which were widely regarded as the disruptive innovation that could solve the existing problems and return scientific advances to a sphere more compatible with the nature of knowledge as a public good. Many contributions have pointed out the advantages that these new publications offer compared to journals based on the *subscriber-pays* model and have urged their extensive adoption (Bergstrom 2001; Parks 2002; Willinsky 2006).

Some authors have even expressed expectations far beyond the realm of technological change, seeing in OA journals a universal remedy that can solve all the problems of scholarly communication (Harper 2009). In similar vein, a British antitrust investigation into presumed anti-competitive practices in the market of scholarly journals in science, technology, and medicine, in the words of the authority chairman Sir John Vickers, concluded that “the market, which operates world wide, has a number of features that suggest that competition may not be working effectively. *However, market forces harnessing new technology may change this without the need for intervention*” (OFT 2002; the italics are ours).

Still, even though there are ample signs of an increasing presence of OA journals and papers in the world of research, the market of traditional

RA journals continues to flourish for a number of reasons (Laakso et al. 2011; Migheli and Ramello 2014). In fact, not only does RA provide incumbents with large and growing profit margins: in many sectors, it also relegates OA journals – with some exceptions – to a subordinate role on the sidelines of scholarly communication.

Therefore, despite the OA model's disruptive potential to better disseminate scientific advances and lower production, printing, and distribution costs (Houghton et al. 2009), the scholarly communication system exhibits structural rigidities tied to the existing system of incentives, essentially stemming from social norms and prudential attitudes on the part of researchers.

Indeed, the persistence and stickiness of the social norms governing researchers' behavior with respect to publication choice can become crucial in determining the uptake of new technological opportunities and the dynamics of scholarly communication (Migheli and Ramello 2013, 2014). In other words, publication choice does not seem to be determined by any specific, immutable preference for a given technology platform but rather by the scientific community's lack of "absorptive capacity," i.e., what the innovation literature defines as the ability of players (here, scholars and research institutions) to recognize and adopt the opportunities offered by major technological breakthroughs, such as OA publishing (Cohen and Levinthal 1990; Feess and Scheufen 2016). The puzzle to unravel therefore concerns the signals produced by the existing RA system, which are not easily replicated in the OA realm without the help of some exogenous intervention or shock. Understanding the dynamics that determine preferences, and hence choices, within a given scientific community is thus an essential starting point for identifying possible drivers of change.

The Long and Winding Road of OA Publishing

The OA movement took off with the information and communication technology (ICT) revolution, which greatly reduced the investment in cost and time required to produce a journal, while also offering unprecedented ease of distribution. Naturally, the new technology is available to all

journals, and it can only bring about change if used in a certain way: commercial publishers see the ICT revolution as merely a new tool for pursuing price discrimination strategies by creating electronic catalogues; OA publishers instead emphasize its ability to offer high-quality peer-reviewed content in digital format without any access restrictions (McCabe 2002; Willinsky 2006).

The OA philosophy is thus a product of the encounter between technological opportunity and a new business model: it is the two innovations together – technological and economic – that enabled a Copernican Revolution which overturned the traditional model of scholarly communication. Scientific knowledge, hitherto "boxed in" by the practical necessities of paper journals – which, through reification of the physical medium, assimilated knowledge to a traditional consumer good – could once again become a public good thanks to marginal costs of duplication and dissemination that are equal to zero.

However, the OA approach also brings back to the fore a characteristic problem for information, that of provision of a public good. Whereas the RA model resolved this in the market, through payment of a subscription fee, in the OA realm, it is necessary to find other solutions that lie outside the market or which are linked to complementary markets.

A variety of models have been proposed: the first is *patronage*, in part also familiar to RA journals, whereby an institution takes charge, entirely or partially, of the costs of the journal. Another model to rely on is *advertising*: as already happens in other media sectors (e.g., in free-to-air television), the publisher operates in a two-sided market that on one side meets the scientific community's demand for content and on the other side sells the audience's attention to firms (Anderson and Gabszewicz 2006). Finally, the model most commonly adopted to date has been to overturn the financial burden, transferring it to the party that most directly and immediately benefits, that is to say, the author. This introduces a novel conception of the market, termed the *author-pays model*, whereby researchers pay for the service of having their results selected and

circulated, rather than for accessing the work of others. Such a model has the advantage that, for an equal amount of expenditure, it eliminates the exclusion through price of part of the audience and lets research results remain a public good. It is therefore a cost-effective solution (Bergstrom and Bergstrom 2004; Houghton et al. 2009). That being said, the *author-pays model* is not immune from criticism because it raises the direct costs of research and excludes those who cannot pay the publication fee (Gass 2005).

While the different alternative financing models are extensively discussed elsewhere (Houghton et al. 2009), we focus here on the unequivocal fact of OA publishing's continued expansion (Laakso et al. 2011). The Directory of Open Archive Journals (DOAJ) is an ever-growing list of the principal OA journals that follow the so-called golden road, meaning they offer all their content free of charge (Harnad et al. 2004). Added to these are the many RA journals that occasionally offer – often on payment by the author – an OA option for individual papers (this is the so-called hybrid form; Björk 2012). Carroll (2011) however notes that hybrid forms are not truly OA since they do not clearly define the rights of readers, such as whether reuse, free circulation, and so forth are permitted. Finally, we can add a large catalogue of freely accessible “working papers,” i.e., preliminary versions of papers published in RA journals, a practice known as the “green road” (Harnad et al. 2004).

These last two forms show how OA has come to increasingly permeate the day-to-day scientific work of researchers, as well as the interest OA has elicited – through the hybrid form – even among traditional journals with a strong for-profit orientation. It is thus all the more intriguing to try find out why gold OA journals have nonetheless had difficulty becoming established in certain fields, in apparent contradiction with the widespread favor they enjoy in the scientific field.

The geographical distribution of OA journals is quite broad (the figures, always in-progress, are provided by the DOAJ listing) published in more than 100 different countries. Of these, around 15% are published in the USA, and if we consider all Anglo-Saxon countries together, they

account for the lion's share, with over 2000 titles. Europe and Latin America also have substantial numbers of OA titles, whereas in the rest of the world, distribution is more fragmented. The “economics” category amounts to a relatively small subset of the DOAJ universe, but this number increases substantially if we also include the titles under “management and business.” Tradition is well established, and the oldest OA journal, *Economics Bulletin*, was established in 2000 (operational from 2001) by John Conley and Myrna Wooders and is archived by the University of Illinois at Urbana-Champaign (Conley and Wooders 2009). Interestingly, the American Economic Association has for the past few years offered all the content of the *Journal of Economics Perspectives* on an OA basis, but in place of this label, the website prefers: “with the compliments of the American Economic Association.”

Publication Choice and the Open Access Paradox

The above description gives us a general idea of the trends followed in recent years. However, they tell us nothing about the reasons for this dynamic nor anything specifically about the role of demand in determining the establishment of OA in scientific publishing.

Behavioral analysis is a useful instrument that can help us penetrate the social black box, understand the trajectory of OA journals, and formulate policy guidelines for promoting the adoption of OA within the scientific community if this – as would appear from the cited cases – is a goal research institutions wish to pursue. However there has been research addressing this question in the fields of medicine, life sciences, physics, economics, and chemistry.

Over the past two decades, there have been a number studies investigating the attitude of scholars to scientific publications in general and to free access publications in particular. However, these investigations were conducted at very different times, in very different contexts, and following widely disparate methodologies. Xia (2010) and Davis and Walters (2011) systematize

a set of empirical studies based on interviews with small samples, and surveys instead conducted on larger populations, generally focused on Europe, North America, and Australia, thus providing a Western-oriented perspective. The samples investigated are fairly diverse and present structural limitations (opportunately noted by Xia 2010; Davies and Walters 2011) but still allow us to detect broad correlations and extract insights for further research. The most noticeable pattern to emerge is scholars' attentiveness to the quality and perceived reputation of journals, confirming the idea that scientific papers are effectively regarded as fundamental assets in the research job market and that community consensus is effectively the pervasive reference framework in which choices are made (Migheli and Ramello 2014).

From the academic career perspective, OA journals should be technologically neutral, since they can just as easily as RA journals produce results aligned with the incentive system of the academic job market. Yet certain peripheral features of this new form of scholarly publishing may elicit diffidence. For example, there tends to be an aversion (though this has been abating over the years) to payment of an author's fee, motivated on one side by the fear that this will increase scholars' direct costs of doing research and on the other by concerns about a possible quality trade-off between maximizing journals' profits and the rigor of the peer review process (Davis and Walters 2011; Migheli and Ramello 2014).

As discussed previously, two important factors driving publication choice are the expected number of citations, both individually and for the journal, and the visibility of published papers (Migheli and Ramello 2018). Both of these appear to be correlated with the number of downloads, which is today the most common method of consuming both OA and RA journals. Accordingly, some studies have sought to measure these values. There has even been a sort of natural experiment conducted by Oxford University Press, which converted its flagship journal on molecular biology (*Nucleic Acid Research*) from RA to OA. For a given journal reputation, this operation first of all yielded increased profits for the publisher but

also an increase in downloads (though with some uncertainty as to the numbers, since these were partly due to search engines). Still, even the estimated increase without counting search engines, of +7–8%, is noteworthy if we consider that this was an already established journal, which means the entire rise can be ascribed to the effect of OA in making content more visible. We do not as yet have any clear indications from this experiment on the effect upon citations (Bird 2008), and in any case the journal in question might be expected to have a reputation and inertia deriving from its previous RA status. Generally, if the equation “more downloads=larger readership=increased citations” was to prove true, it ought to give OA journals a formidable lever for building their ascent while allowing individuals to comfortably base their publication choices on a self-enforcing prophecy.

This suggests the hypothesis that, in the absence of clear indicators, the choice of preferring traditional journals (generally RA) is a behavior aimed at minimizing risk under the currently existing social norms. Such a hypothesis could provide a reasonable explanation for an intriguing ambivalence observed in various studies, which might be termed the “open access paradox”: scholars generally express strong support for OA and use it extensively, also via the green road, but this enthusiasm is not later transferred to their publication choices (Park and Qin 2007; Hubbard et al. 2011; Migheli and Ramello 2014). Although the stated system of preferences should theoretically influence individual actions, the expressed desirability of OA journals is contradicted by researchers' reluctance to submit their papers to those journals.

This seeming contradiction between assertions of general principle and individual behavior poses the interesting puzzle of a possible divergence between individuals' incentives and what they regard as the collective interest. The solution to the paradox probably involves rational choice prevailing over the emotional sphere, suggesting that policies aimed at changing the status quo must necessarily follow the route of altering the economic incentives. In other words, we can reasonably expect OA journals to gain market

share and become substitutes for other titles, provided that a self-enforcing prophecy is triggered which causes an ex post outcome in the reference scientific community that is compatible with the incentives that existed ex ante.

It is therefore a question of identifying the factors that determine how individual researchers choose and evaluate journals. These include the researcher's position in the academic hierarchy, gender (because of the discrimination that women often suffer, also in academia; Leahey 2007) and working context (including the geographical location), the expected impact of OA journals in yielding visibility and reputational capital, and so forth (Migheli and Ramello 2013, 2014).

Cross-References

- [Cartels and Collusion](#)
- [Concentrated Ownership](#)
- [Market Failure: Analysis](#)
- [Merger Control](#)

References

- Anderson SP, Gabszewicz JJ (2006) The media and advertising: a tale of two-sided markets. In: Ginsburg VA, Throsby D (eds) *Handbook of cultural economics*. Elsevier, Amsterdam
- Bergstrom TC (2001) Free labor for costly journals? *J Econ Perspect* 15:183–198
- Bergstrom TC, Bergstrom CT (2004, May 20) Can 'author pays' journals compete with 'reader pays?'. *Nature Web Focus*. <http://www.nature.com/nature/focus/accessdebate/22.html>. Lastly accessed 20 June 2018
- Bird C (2008) Oxford journals' adventures in open access. *Learned Publish* 21:200–208
- Björk BC (2012) The hybrid model for open access publication of scholarly articles: a failed experiment? *J Am Soc Inf Sci Technol* 63:1496–1504
- Carroll MW (2011) Why full open access matters. *PLoS Biol* 9(11):e1001210. <https://doi.org/10.1371/journal.pbio.1001210>. Lastly accessed 20 June 2018
- Cavaleri P, Keren M, Ramello GB, Valli V (2009) Publishing an e-journal on a shoe string: is it a sustainable project? *Econ Anal Policy* 39:89–101
- Clemens ES, Powell WW, McIlwaine K, Okamoto D (1995) Career in print: books, journals, and scholarly reputations. *Am J Sociol* 101:433–494
- Cohen WM, Levinthal DA (1990) Absorptive capacity: a new perspective on learning and innovation. *Adm Sci Q* 35:128–152
- Conley JP, Wooders M (2009) But what have you done for me lately? Commercial Publishing, Scholarly Communication, and Open-Access. *Econ Anal Policy* 39:71–87
- Comes R, Sandler T (1996) *The theory of externalities, public goods, and club goods*. Cambridge University Press, Cambridge
- Davis PM (2009) Reward or persuasion? The battle to define the meaning of a citation. *Learned Publishing*, 22(1):5–11
- Davis PM, Walters WH (2011) The impact of free access to the scientific literature: a review of recent research. *J Med Libr Assoc* 99:208–217
- Feess E, Scheufen M (2016) Academic copyright in the publishing game: a contest perspective. *Eur J Law Econ* 42:263–294
- Gass A (2005) Paying to free science: cost of publication as costs of research. *Ser Rev* 31:103–106
- Harnad S, Brody T, Vallieres F, Carr L, Hitchcock S, Gingras Y, Oppenheim C, Hajjem C, Hilf E (2004) The access/impact problem and the green and gold roads to open access: an update. *Ser Rev* 34:36–40
- Harper G (2009) OA and IP: open access, digital copyright and marketplace competition. *Learned Publish* 22:283–288
- Houghton J, Rasmussen B, Sheehan P, Oppenheim C, Morris A, Creaser C, Greenwood H, Summers M, Gourlay A (2009) Economic implications of alternative scholarly publishing models: exploring the costs and benefits. Loughborough University, Loughborough
- Hubbard B, Hodgson A, Fuchs W (2011) Current issues in research communications: open access – the view from the Academy. Centre for Research Communications, University of Nottingham. http://eprints.nottingham.ac.uk/1480/1/RCS_March_2011.pdf. Lastly accessed 20 June 2018
- Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C et al (2011) The development of open access journal publishing from 1993 to 2009. *PLoS One* 6(6):e20961. <https://doi.org/10.1371/journal.pone.0020961>
- Leahey E (2007) Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings. *American Sociological Review* 72:533–61
- Liebowitz SJ (1986) Copyright law, photocopying, and price discrimination. *Res Law Econ* 8:181–200
- Migheli M, Ramello GB (2018) The market of academic attention *Scientometrics*, 114(1):113–133
- McCabe MJ (2002) Journal pricing and mergers: a portfolio approach. *Am Econ Rev* 92:259–269
- Merton RK (1988) The Matthew effect in science, II. Cumulative advantage and the symbolism of intellectual property. *Isis* 79:606–623
- Migheli M, Ramello GB (2013) Open access, social norms & publication choice. *Eur J Law Econ* 35:149–147
- Migheli M, Ramello GB (2014) Open access journals & academics' behaviour. *Econ Inq* 52:1250–1266
- OFT (2002) *The market for scientific, technical and medical journals*. Office of Fair Trading, London
- Origgi G, Ramello GB (2015) Current dynamics of scholarly publishing. *Eval Rev* 39:3–18

- Panitch JM, Michalak S (2005) The Serials crisis, UNC-Chapel Hill Scholarly Communications Convocation. <http://www.unc.edu/scholcomdig/whitepapers/panitch-michalak.html>. Lastly accessed 20 June 2018
- Park JH, Qin J (2007) Exploring the Willingness of Scholars to Accept Open Access: A Grounded Theory Approach. *Journal of Scholarly Publishing* 38:55–84
- Parks RP (2002) The Faustian Grip of Academic Publishing. *Journal of Economics Methodology* 9:317–35
- Polanyi M (1967) The republic of science. *Minerva*, 1(1): 54–73
- Popper KR (1959) *The logic of scientific discovery*. Basic Books, New York
- Willinsky J (2006) *The access principle: the case for open access to research and scholarship*. MIT Press, Cambridge, MA
- Xia J (2010) A Longitudinal Study of Scholars' Attitudes and Behaviors toward Open-Access Journal Publishing. *Journal of the American Society for Information Science and Technology* 6:615–24

Securities Class Action

Lucia dalla Pellegrina and Margherita Saraceno
DEMS, University of Milano-Bicocca and BAFFI
CAREFIN Centre, Università Bocconi,
Milan, Italy

Definition

Securities class actions (SCAs) are lawsuits brought before a court on behalf of a group of investors, whose rights have allegedly been violated by the company in which they have a stake. Usually, allegations concern the fact that the company and some of its officers violated federal or state securities laws. In this article, we overview the procedural characteristics of SCAs in the USA and other countries that have established this type of legal remedy, discussing their role as compensation, deterrence, market regulation, and corporate governance tools.

Introduction

This article overviews the main features of securities class actions (SCAs, hereafter), both in the USA and in other countries, discussing their role as compensation, deterrence, market

regulation, and corporate governance tools. Relying on empirical contributions that investigated the relationship between the event of an SCA and target companies' behavior, this contribution supports the view according to which SCAs seem to be a weak tool in terms of investors' compensation and deterrence. Rather, their substantial impact on the performance of a company suggests SCAs have an important role as market discipline tools. Consequently, several scholars are increasingly convinced that SCAs are to be classified as corporate governance instruments.

Securities Class Action: An Overview

USA is the ultimate country for SCA. On the other hand, other countries have similar procedures. In the subsections below, we describe SCA in the USA since it can be considered as a benchmark and then we briefly discuss procedures in other countries.

US Securities Class Action

An SCA is a lawsuit brought pursuant to Federal Rule of Civil Procedure 23 on behalf of a group of investors who purchased the securities of a specific company (*target company*) during a given period of time (*class period*). Usually, allegations concern the fact that the company *and/or* some of its officers violated federal or state securities laws. The main federal laws that are involved in SCAs are the Securities Act of 1933 and the Securities Exchange Act of 1934. In addition, the so-called state *blue sky laws* – aimed at protecting individual investors – may be alleged to be violated in SCA cases.

An SCA can be considered as a specific type of class action (CA, hereafter). With respect to many issues, considerations provided by Ramello (2014) extend to SCAs. In fact, the suit is filed as a CA because class members are so many that a joinder of all plaintiffs is unattainable.

As for any kind of CA, proceeding as a class must be superior to other available methods for a fair and efficient adjudication whereas the prosecution of separate individual actions could lead to inconsistent adjudications. According to Coffee (2006), SCAs represent the largest category of

CAs, absorbing relevant public resources of the judiciary subsidized by taxpayers.

On the other hand, SCAs are characterized by relevant specificities. One peculiarity is that, besides the usual rules governing CAs, SCAs are regulated according to more stringent requirements. In particular, after the Private Securities Law Reform Act (PSLRA) of 1995 and the Securities Litigation Uniform Standards Act (SLUSA) of 1998, class certification requires strict compliance with legal provisions that are not applied in the case of other types of CAs. The PSLRA was explicitly designed to limit frivolous securities lawsuits. Very briefly, under the PSLRA, plaintiffs need relevant proof of fraud before they can initiate an SCA. After filing, judges decide the most adequate lead plaintiff and a more active monitoring role of the court is prescribed; stringent discovery rules apply and full disclosure to investors of proposed settlements – including the amount of attorneys' fees – is mandated; bonus payments to favored plaintiffs are banned. The SLUSA was mainly aimed at curtailing the plaintiffs' practice of filing securities cases in state courts to avoid the proscriptions of the PSLRA. The SLUSA actually posed all the SCAs under both the jurisdiction and the strict surveillance of federal courts.

Beyond the procedural specificities that distinguish SCA from other types of CA, intrinsic differences should be carefully discussed to understand the everlasting debate about merits and limitations of SCAs.

As well presented by Coffee (2006), substantial differences exist between SCAs and all the other forms of both private litigation, including CAs, and public prosecution for legal violations. Main peculiarities are related to who suffers for a company's misbehavior and who bears the costs of damage compensation.

Usually, when a corporation incurs liability (even in a CA) because of a defective product, environmental pollution, or other misbehavior, the society imposes a cost on the corporation (and indirectly on its shareholders). Typically, shareholders (and other investors who have stakes in the target company) are not themselves the primary victims of the offense. Thanks to private litigation, the liability system is

enforced and the negative externality is internalized (at least partially) by the firm.

In most part of the SCAs, some shareholders (those who traded the company's securities in the class period) decide to proceed against the target company because of an alleged legal violation committed by corporate agents (executives, directors, auditors, other insiders) resulting in a damage for the shareholders themselves. Typically, the victims and the shareholders are largely the same.

The oddity of the SCA emerges quite evidently. A CA against a target company serves the rationale of the enterprise liability since it is assumed that corporate agents might engage in misconduct like production of defective products or environmental pollution to maximize profits for the company and its shareholders. Conversely, securities fraud appears to be primarily motivated by the manager's own personal interests. Except when the company is issuing new shares, inflating firm's earnings do not benefit either the company or its shareholders. Instead, executives and insiders could have self-interested reasons to inflate earnings when securities are traded on the secondary market; likewise, they have incentives to hide bad news, overstate good developments, understate risks to maximize both the value of their stock option and other equity compensations, and, eventually, improve their reputation.

In the case of securities frauds committed by corporation's executives, not only (some) shareholders are the main victims, but – in the case of an SCA – they also bear the burden of direct and indirect costs related to the legal remedy.

Obviously, SCAs should not be confused with shareholder derivative suits, which are lawsuits brought by a shareholder on behalf of a corporation against a third party. Often, the third party is an insider of the corporation, such as an executive officer or director. Shareholder derivative suits allow a shareholder to initiate a suit when the management has failed to do so.

Some Statistics: The Securities Class Action Clearinghouse Database

Scholars and practitioners who are interested in understanding the reality of SCAs can take

advantage of an invaluable asset: the comprehensive and compelling website of the Securities Class Action Clearinghouse (SCAC). The SCAC Database tracks all the SCAs filed in Federal Courts after the PSLRA. Information about plaintiffs, defendants, allegations, lawyers, court districts, judges, and settlement details involved in each case are summarized and easily available.

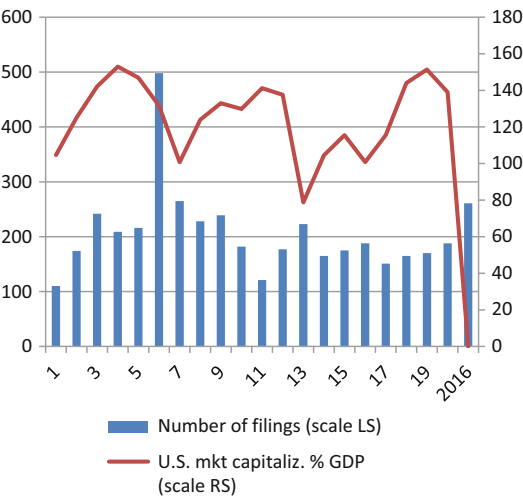
According to the database, from 1996 to 2016, 4,347 SCAs were filed (see Fig. 1). The impressive peak of 2001 seems related to the crash of the *dot com bubble*, given that about 58% of total filings have been filed against companies in the

ICT sector. Another specificity is observed in correspondence of the *subprime mortgage crisis*: in 2008, around 51% of total filings (46% in 2009) were against companies in the financial sector. By comparing market capitalization/GDP with the trend of SCAs, the latter seems to be somehow countercyclical.

In detail, 2,159 cases (49.7% of total filings) have been settled, for a total amount of all settlements of more than 89 billion dollars; 1,685 cases have been dismissed (38.8%), while 503 filings are still ongoing (11.6%). The largest settlement (\$7,227,390,000) has been paid by Enron Corporation in 2008 after ruling summary judgment but before trial verdict. Fees and expenses approved by the court represented 9.5% of the total settlement.

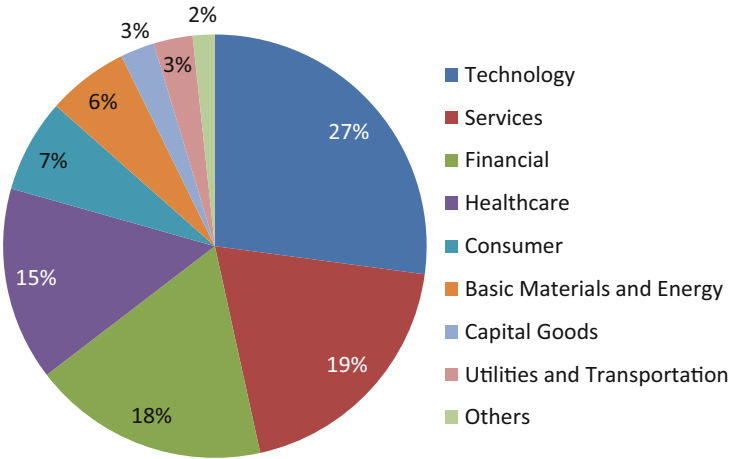
The most frequently sued sector is technology, recording 27% of total filings, services (19%), the financial sector (18%), and healthcare (15%) (see Fig. 2). The most frequently sued issuers are on the NASDAQ. The most active district court involved in SCAs is the Southern District of New York; the 2nd Circuit (New York) and the 9th (San Francisco) decide more than half of all the SCAs (31% and 23%, respectively).

This brief overview gives only a sketchy idea of the complexity and relevance – especially from a legal and economic perspective – of this kind of legal remedy. Section “[Empirical Evidence About Causes and Effects of SCAs](#)” summarizes various empirical contributions that investigated these complexities.



Securities Class Action, Fig. 1 SCAs and U.S. market capitalization as % of GDP from 1996 to 2016, by year

Securities Class Action, Fig. 2 SCAs from 1996 to 2016, by sector of target company



Other Countries

A few countries have forms of representative securities litigation working similarly to the US SCA, mainly Australia and Canada. In other countries, including England and Wales, South Korea, Germany, and the Netherlands, there are litigation tools that might be somehow assimilated to SCAs, but these significantly depart from the standards of the US Federal Rule of Civil Procedure 23. Below, only a brief selection of representative countries is considered.

Australia

In Australia, SCAs are now an established part of the legal landscape. After the first filing in 1999, several cases have been commenced. The Australian model is similar to the US SCA, but there are some relevant differences: no formal motion for class certification is required; moreover, the procedure is based on an *opt in* standard since interested claimants are required to register at the outset of the case. Like other litigation procedures in Australia, fee shifting is based on the English rule while contingent fees are prohibited (third-party litigation funding is allowed and is considered an effective substitute of the contingent fee system). See also Murphy and Cameron (2006).

Canada

SCA is available in Canada since 2004. SCAs are not regulated at the federal level, but there are province procedures that may differ and SCAs can simultaneously arise in different provinces. Until 2006, in the majority of provinces, SCA was applied only for frauds related to the primary market. However, since 2006, most provinces have amended their legislation to introduce a right of action through SCA for misrepresentations in the secondary market. The possibility to proceed for secondary market frauds significantly boosted SCA litigation. In many cases, procedural rules prescribe court approval before the action commencement, damages limitations, and the English rule for fee shifting. Sarra and Pritchard (2010) provide a detailed analysis of SCA in Canada.

England and Wales

Since 2000, England and Wales (but not Scotland) introduced civil litigation rules for the management of group litigation, including securities litigation. Unlike the US system, the English procedure is based on an *opt in* standard; the class may be represented by a lead lawyer, but also individual members of the group can have personal lawyers. The action commences with a Group Litigation Order establishing a register of the class and the involved claims; the court manages the entire procedure. The English rule of fee shifting and the unavailability of contingent fees probably limited the diffusion of SCA in England and Wales. However, as in Australia, the emergence of third-party litigation funding may favor the establishment of SCA. See also Andrews (2001).

South Korea

In 2005, SCA was introduced in South Korea. Under the local legislation, an SCA may be brought against a company for illegal acts connected with securities issued by the company including false statements, insider trading and market manipulation, and negligent audit. To proceed as a class, a minimum of 50 shareholders owning at least 0.01% of a company's equity must join as plaintiffs on behalf of the other shareholders. The CA must be certified by the court. Strict rules of representations apply: terminating the action, relinquishing a right granted under the law, or settling the case without court approval is prohibited. See Chung (2004).

Germany

The German law does not provide any specific type of representative group litigation. However, in 2005, new rules were enacted to facilitate investors' action for misstatements and omissions by issuers. In particular, according to the Capital Investor's Model Proceeding Act (Kapitalanlegermusterverfahrensgesetz, CIMPA, hereafter), amended in 2012 and valid till 2020, investors can proceed for securities claims according to a *model case*. Notably, with the CIMPA, the German legislator has provided a way to handle capital market mass proceedings

without transferring existing models of group litigation from foreign jurisdictions.

According to the procedure, both the plaintiff and defendant may enter an application for establishing a *model case* to the court. Admissible applications are publicly announced by the court in an electronic Complaint Registry. Only if at least ten applications referring to the same matter are entered in due time in the Registry, the court proceeds according to the CIMPA. While the model case proceeding is decided by the Higher Regional Court, all pending related proceedings are automatically suspended. Once decided by the Higher Regional Court, the model case is binding on all the courts dealing with that matter.

The procedure applies only to claims related to false, misleading or omitted public capital markets information, or due to breaches of the Securities Acquisition and Takeover Act. It only applies to parties who have already filed a suit and does not allow a claim to be brought in the name of an unknown group of claimants. The CIMPA, in fact, introduces a procedure which is not properly a representative lawsuit, but rather a representative case; note that the procedure is based on an *opt in* system since only claimants who choose to file an action are bounded by the model.

Litigation costs are allocated according to the English rule. Punitive damages or other kinds of noncompensatory damages are not permitted. The courts will then decide each individual proceeding itself, including the decision on individual damages. Contingency fees are available only in special circumstances. As a result, litigation funding/insurance is a common component in these suits.

The Netherlands

In 1994, the Netherlands enacted a legislation that permitted collective actions by consumer associations. In 2005, the Civil Code was amended by the Collective Settlement of Mass Damage Claims Act. Like for the US CAs, the latter ensures that class members will be bound by settlements under the Act unless they affirmatively decide to opt out. In 2007, thanks to a few cases related to incorrect financial information and financial

frauds involving the Dutch investor association VEB as plaintiff, the existing form of representative procedure extended to cases of securities litigation. As in the German case, the stockholders involved are still required to bring individual proceedings on reliance and damages issues. Under the Dutch procedure, only court-approved representative organizations (i.e., investor associations) can pursue an SCA on behalf of the investors. Furthermore, such organizations cannot seek damages in court. The court can only certify the class and decide whether to approve an out-of-court settlement between the parties. This could limit the potential of the Dutch representative action.

Finally, although this is beyond the scope of this article, consider that, today, financial markets and securities are globalized; investors, especially those of large public companies, live and trade securities in many countries. This obviously poses relevant problems of both jurisdiction and applicability of SCA procedures. The debate about these issues, and particularly about the extraterritorial effect of US SCA, has been significantly stimulated by the famous US Supreme Court's *Morrison* decision (*Morrison v. National Australia Bank*, 561 U.S. 247, 2010).

This ruling stated that the US securities laws apply only to securities transactions within the USA. In fact, the *Morrison* decision excluded “foreign-cubed” investors and “foreign-squared” cases (cases brought by domestic plaintiffs against foreign issuers and pertaining transactions on foreign exchanges) from US SCAs.

A Rationale for SCA: Compensation, Deterrence, Market Regulation, or Corporate Governance?

CA (see Miller 1979), in general, and SCA, particularly, are very controversial procedures. The main reason why the debate about the functions of securities litigation is even more vivid than in the case of other types of CA is due to the intrinsic *circularity* characterizing SCAs. As explained

above, shareholders are, at the same time, the possible beneficiaries of damage compensation inasmuch victims of company misbehavior and those who finally bear all direct and indirect consequences of an SCA.

Concerning the usual judicial functions (access to justice, compensation, and deterrence) typically ascribed to CAs, *access to justice* is the only one that is not significantly dismantled. SCAs may represent an affordable option (especially given the contingent fee arrangements) to many investors to proceed against the target company and their officers for securities frauds. As for CAs related to mass torts cases, environmental and product damages, securities frauds often affect fragmented and widespread claimants. Therefore, SCAs can be a means to improve the effective protection of collective interests and individual rights. By leading to a decision (either a court pronouncement or, usually, a settlement) against the liable injurer, SCAs offer a legal remedy in securities cases. Hence, group litigation performs the social function of granting every person to seek forms of judicial protection.

Circularity seriously jeopardizes the *compensation* function of SCAs. A first criticism is related to the fact that the net redress for class plaintiffs might be insufficient to compensate losses because of the significant fee amount to be paid to counsels and/or because of unfair settlements (Coffee 1995). A second criticism is that in SCAs the compensatory function is significantly undermined by the substantial correspondence between victims and those who finally bear direct and indirect consequences of the legal action on the target company. In fact, defendants very rarely pay for all the costs involved in an SCA by using personal resources. Typically, company's resources and D&O insurances are used to cover defense costs, damage compensation, and settlement. Even possible indirect consequences of a successful SCA, including loss of reputation and stock price drops, are borne by shareholders in the end. To the extent that an SCA imposes its costs almost exclusively on the corporation and insurers, it does not seem to benefit shareholders anyhow (Baker and Griffith 2009).

For the same reasons, the *deterrence* function of SCA seems to be substantially frustrated (Baker and Griffith 2009; Coffee 2006). Although the cost-internalization function of CA is independent of its compensatory function – and may be served even when the latter cannot – SCAs might effectively force the internalization of externalities and improve deterrence only by bringing damages back to the actual fraud perpetrators. Conversely, officers, executives, and insiders are often able to escape from direct consequences of SCA thanks to both D&O insurance policies and the economic shield provided by the target company.

According to a slightly different vision, some authors support the idea that SCAs can be an effective tool for *private enforcement of market regulation*. In SCAs, private attorneys working on behalf of investors and seeking damage compensation can promote the public interest and law enforcement (see Burch 2008; Porrini and Ramello 2011). From this perspective, it must be noted that SCAs usually result in settlement amounts that are incomparably higher than the total amount of public monetary sanctions. Moreover, Cox et al. (2003) find that overlapping between SCAs and the SEC provisions appears infrequent. This supports the idea that public regulation and private litigation through SCA might be considered as complementary means to enforce securities laws. However, once again, if wrongdoers do not pay for their misbehavior, even the private enforcement rationale becomes weaker.

A further relevant limitation of the deterrence function of SCA – and consequently of its aptitude to be a private enforcement means – is related to the fact that lawyers proceed only when expecting large recovery and then sufficiently large fee awards. For this reason, SCAs typically target only *deep-pocket* firms (large and rich corporations). This means that an SCA cannot be considered as an overall means of private enforcement since it does not deter misbehavior in small and not very well capitalized firms.

In the same fashion, some authors suggest that SCAs in the banking sector might be numbered among the *market discipline* tools. In fact, many types of abuses and negligence in the banking

sector not only affect bank safety and stability but also fall in the SEC jurisdiction and cause SCAs: if investors – or their counsels – recognize that a bank has taken on pathological risk or engaged in abuses like insufficient capitalization, negligent misappraisal of risk exposure, or leverage, they may try to control it through an SCA. When an SCA takes place in the banking sector, it acts as a tool to empower private investors to collect information, thus exerting an effective influence on targeted banks (Dalla Pellegrina and Saraceno 2011).

Even though the SCA is a very controversial procedure, a change in perspective provides it with a new rationale. Some authors, indeed, suggest that, though not traditionally included among corporate governance instruments, SCAs have relevant *corporate governance* effects.

From this perspective, some points need to be raised. First, an SCA is a critical event in the life of a corporation: filing, certification or dismissal, and settlement represent public sensitive information. Second, but in the case of palpably frivolous SCA, filing implies that allegations of negligent or fraudulent behavior possibly harmful to both investors and corporate value are ascribed to the corporation itself and its management. Third, an SCA is an unpleasant – primarily for the targeted company and its stakeholders including those who proceed against the company – result of some managerial choices (Johnson et al. 2000). Since securities litigation usually has a negative impact on stock prices and market perception of the targeted company, it is reasonable to think that SCAs is used to put pressure on reluctant boards of directors to rectify poor performance.

Because of all these reflections, not surprisingly, an SCA seems able to produce effects in terms of *corporate control*. In particular, it can be seen as a means for both shareholders and bondholders to influence the governance of a company. Plausibly, SCAs intervene when other standard corporate governance tools fail. This may happen when the interests of controlling shareholders and CEOs are aligned but in contrast with those of minority/dispersed shareholders and/or bondholders. Otherwise, shareholders would use less costly *voice* tools to realign their interests with

those of the CEO (Dalla Pellegrina and Saraceno 2016). In this respect, SCAs provide minorities with a sort of *voice*, besides the usual *exit* tools. According to Strahan (1998), SCAs are useful devices to manage residual agency problems by inducing managerial turnover. In general, the literature suggests that they are likely to both reinforce monitoring on managerial activities and induce disciplinary takeovers, cuts in corporate compensations, and CEO turnover (Ferris et al. 2007; Fich and Shivdasani 2007; Humphery-Jenner 2012).

In light of the discussion above, it becomes particularly important to consider the empirical evidence about the relations between the occurrence of an SCA and features and behavior of the target company.

Empirical Evidence About Causes and Effects of SCAs

As it will be clearer to the reader by reading this section, the behavior of a target company can both *cause* SCAs and *change because of* SCAs.

Factors Affecting SCAs

The empirical literature on SCAs well established that filings are more likely against *deep-pocket* corporations, on the one hand, appearing financially vulnerable, riskier, inefficient, and/or characterized by weak governance, on the other hand.

The significant and positive relation between the incidence of SCAs and firm's assets has been observed by the empirical literature, suggesting that class lawyers decide to proceed only against firms having sufficient resources to pay a possible settlement (see, among others Francis et al. 1994; Ferris and Pritchard 2009; Gande and Lewis 2009).

According to several studies, attorneys target firms with highly volatile stock return histories. On the one hand, the susceptibility to be targeted by an SCA results to be positively related to low stock returns (see among others, Ferris and Pritchard 2009; Gande and Lewis 2009). On the other hand, a higher market-to-book ratio is negatively related to litigation risk (Strahan 1998).

Gande and Lewis (2009) further show that low profitability and inefficient financial management significantly increase firm's litigation risk. Strahan (1998) documents a significant positive relationship between the propensity to be sued and firm's risk. This evidence is confirmed for the financial sector by Dalla Pellegrina and Saraceno (2011). High leverage and free cash flow are also associated with a higher probability of being sued (Ferris and Pritchard 2009).

The incidence of SCAs is higher in firms characterized by a greater likelihood of agency conflicts: poor corporate governance is often a catalyst for SCA (Strahan 1998; Ferris et al. 2007). In particular, there is evidence that incentives paid to executives in the form of options and bonuses are associated with a higher probability of being sued through an SCA (Peng and Röell 2008; Dalla Pellegrina and Saraceno 2016).

Furthermore, some authors find that nondividend-paying firms exhibit a significantly higher probability of being sued compared to companies which regularly pay dividends (Strahan 1998). On the contrary, Francis et al. (1994) find that firms being sued have higher dividend pay-out. Regarding board size, instead, no clear-cut evidence exists about the features of the board and the incidence of SCAs.

Effects of SCA

Concerning the impact of SCAs on the target companies, scholars identify both negative "external" effects (mainly related to the market reaction and reputational effects) and "internal" corrective effects of lawsuits.

A vast literature deals with the stock-price effects of SCA: filings are associated with significant negative price reactions (Griffin et al. 2004; Ferris and Pritchard 2009; Fich and Shivdasani 2007; Gande and Lewis 2009), whereas Ferris and Pritchard (2009) find no significant price reaction at the end of the procedure. This might suggest that the stock market immediately discounts the information related to the commencement of an SCA.

SCA can significantly heighten the cost of capital (both equity and debt) and, more in general, raising external capital becomes more

difficult for the target company because of the reputational losses related to the lawsuits (DuCharme et al. 2004; Deng et al. 2014).

SCA seems to significantly influence also investment decisions of the target company. McTier and Wald (2001) find that sued firms are more likely to increase cash holdings and leverage while decreasing overinvestment and payouts to shareholders. Dalla Pellegrina and Saraceno (2011) find that SCA induces target financial intermediaries to increase the accumulation of loan loss reserves, recapitalize, and reduce dividend pay-outs. They also provide evidence of a positive contemporaneous relationship between bank profitability – measured by either the Z-score or the Sharpe ratio – and the event of an SCA. However, the study also shows that, after few months, a more cautious attitude toward risk tends to replace the search for profits. Arena and Julio (2015) confirm both that target firms accumulate more cash in anticipation of future settlements and other costs related to litigation and reduce capital expenditures.

Many authors suggest that SCAs can enhance corporate disclosure requirements and managerial duties of loyalty and transparency, since they represent a residual monitoring device in the hands of shareholders (see, among others, Mitchell 2009). Empirical evidence supports the idea that shareholders use SCAs as a disciplinary force on managers and executives (Cheng et al. 2010).

CEO turnover has also been linked to SCA (Strahan 1998; Helland 2006; Fich and Shivdasani 2007; Humphery-Jenner 2012). Humphery-Jenner (2012) also provides insight that SCA can both induce disciplinary takeovers and make tougher prospects in the executive labor market for fired CEOs. These adverse effects, directly suffered by executives, may partially restore both the deterrence function and the private enforcement function of SCA. Indeed, Ferris et al. (2007) find improvements in the board of directors, including an increase in the proportion of external directors, while Humphery-Jenner (2012) and Dalla Pellegrina and Saraceno (2016) observe negative effects on executive compensation.

Conclusions

The main message to be learned from the dense literature on SCAs is that – because of the *circularity* mechanism that involves stakeholders as damaged claimants suing the company in which they have their own stake – these tools seem to be ineffective for the purpose of compensation and deterrence. It is undeniable, instead, that they can influence the performance of a company, causing stock price variations, changing the attitude towards risk undertaking, and affecting the reputation of its management. These effects, however, if strong enough, can return to the SCA a role of deterrence through a sort of market discipline mechanism. Several scholars are, therefore, increasingly convinced that SCAs are to be included among the corporate governance instruments.

Cross-References

- ▶ [Class Action and Aggregate Litigations](#)
- ▶ [Mass Tort Litigation: Asbestos](#)

References

- Andrews N (2001) Multi-party proceedings in England: representative and group actions. *Duke J Comp Int Law* 11:249–271
- Arena M, Julio B (2015) The effects of securities class action litigation on corporate liquidity and investment policy. *J Financ Quant Anal* 50:251–275
- Baker T, Griffith SJ (2009) How the merits matter: D&O insurance and securities settlements. *Univ PA Law Rev* 157:755–832
- Burch EC (2008) Securities class actions as pragmatic ex post regulation. *Georgia Law Rev* 43:63–131
- Cheng CSA, Huang HH, Lobo G (2010) Institutional monitoring through shareholder litigation. *J Financ Econ* 95:356–383
- Chung DH (2004) Introduction to South Korea's new securities-related class action. *J Corp Law* 30:165
- Coffee JC Jr (2006) Reforming the securities class action: an essay on deterrence and its implementation. *Columbia Law Rev* 106:1534–1586
- Coffee JC Jr (1995) Class Wars: The Dilemma of the Mass Tort Class Action, 95 *Columbia Law Review*, 1343–1465
- Cox JD, Randall TS, Kiku D (2003) SEC enforcement heuristics: an empirical inquiry. *Duke Law J* 53: 737–780
- Dalla Pellegrina L, Saraceno M (2011) Securities class actions in the US banking sector: between investor protection and bank stability. *J Financ Stab* 4:215–217
- Dalla Pellegrina L, Saraceno M (2016) Can shareholder litigation discipline CEO bonuses in the financial sector? The role of securities class actions. *Econ Notes* 45:3–36
- Deng S, Willis RH, Xu L (2014) Shareholder litigation, reputational loss, and bank loan contracting. *J Financ Quant Anal* 49:1101–1132
- DuCharme LL, Malatesta PH, Sefcik SE (2004) Earnings management, stock issues, and shareholder lawsuits. *J Financ Econ* 71:27–49
- Ferris SP, Pritchard AC (2001) Stock price reactions to securities fraud class actions under the Private Securities Litigation Reform Act. The John M Olin Center for Law & Economics, University of Michigan, WP No. 01–009
- Ferris SP, Jandik T, Lawless RM, Makhija A (2007) Derivative lawsuits as a corporate governance mechanism: empirical evidence on board changes surrounding filings 2001. *J Financ Quant Anal* 42:143–165
- Fich EM, Shivdasani A (2007) Financial fraud, director reputation, and shareholder wealth. *J Financ Econ* 86:2306–2336
- Francis J, Philbrick D, Shipper K (1994) Shareholder litigation and corporate disclosures. *J Account Res* 32:137–164
- Gande A, Lewis CM (2009) Shareholder-initiated class action lawsuits: shareholder wealth effects and industry spillovers. *J Financ Quant Anal* 44:823–850
- Griffin PA, Grundfest JA, Perino MA (2004) Stock price response to news of securities fraud litigation: an analysis of sequential and conditional information. *Abacus* 40:21–48
- Helland E (2006) Reputational penalties and the merits of class-action securities litigation. *J Law Econ* 49:365–395
- Humphery-Jenner ML (2012) Internal and external discipline following securities class actions. *J Financ Intermed* 21:151–179
- Johnson MF, Kasznik R, Nelson K (2000) Shareholder wealth effects of the Private Securities Litigation Reform Act of 1995. *J Account Res* 5:2013–2233
- McTier BC, Wald JK (2001) The causes and consequences of securities class action litigation. *J Corp Finan* 17:649–665
- Miller AR (1979) Of Frankenstein monsters and shining knights: myth, reality, and the class action problem. *Harv Law Rev* 92:664–694
- Mitchell LE (2009) The “innocent shareholder”: an essay on compensation and deterrence in securities class-action lawsuits. *Wis Law Rev* 2009:244–293
- Murphy B, Cameron C (2006) Access to justice and the evolution of class action litigation in Australia. *Melb Univ Law Rev* 30:399–440
- Peng L, Röell A (2008) Executive pay and shareholder litigation. *Rev Finance* 12:141–184
- Porrini D, Ramello GB (2011) Class action and financial markets: insights from law and economics. *J Financ Econ Policy* 3:140–160

- Ramello GB (2014) Class action and aggregate litigations. In: Encyclopedia of law and economics. Springer, New York, pp 1–7
- Sarra JP, Pritchard AC (2010) Securities class actions move north: a doctrinal and empirical analysis of securities class actions in Canada. *Alta Law Rev* 47:881–926
- Strahan PE (1998) Securities class actions, corporate governance and managerial agency problems. Federal Reserve Bank of New York. Federal Reserve Bank of New York, New York. WP June-212

Seized Money

► Counterfeit Money

Self-interest

► Intrinsic and Extrinsic Motivation

Separation of Power

Katarzyna Metelska-Szaniawska
Faculty of Economic Sciences, University of
Warsaw, Warsaw, Poland

Definition

Separation of powers is a principle pertaining to state governance, envisaging that responsibilities of the government are divided into distinct branches, in order to assure that none of these branches exercise the key functions of another. In consequence, state power is deconcentrated, and checks and balances are provided for. In economics, this principle is primarily the focus of the research program of constitutional economics (and, more broadly, public choice), emphasizing its role as a form of specialization, as well as a source of constraints enhancing the constitutional commitment mechanism. We present the theoretical economic approaches to the separation of powers, illustrating them with examples of empirical studies.

Introduction

For centuries, the structure of power in the state raised the interest of political and social thinkers, as well as practitioners. Separation of powers constituted an important focus of these considerations. The crucial aspect of this principle involves distinguishing generically distinct spheres of state action and providing that a separate body, or a group of state bodies, is responsible for operation in each of these areas. In this way, concentration of power is prevented, which otherwise could potentially lead to power abuse and excessive restraints on fundamental rights and liberty. A classic element of the principle of separation of powers is the system of checks and balances. As opposed to complete concentration (fusion) of power, separation of powers is considered an essential element of a democratic political system and “good” governance.

The doctrine of the separation of powers has its origins in ancient Greece, where the idea of governmental functions, as well as the theories of mixed and balanced government, first developed (for a detailed historical account, see Vile 1967). The term “separation of powers” itself is attributed to the eighteenth-century French lawyer and political philosopher Montesquieu and can be found in his famous *Spirit of the Laws* (1748). According to Montesquieu’s model, state authority should be divided into three branches with distinct responsibilities: the legislative, executive, and judicial powers. The first is responsible for enacting laws and the second is for their implementation, while the third is for interpreting and applying them. To promote liberty most effectively, these three powers must be separate and act independently. The tripartite division is, however, not the only possibility – other approaches have also been proposed in political–legal thought over the centuries (e.g., concepts including separate monarchy and/or local municipality powers). In fact, the structure of power in different countries is usually much more complicated. For example, the executive branch is often entitled to create certain types of legal acts or, at least,

to initiate legislative proceedings. Several areas of government activity escape the classical divisions, e.g., government performance auditing, independent central banks, and other independent agencies such as media regulators. Furthermore, *de facto* separation of powers depends not only on legal solutions contained in countries' constitutions but also on the actual political situation in a country.

Economic analysis of the separation of powers is primarily the domain of constitutional political economy (constitutional economics) or, more broadly, public choice. Viewed from an economic perspective, separation of powers is "a structural element, which is a necessary (but not sufficient) condition to protect individual liberty when citizens (principals) have delegated powers to political decision makers (agents) in order to benefit from specialization (division of labor)" (Salzberger and Voigt 2009, p. 198). As these authors emphasize, this concept may operate at various levels: (1) the separation of functions (i.e., separation between the constitutional and post-constitutional stages, as well as between "rule making, rule execution and rule adjudication" in the latter stage), (2) the separation of agencies (between the legislature, executive, and judiciary active in the post-constitutional stage, as well as the constitutional court which enforces the constitution in the first place), (3) the separation of persons, and (4) the structure of interrelations between the divided branches of powers (either pure separation or checks and balances and the latter either on the level of decision making as, generally, in presidential systems or on the level of persons as in many parliamentary systems). Constitutional economics sees the main function of the constitution for the economy as a commitment mechanism; therefore, in studying the economic effects of constitutions, proponents of this approach devote particular attention to constraints which the constitution places on those in power in order to ensure that the promises that they make do not remain mere empty declarations. A fundamental source of such constraints is the separation of powers.

In this entry, we present the theoretical economic approaches to separation of powers, illustrating some of them with examples of empirical studies.

Economic Approaches to the Separation of Power

In theoretical economic studies, the problem of separation of powers is viewed from various perspectives. The latter are then often allowed to interpenetrate themselves as fundamental frameworks for more detailed empirical studies of the determinants and economic effects of separation of powers. Therefore, we supplement the theoretical discussion provided below with some examples of empirical studies.

Firstly, the constitutional principle of separation of powers is approached by economists as a certain type of specialization, where the legislature constitutes a forum for political bargaining at relatively low transaction costs, the hierarchically constructed executive allows for mitigating the costs of law implementation, and the judiciary supplies neutral interpretation of law ensuring that commitments of members of the legislature made during the bargaining process gain credibility.

Secondly, separation of powers is treated as separation of a political cartel. This classical approach of economists to the analysis of the structure of power within the state stems from the assumption that from an economic point of view the state is a monopolist in the exercise of institutionalized coercion. Separation of powers causes destabilization of the political cartel that could emerge in the event, whenever a certain group (e.g., political party) took control over this monopoly. To identify the particular effects of introducing separation of power for various actors within the state, it is, however, crucial to distinguish whether this separation resembles horizontal or vertical (functional) separation of a monopoly (this important observation is particularly attributed to Brennan and Hamlin (1994)). Economists typically agree that while executive–legislature relations, bicameralism,

and independence of the judiciary are mainly expressions of horizontal separation, federalism and decentralization are considered primarily as forms of vertical separation of powers. Empirical studies concentrate, in particular, on economic effects of presidential versus parliamentary systems (e.g., Persson and Tabellini 2003; Blume et al. 2009), as well as decentralization (e.g., Thießen 2003; Enikolopov and Zhuravskaya 2007). For more examples and a further discussion, see Voigt (2011).

Thirdly, separation of powers is analyzed in relation to the model of demand and supply of legislation and the role of interest groups in this setting. Two conflicting views emerge in this respect from an old debate on the constitution-making in the United States (see, e.g., on the one hand Beard (1913/1986), Posner (1986), and Tollison (1988) and on the other hand Macey (1987)). The first one regards separation of powers as the centerpiece of interest group bargains which the constitution itself is thought to represent. The effects of this are that the constitution hinders progressive change and the general society's voice in politics (Beard 1913/1986), increases the persistence of special-interest legislation (e.g., Crain and Tollison 1979), and ultimately furthers transfer of wealth from the general population to well-organized special-interest groups. According to an alternative view, separation of powers constitutes the basis for a "credible constitutional method of constraining rent seeking" undertaken by special-interest groups (Macey 1987, p. 76). Constitutions, establishing the structure of power, lay down the meta-rules that interest groups must conform to in order to successfully pursue the enactment of laws that they seek for their clients. These rules, including in particular the separation of powers, determine the level of transaction costs that such groups must bear when they attempt to accomplish the enactment of desired legislation. The higher these costs are at the constitutional level (i.e., the more it costs an interest group to obtain desired legislation), the lower the supply of special-interest legislation in a society and the less detrimental their activity to the working of the economy overall.

Fourthly, separation of powers also affects policy making by exerting influence upon the manner of exercising power within the state. Instead of hierarchical relationships, it introduces equivalency of actors and the necessity to bargain in order to implement new solutions. The need for cooperation between branches of power causes that they begin to act strategically vis-à-vis each other. The view at separation of powers from the perspective of bargaining models, describing strategic interactions between veto players (organs, powers, actors) influencing policy decisions, shows in what way these decisions depend on preferences of branches of power participating in the enactment of laws and the structural rules of these interactions defined in the constitution (for a detailed overview, see Cooter 2000).

Finally, literature concerning the significance of the number of veto players for policy making and its effects is naturally closely bound with the discussion presented here relating to constitutional separation of powers. In fact, Voigt (2011, p. 234) refers to the number of veto players in the state as "the new separation of powers." According to this approach, policy change or reform is more difficult when there exist numerous veto players. Each additional veto player, understood as a political actor capable of blocking policy change, "either shrinks the set of policies that can defeat the *status quo* or leaves it unchanged" (Gehlbach and Malesky 2010, p. 957). Therefore, policies are expected to be more stable when the number of veto players increases (Tsebelis 1995, 2002). One should however bear in mind that this may be desirable when commitment to a given policy is needed (Keefer and Stasavage 2003), but it will be unfavorable when policy change is required (Cox and McCubbins 2001). To address this problem, Gehlbach and Malesky (2010) provide a formal model and show that as the number of veto players increases, it becomes less likely that policies bringing benefits to special-interest groups will supersede the *status quo*. This is because such narrow, organized groups must compensate each veto player for opting for policy change instead of preserving the *status quo*. In certain contexts, this implies that changes in policy will actually be

greater when veto players are numerous than when their number is not large. Empirical literature initiated by Henisz (2000) argues that higher numbers of veto players correspond to better economic performance in terms of investment (more veto players result in stronger political constraints and less discretion in political decision making) and, ultimately, also economic growth.

An important place in the considerations of separation of powers from an economic perspective is devoted to the problem of independence of the judiciary and its position within the structure of power in a state. In this context, an independent (constitutional) judiciary is often regarded as the final stage in converting promises made by representatives of state power into credible commitments or a solution to the “dilemma of the strong state,” i.e., protection of citizens from the discretionary infringement of their rights by the state (Feld and Voigt 2003; Voigt et al. 2015). An independent judiciary can also be regarded as an additional veto player in the structure of power and therefore may constitute a further constraint on this power enforcing credibility of policy makers’ commitments. Empirical studies (e.g., conducted by the aforementioned authors) confirm the relevance of *de facto* judicial independence for economic growth in a global sample; this factor is also found to matter for successful economic transition of post-socialist countries of Europe and Asia (see Metelska-Szaniawska 2016).

Theoretical and empirical economic studies also concern the delegation of powers to independent bodies within the system of powers in the state which escape the classical tripartite division, such as central banks (in particular, in the context of their relevance for inflation – see, e.g., de Haan and Klomp 2010), prosecutors (e.g., van Aaken et al. 2010), auditing bodies (e.g., Blume and Voigt 2011), and other regulatory agencies specializing, e.g., in supervising liberalization, privatization, and regulation (in the field of telecommunications, the media, electricity, food safety, etc.). The establishment of independent regulatory agencies is based upon the assumption that their independence will create peer pressure and reputational incentives, leading to more

expertise-based decisions instead of the pursuit of short-term policy-driven objectives (Majone 2001). The independence of such regulators, perceived as trustees acting in the public’s interest and enhancing credible commitment of the government, rather than being mere agents of political parties in power, eventually leads to a higher quality of regulatory decisions.

Conclusions

Economists have developed a number of approaches to the analysis of separation of powers, which considerably enrich the traditional constitutional–legal scholarship. Some of these studies focus on providing new theoretical arguments and strengthen (or, in some cases, question) this principle’s conventional justification. Other works, including empirical analyses, bring valuable original findings regarding the manifold consequences of the functioning of separation of powers in constitutional systems around the globe, linking developments from the political sphere with social and economic outcomes. While, as presented in this entry, many fundamental issues have already been resolved, as new questions keep arising (concerning, e.g., supranational governance structures), the future outlook for research in this area is promising.

Cross-References

- [Constitutional Political Economy](#)
- [Independent Judiciary](#)

References

- Beard Ch (1913/1986) *An economic interpretation of the constitution of the United States; with a new introduction by Forrest McDonald*. Free Press, New York
- Blume L, Voigt S (2011) Does organizational design of supreme audit institutions matter? A cross-country assessment. *Eur J Polit Econ* 27(2):215–229
- Blume L, Müller J, Voigt S, Wolf C (2009) The economic effects of constitutions: replicating – and extending – Persson and Tabellini. *Public Choice* 139:197–225

- Brennan G, Hamlin A (1994) A revisionist's view of the separation of powers. *J Theor Polit* 6(3):345–368
- Cooter RD (2000) *The strategic constitution*. Princeton University Press, Princeton
- Cox GW, McCubbins MD (2001) The institutional determinants of economic policy outcomes. In: Haggard S, McCubbins MD (eds) *Presidents, parliaments, and policy*. Cambridge University Press, Cambridge, pp 21–63
- Crain WM, Tollison RD (1979) Constitutional change in an interest-group perspective. *J Leg Stud* 8(1):165–175
- de Haan J, Klomp J (2010) Inflation and central bank independence: a meta-regression analysis. *J Econ Surv* 24(4):593–621
- Enikolopov R, Zhuravskaya E (2007) Decentralization and political institutions. *J Public Econ* 91(11):2261–2290
- Feld LP, Voigt S (2003) Economic growth and judicial independence: cross-country evidence using a new set of indicators. *Eur J Polit Econ* 19(3):497–527
- Gehlbach S, Malesky EJ (2010) The contribution of veto players to economic reform. *J Polit* 72(4):957–975
- Henisz WJ (2000) The institutional environment for economic growth. *Econ Polit* 12(1):1–31
- Keefer P, Stasavage D (2003) The limits of delegation: veto players, central bank independence, and the credibility of monetary policy. *Am Polit Sci Rev* 97(3):407–423
- Macey JR (1987) Competing economic views of the constitution. *George Wash Law Rev* 56(1):50–80
- Majone G (2001) Two logics of delegation: agency and fiduciary relations in EU governance. *Eur Union Polit* 2:103–122
- Metelska-Szaniawska K (2016) *Economic effects of post-socialist constitutions 25 years from the outset of transition. The constitutional political economy approach*. Peter Lang International Academic Publishers, Frankfurt am Main
- Persson T, Tabellini G (2003) *The economic effects of constitutions*. The MIT Press, Cambridge, MA
- Posner R (1986) *Economic analysis of law*, 3rd edn. Little, Brown and Company, New York
- Salzberger E, Voigt S (2009) Separation of powers: new perspectives and empirical findings – introduction. *Const Polit Econ* 20(3):197–201
- Thießen U (2003) Fiscal decentralization and economic growth in high income OECD countries. *Fisc Stud* 24(3):237–274
- Tollison RD (1988) Public choice and legislation. *Va Law Rev* 74:339–371
- Tsebelis G (1995) Decision-making in political systems: veto players in presidentialism, parliamentarism, multicameralism and multipartyism. *Br J Polit Sci* 25(3):289–325
- Tsebelis G (2002) *Veto players: how political institutions work*. Princeton University Press, Princeton
- van Aaken A, Feld LP, Voigt S (2010) Do independent prosecutors deter political corruption? An empirical evaluation across seventy-eight countries. *Am Law and Econ Rev* 12(1):204–244
- Vile MJC (1967) *Constitutionalism and the separation of powers*. Clarendon Press, Oxford
- Voigt S (2011) Positive constitutional economics II – a survey of recent developments. *Public Choice* 146(1–2):205–256
- Voigt S, Gutmann J, Feld L (2015) The effects of judicial independence 10 years on: cross-country evidence using an updated set of indicators. *Eur J Polit Econ* 38:197–211

Sex

Samuel Cameron

Division of Economics, University of Bradford,
Bradford, UK

Definition

This entry looks at the law and economics of sex defined as the exchange of sexual acts which may be intercourse for procreative reasons or various forms of recreational exchange. The key issue in regulation is the matter of consent although this has frequently been attenuated by religious and moral considerations.

Commodity or Service

The economic analysis of sex is to treat it as a commodity or service. This is now reflected in the economic literature on happiness some of which includes frequency of sexual outlets as a determinant of subjective well-being (Blanchflower 2003). Much of early and current literature on the law and economics of sex is concerned with the normative issue of whether prostitution should be legal or not.

Some would take the position that sexual services are inalienable rights which should not be traded. This can also be applied to the selling of one's eggs or semen to others to overcome infertility. The regulatory issue here is whether money should exchange hands in return for sexual services. The other key transactional issue, in the area of sex, is consent between the parties. In the

Samuel Cameron has retired.

case of prostitution, the presence of monetary exchange, in itself, implies consent. Prostitution represents a conventional economic market where sex is bought and sold. However, making it illegal makes it difficult to draw up and enforce contracts regulating the transaction (see Della Giusta (2010); Jakobsson and Kotsadam (2013); Kotsadam and Jakobsson (2014); Immordino and Russo (2015); Scandizzo and Ventura (2015); Cho (2016)). Sexual transactions, outside of prostitution, operate in an informal market where there may be implicit contracts. That is, gifts or economic support, for example, may be given with an expectation of sexual reciprocation. The ambiguity of such situations may be connected to the issue of rape which can be seen as a transaction completed by coercion (physical or otherwise) against the will of one of the parties thus meaning there was no consent.

Sex Market

The implicit sex market is further complicated by the institution of marriage. Sexual trade may reflect an enforced case of unjustly unequal exchange where individuals are driven by a power imbalance into surrendering their inalienable rights. This takes us to the Marxist-Feminist position that marriage has been, in many times and places, a form of prostitution hidden behind a veil of respectability.

Marriage is traditionally connected to non-recreational sex that is sex for the purpose of producing progeny. One economic view of marriage is that it may be an efficient means of internalizing transactions costs, of obtaining recreational sex, in an ongoing relationship. We can see a logical economic rationale for gay marriage, in addition to any arguments of gender equity. This may be contrasted with the striking remarks of economist and Supreme Court Judge Richard Posner that it is not rational to choose to be homosexual (Posner 1992).

Traditional legally sanctioned monogamous marriage is a means of internalizing transactions costs, of sexual transaction, but this is not the purpose for which it was devised. The origins of

state-regulated marriage arrangements lie in religions some of which support nonmonogamous outcomes. It is also geared primarily toward the rearing of children. It is logical that universal monogamous contracted marriage may not be the most socially efficient institution. Accordingly we find other arrangements most notably cohabitation but also ad hoc adultery/infidelity, polygamy, LATs (living alone together separately), polygyny, orgies, polyandry, and “swinging” (organized temporary partner swapping within a group of couples). We also find other informal but semiorganized markets such as the case of “groupies” in popular music and showbiz where the groupie fan might offer sex as an informal payment to a paid protector of the star as a means of increasing the chances of offering free sex to the star.

Institutional Forms of Sex Trade

The use of different institutional forms of sex trade can be attributed to three interrelated factors – frequency of sexual activity, desire for variety of partners, and desire for variety of sexual performance. Partners may differ with respect to desired frequency of sexual activity. If this is so in a monogamous union, we have a disequilibrium in a situation of bilateral monopoly. A solution to this can be sought by recourse to either a secondary informal market – what is conventionally termed “adultery” or “infidelity” or may more neutrally be termed “partner out trading” (Cameron 2002) or the formal market of prostitution. These are consensual extended markets. The disequilibrium might also be seen as a driving force in nonconsensual additions to sexual activity, that is, rape. There is a long historical tradition of arguing that prostitution might be socially desirable as a means of providing a substitute good for rape for the disequilibrated consumer. The same arguments about frequency apply to desire for variety of partners.

Desire for a variety of partners may, to some extent, be a desire for variety of sexual acts as the designated partner may not wish to engage in the full menu of sexual acts demanded by the variety-

seeker. Thus we have types of sex. Economists can treat each type as a separate commodity or service, or they can regard the different types as product differentiation where each type has different characteristics in different proportions. Partners implicitly trading sex can differ in terms of their desires along the spectrum of such characteristics. They can also differ in their preferences for the relative amount of recreational versus reproductive sex. This will be reflected in conflict about the optimal amount of use of contraception.

Sex is not a unified product and can be performed in a number of different ways. These include buggery, bondage, use of capital goods (such as sex toys) for enhancement, and sadomasochism. Some aspects of these have regularly been subject to strong legal regulation based on social norms rather than law and economics reasoning.

Much of the historical debate around the status of buggery is related to issues of sexual identity and religion. In England, the Buggery Act of 1553 was predicated on the idea that unnatural sexual acts against the will of God should be prohibited. The Act was not fully repealed until 1967 when homosexual usage ceased to be an offence. The US Supreme Court ruled in 2003 that sodomy laws were unconstitutional; however, some states and the US military continued to attempt to have their own sodomy laws. In Germany, the laws were repealed in the late 1960s but not fully (i.e., removal of punishment of homosexuals) until 1989 (East) and 1994 (West).

Sadomasochism has proved a problematic area because of the matter of consent (Bennett 1994). Traditional microeconomic theory and welfare economics would seem to suggest that consensual sadomasochistic sex is potentially a socially efficient arrangement. The landmark case for the contest of this view is the “Spanner” action (the case of *RvBrown*) in the UK (Thompson 1994). The appeal against the decision to convict consensual sadomasochists for assault was heard in March 1993. The backdrop to this was that SM was not illegal as there was no law on it, but there was a precedent in the Donovan case of 1934 where a man had lost his appeal against an assault

conviction. In that case, the other party had consented to the treatment and continued to maintain this consent. There was also no accusation of any kind of what we would now call “grooming” which would fall under the category of coercion and thus erode consent. There has been continued debate about this case, but the UK continues to maintain the stance of illegality in contrast to more liberal regimes like Germany and Australia. In the USA, there is a lack of a clear federal law, but individual states continue to prosecute under the same assault grounds, as in England, even when consent is clearly present.

So, the basis for these convictions as assault, rather than legitimate sexual exchange, would seem to be on grounds of inalienable body rights as with blood, semen, etc. That is, an individual does not have the right to choose to be assaulted for the purpose of sexual gratification. The case of bondage without sadomasochism is different if an individual who has been rendered unable to move during the sexual acts has consented to be rendered immobile, then they cannot be deemed to have been coerced.

Any of the above variations in sexual activity is subject to regulation on grounds of by-products of them being “indecent” if the assault claim is not used or cannot be used. For example, if the parties produce any media (such as films or writing) as a by-product of these activities, these can be treated as pornography.

There are, as we have seen, a wide variety of formal and informal market arrangements for sexual trade. The Chicagoan law and economics approach would argue that the regulation of these should seek to produce a social welfare-maximizing outcome. In practice, this is attenuated by the historical legacy in legal systems of religious and ethical traditions that may not necessarily conform to socially rational choice.

Cross-References

- [Commercial Sex and Health](#)
- [Crime: Organized Crime and the Law](#)
- [Prostitution, Demand and Supply of](#)
- [Sex Offenses](#)

References

- Bennett T (1994) Sadomasochism under the human rights (sexual conduct) Act 1994. *Sydn Law Rev* 35:541–563
- Blanchflower DG (2003) Money, sex and happiness: an empirical study. *Scand J Econ* 106(3):393–415
- Cameron S (2002) The economics of partner out trading. *J Bioecon* 4:195–222
- Cho S-Y (2016) Liberal coercion? Prostitution, human trafficking and policy. *Eur J Law Econ* 41(2):321–348
- Della Giusta M (2010) Simulating the impact of regulation changes on the market for prostitution services. *Eur J Law Econ* 29(1):1–14
- Immordino G, Russo FF (2015) Laws and stigma: the case of prostitution. *Eur J Law Econ* 40(2):209–223
- Jakobsson N, Kotsadam A (2013) The law and economics of international sex slavery: prostitution laws and trafficking for sexual exploitation. *Eur J Law Econ* 35(1):87–107
- Kotsadam A, Jakobsson N (2014) Shame on you, John! Laws, stigmatization, and the demand for sex. *Eur J Law Econ* 37(3):393–404
- Posner RA (1992) Sex and reason. Harvard University Press, Cambridge, MA
- Scandizzo PL, Ventura M (2015) Organized crime, extortion and entrepreneurship under uncertainty. *Eur J Law Econ* 39(1):119–144
- Thompson B (1994) Sadomasochism. Cassell, London

sex offender postrelease regulations. With respect to postrelease regulations, we discuss potential reasons why sex offenses are treated differently and review the literature on the theoretical and empirical effects of these policies for crime rates and recidivism as well as their collateral consequences.

Synonyms

[Child molestation](#); [Child pornography](#); [Crimes against nature](#); [Prostitution](#); [Rape](#); [Sexual assault](#); [Sexual offenses](#)

Definition

The commission of an act that is defined by the law as sexual (usually involving a sexual act or touching or display of a part of the body typically associated with sexual behavior) and to which the victim has not given consent or is unable under the law to give consent. The legal absence of consent often turns on the use of force or threat of force against the victim or another person; the unconsciousness, impairment, or intoxicated state of the victim; the use of fraud in obtaining consent; the age of the victim; or the use of illegal means of obtaining consent (e.g., money). This class of crimes includes rape, statutory rape, child molestation, sexual assault, prostitution, peeping-tom offenses, and the creation, possession, and distribution of child pornography.

Sex Offenses

Amanda Agan¹ and J. J. Prescott²

¹Princeton University, Princeton, NJ, USA

²Law School, University of Michigan, Ann Arbor, MI, USA

Abstract

Sex offenses today may be generally defined as acts of a sexual nature to which a victim has not given legal consent. This broad definition covers many different types of behaviors, but almost all individuals in the U.S. (and in a few other jurisdictions) who are convicted of committing any crime in this category are subject to unique postrelease regulations such as registration, public notification, and residency restrictions. In this entry, we briefly examine the scope of sexual offending in the U.S., the utility of the economic model of crime in understanding sex offender behavior, and the rise of

Introduction

Sex offenses constitute an ancient and important category of crime. Over the course of history, many kinds of sexual acts have been criminalized, including encounters in which consent was not at issue, such as sodomy, adultery, homosexuality, and fornication. In a large number of jurisdictions today, acts of this sort are no longer criminally prohibited; the criminalization of a sexual act instead turns primarily on lack of consent. For purposes of this entry, a sex offense is defined as

the occurrence of a sexual act, a sexual touching, or a display of body parts considered to be sexual in nature, coupled with actual or legally imposed lack of consent on the part of the victim or society.

This definition evolved from a prototypical sex offense, rape, which was defined in the English Common Law as “the carnal knowledge of a woman [by a man] forcibly and against her will” (Tucker 1803, p. 209). Lack of consent is no longer defined to include only those situations in which an offender uses physical force to overcome active resistance by a victim. Depending on the jurisdiction, lack of consent may include anything short of actual verbal and physical consent to the activity, and even actual consent is not sufficient when the victim is considered legally incapable of giving consent. Intercourse is not a required element of many sex offenses; touching, display, or even the possession of a display can suffice. Increasingly, sex offense statutes treat men and women symmetrically.

As a class, sex offenses include crimes as varied as rape, prostitution, child molestation, possessing child pornography, and peeping-tom behavior. Despite this diversity, in the United States, those who commit these crimes, upon conviction and release, are often subjected to the same novel and extensive postrelease regulations, which may include registration, public notification, civil confinement, and residency restrictions, among others. The scope and severity of these laws makes studying their predicted and actual consequences not just academically interesting, but important from a policy perspective. Perhaps it is not surprising that the traditional economic approach to crime has been questioned when applied to sex offender behavior, as sex crimes appear to some to result from inexplicable impulses or other irrational drives and thus may be undeterrable. This distinction may explain why convicted sex offenders are both uniquely and uniformly subject to these unorthodox postrelease regulations: if sex offenders are irrational and incorrigible, postrelease laws ought to focus on incapacitation rather than deterrence.

This entry will begin in Part II by providing some rough statistics on the relative frequency and diversity of sex offenses, primarily relying

on data from the U.S. Part III will briefly recount the origin and evolution of postrelease sex offender regulations and the current state of these approaches in the U.S. and other countries. Part IV will apply the traditional economic model of crime to sex offending behavior, with a focus on first describing and then understanding the advantages and disadvantages of the postrelease regulation of sex offenders. An important preliminary to this discussion will be the basic assumptions of the economic model of crime and how alternative models of sexual deviance can be useful for understanding relevant policy developments, particularly why punishments for sex offenses often differ from those of other crimes. Finally, Part V will describe the major studies and the current state of knowledge on the broad impact of sex offense laws, with a focus on the consequences of postrelease laws on crime rates, victim behavior, the geography of victimization, offender welfare, and other collateral consequences, such as housing price effects.

The Relative Frequency of Sex Offenses

Violent crime rates in the U.S. plunged over the course of the 1990s and have continued their general decline throughout the 2000s. Sex offense rates have followed roughly the same trend. Data from the FBI’s Uniform Crime Reports (UCR) show that reports of rape incidents declined from 39.2 per 100,000 in 1994 to 25.2 per 100,000 in 2013. Furthermore, the number of persons arrested for rape, prostitution, and other sex offenses each fell by over 30% between 2004 and 2013: specifically, arrests for rape fell by 30.3%, arrests for prostitution and commercialized vice fell by 35.8%, and arrests for all other sex offenses fell by 34.4% (United States Department of Justice 2014).

Not all crime incidents are reported to the police, however, and sex offense statistics may be particularly vulnerable to downward reporting bias. Indeed, according to data from the Bureau of Justice Statistics’ (BJS) National Crime Victimization Survey (NCVS), the fraction of rapes and sexual assaults reported to police was similar to

other crimes in 2003 at 56%, but by 2011, this rate had fallen precipitously to 27%, which is significantly lower than reporting rates for other violent crimes (Truman et al. 2013, Table 1). Nevertheless, NCVS data also show a decline of even greater magnitude in *victimization* rates for rape or sexual assault – from 4.3 victims of rape/sexual assault per 1,000 in population over age 12 in 1993 to 1.3 per 1,000 in population over age 12 in 2012 (Truman et al. 2013, Fig. 3).

Many distinct crimes are included under the broad definition of sex offenses; unfortunately, data on sex offenses are often only collected and/or labelled at relatively aggregate levels, such as “rape” or “other sexual offenses.” For instance, UCR incidents are reported only for rape; UCR arrests are reported for rape, prostitution, and other sexual offenses; and NCVS victimization data include rape, sexual assault, and sexual contact with or without force. As a consequence, identifying and measuring trends with respect to specific sex offenses in the U.S. can be difficult. The FBI’s National Incident Based Reporting System (NIBRS) collects information on a more detailed list of offenses (specifically, forcible rape, forcible sodomy, sexual assault with an object, forcible fondling, statutory rape, incest, prostitution, and pornography), but even as of 2012, NIBRS covers only 33% of all UCR reporting agencies, making reliable comparisons over time difficult.

Sex offenses may be committed by first-time sex offenders or by recidivist sex offenders who have committed similar crimes in the past. This distinction is of course not unique to sex offenses, but it is an important one in this context because many laws in the U.S. specifically target reducing sex offender *recidivism* rather than deterring first-time offenders. Using administrative data from New York State, Sandler et al. (2008) show that 95% of sex offense arrestees between 1986 and 2006 were *first-time* sex offenders. More comprehensive evidence exists on the fraction of convicted sex offenders who commit another sex crime after leaving prison: using detailed information on all sex offenders released from prison in 15 states in 1994, Langan et al. (2003) find that, within 3 years of release, 5.3% of these previously

convicted sex offenders had been rearrested for a sex crime. The number was somewhat lower for convicted child molesters, 3.3% of whom were rearrested for another sex crime against a child within 3 years of release. To provide some measure of comparison, the overall same-crime rearrest rate for robbery was 13.4% and for assault 22.0%; homicide was smaller at 1.2% (Langan and Levin 2002). Overall sex offender recidivism, as opposed to same-crime recidivism, was much higher, with 43% of released sex offenders being rearrested for some crime within 3 years. However, even this rate is substantially lower than the analogous rate for other kinds of convicted offenders, who had an overall rearrest rate of 68%. There can, of course, be different definitions of recidivism and different follow-up periods, which might imply higher recidivism rates. For a comprehensive review of relatively recent literature on sex offender recidivism, see Soothill (2010).

Postrelease Sex Offender Laws

Even against a backdrop of declining sex offense rates and relatively low recidivism levels, the policies that legislatures and law enforcement agencies in the U.S. have employed in recent years to reduce the frequency of sex offenses differ significantly from those traditionally used to combat crime. Criminal punishment in the U.S. and other western countries almost exclusively involves incarceration (and postrelease supervision) or fines. While these forms of punishments are also deployed against sex offenders, post-release sanctions play a much larger role in policy discussions and scholarly research. These regulations were enacted in hopes of dramatically curtailing recidivism by essentially incapacitating convicted sex offenders through residential isolation and additional monitoring and by notifying the public about these individuals in the hopes that potential victims might take precautions to protect themselves.

The Rise of Sex Offender Laws in the U.S.

The first modern wave of concern over sex offenses in the U.S. ran from the 1930s through

the 1960s (Thomas 2011). According to Thomas, there is no evidence that this concern was prompted by an increase in sex offense rates; rather, the image of the “sexual psychopath” somehow gained currency and fear of and focus on sex offenses became commonplace. The first state-level registration laws specifically targeting sex offenders were enacted during this period. Registration laws can be generally defined as laws imposing a requirement that individuals at higher than average risk of committing a crime – i.e., individuals previously convicted of certain sex offenses – update law enforcement authorities at regular intervals with their current addresses, activities, and identifying information. Despite this earlier era of fear and legal innovation, it was only much more recently – not until the late 1980s and early 1990s – that state legislatures became especially interested in the threat posed by sex offenders, often as a result of well-publicized abductions of children by individuals who had previously been convicted of similar crimes. By 1989, 12 states had created sex offender registries.

U.S. congressional involvement in this area began in earnest in the aftermath of a 1989 attack on 11-year-old Jacob Wetterling, which led ultimately to federal sex offender registration legislation in 1994 (The Jacob Wetterling Crimes against Children and Sexually Violent Offender Registration Act). This law mandated that states build sex offender registries that, at a minimum, recorded the names and addresses of offenders convicted of offenses against children or sexually violent crimes for a minimum of 10 years, for the confidential use of law enforcement (and other government authorities, in some instances), with offenders required to verify their information every year.

In 1996, the U.S. Congress enacted Megan’s Law (the name of the law being a tribute to Megan Kanka, a young victim of a previously convicted sex offender), which effectively required the *release* of relevant sex offender registration information to the public. Such community notification laws make mandatory (or, in some cases, simply allow) the dissemination of registration information to the public, either actively or passively, with

the application of the law sometimes limited only to high-risk classes of offenders. Notification can occur in several different ways under these laws, including by allowing the public to physically inspect paper registries, by publishing notices in newspapers, and/or by actively informing potentially interested residents through the mail or through an in-person visit by a law enforcement officer. Primarily, however, community notification is handled via easily accessible webpages, which the public can use to search for offenders by name or for offenders living near particular addresses. Meanwhile, the Pam Lychner Sexual Offender Tracking and Identification Act of 1996 created a National Sex Offender Registry (NSOR) to address the difficulties that result from interstate migration by convicted sex offenders. Additional federal laws were passed during the late 1990s to extend the scope and increase the coverage of registration and notification and to address concerns related to child pornography on the internet.

The turn of the millennium saw the U.S. enact yet more extensive federal sex offender legislation. For instance, the Campus Sex Crimes Prevention Act of 2000 obligated all sex offender registrants taking courses or working at a university to affirmatively notify the administration of their registration status. The 2003 Prosecutorial Remedies and Other Tools to end the Exploitation of Children Today (PROTECT) Act required that states maintain online websites for their registries and worked to make these websites more accessible. In 2006, as part of the Adam Walsh Child Protection and Safety Act, the U.S. Congress enacted the Sex Offender Registration and Notification Act (SORNA), which further expanded the scope of registration and notification laws, and which developed a three-tier system of classifying sex offenders. Offenders are ranked based on their convictions; tier one offenders must register for 15 years minimum, second tier offenders for 25 years, and third tier offenders for life.

The Diversity and Intrusiveness of Sex Offender Laws in the U.S.

Although federal involvement has been substantial in this area, most innovation has occurred at

the state or local levels of government (indeed, the federal laws were built on state models). Post-release restrictions of seemingly every imaginable variety have either been experimentally implemented or currently exist in the U.S. at the state or local level.

For example, one common trend during the last 20 years has been the increasingly intensive use of sex offender residency restrictions. These restrictions are mandatory limits on where registered sex offenders may reside, with permissible areas defined by whether the residence is more than a minimum distance away from all places that have certain characteristics or fit into certain categories (e.g., schools, day care centers, etc.). As of 2014, approximately 33 states have implemented sex offender-related residency restrictions. A number of jurisdictions now also enforce lifetime GPS monitoring of convicted sex offenders. Other jurisdictions impose extensive employment restrictions. For example, in Louisiana, registrants cannot operate any bus, taxicab, or limousine, or work in any service position that would involve entering a residence (La. Rev. Stat. Ann. § 15:553). Travel restrictions are also common: in Illinois, registrants must provide law enforcement with a detailed travel itinerary if they intend to be away from their registered residence for more than two days (730 Ill. Comp. Stat. 150/3). Other jurisdictions ban registrants from public libraries, hurricane shelters or emergency shelters, or parks and playgrounds (see Iowa Code § 692A.113(f), Associated Press (2005), and Warren Mun. Code 22-140, respectively).

Other innovations by state and local authorities appear to be uncommon, but nonetheless intrusive and sometimes inexplicable. In Tennessee, certain registrants may not “[p]retend to be, dress as, impersonate or otherwise assume the identity of a real or fictional person or character or a member of a profession, vocation or occupation while in the presence of a minor” (Tenn. Code Ann. § 40-39-215(a)(1)). In California, among a number of other states, a registrant on parole may not live with any other registrant in a “single family dwelling,” although a statutory exception allows married or related registrants to cohabitate (Cal. Penal Code § 3003.5(a)). In Florida, Tennessee, and

Texas, sex offender registrants must carry special driver’s licenses or identification cards that clearly indicate that they are convicted sex offenders and that they are required to register (Fla. Stat. § 322.141; Tenn. Code Ann. § 55-50-353; Tex. Transp. Code Ann. § 521.057).

After being enacted, these restrictions have often been retroactively applied, meaning that many sex offenders released years – even decades – earlier are still required to comply, notwithstanding the fact that the law was passed many years after these offenders committed their offenses. Governments are free to do this because postrelease sex offense regulations are not formally considered punishments, although there is little doubt that they are punitive in effect and courts and scholars in the U.S. increasingly view them as forms of punishment. Rather, these post-release laws are considered regulations, designed to reduce the likelihood that a convicted sex offender will reoffend upon his release from prison. While registration, community notification, and residency restrictions may be the best known of these laws, perhaps the most extreme instance can be found in the broad civil confinement authority granted to corrections officials by some state legislatures. Specifically, because constitutional or fairness considerations bar continuing *punitive* incarceration beyond the length of an offender’s official sentence, legislatures have simply enacted civil confinement measures that allow authorities to continue to imprison a sex offender even after his sentence has expired if he is still deemed to pose a threat. The U.S. Supreme Court upheld this practice in *Kansas v. Hendricks*, 521 U.S. 346 (1997).

During this period of steady growth in the scope and intensity of postrelease sex offender regulation, many of those allegedly harmed by these laws have challenged their legality, principally arguing that their retroactive application is a violation of the Ex Post Facto Clause of the U.S. Constitution. In 2003, the U.S. Supreme Court set the tone for these kinds of arguments in *Smith v. Doe*, 538 U.S. 84 (2003), in which a majority of the Justices ruled that the Alaska registration law before the Court was not a form of punishment, but rather a regulation in the interests

of public protection and, therefore, could be applied retroactively. As the growth in the breadth (and number) of postrelease sex offender laws has continued apace in recent years, *ex post facto* legal challenges have met with more success, at least in some jurisdictions.

Sex Offender Laws: International Context

The U.S. is not alone in its growing alarm over sex offender recidivism and in at least contemplating the use of postrelease sanctions designed to reduce such recidivism. Thomas (2011) provides extensive detail on these international developments through 2010. At a minimum, the United Kingdom, Ireland, France, Canada, Australia, South Africa, Kenya, Jamaica, and South Korea have enacted some form of registration law. The European Union allows for a comprehensive approach to sanctions for certain sex offenses and has attempted to enforce criminal record sharing and reciprocal bans for convicted sex offenders from working with children (Smith 2011). However, none of these E.U. directives have created anything like the registry or notification procedures that one sees in the U.S.

As of 2014, the U.S. allows greater access to registration information than any other country. In a few countries (e.g., the U.K.), laws do permit the disclosure of information about an offender's status and criminal history to people who have a close connection to a child who may have unsupervised contact with someone suspected of being a sex offender. However, this information is considered confidential and cannot be shared with others. In 2001, South Korea began publicly disclosing information about high-risk offenders, but in 2003, this practice was deemed unconstitutional. Today, however, South Korea does rely on GPS tracking of convicted sex offenders. In Manitoba, Canada, the public can access limited information on registered offenders who pose a great risk of reoffending. Western Australia began disclosing information about convicted sex offenders to the general public in 2012 (Whitting et al. 2014). In 2014, a bill in Canada that proposed tougher penalties for sexual predators included a provision allowing for the public disclosure of names and addresses of some registered

sex offenders, in order "to encourage protecting children from past proven behaviours" (Connolly 2014), but at the time of this writing, this proposal has not been made law.

Sex Offender Laws and the Economic Model of Crime

The economic model of crime conceptualizes a rational potential offender who chooses to engage in crime when the expected benefits of committing the crime outweigh its expected costs. Expected costs in most contexts are a function of the probability of punishment, the punishment if convicted (e.g., prison, fines, or social sanctions, including shaming punishments and isolation), and the direct costs of committing the crime (e.g., travel, or procuring a gun). Expected benefits can either be tangible (e.g., money or goods) or intangible (e.g., pleasure or other psychological benefits). A more severe punishment or a higher likelihood of detection should reduce the proportion of potential offenders who choose to commit a crime in this model. Thus, while other purposes of punishment (such as incapacitation, rehabilitation, or retribution) may reduce crime or improve social welfare, the economic model of crime focuses primarily on the potential deterrent effect of criminal policies and regulations.

Are Sex Offenders Economically Rational?

The relevance of the economic model of crime to the commission of sex offenses and to the understanding of sex offender behavior is not at all obvious. At least in part, this observation stems from the fact that, to many, sex offenses appear simply inexplicable – i.e., people have trouble comprehending why anyone would benefit *at all* from committing most such crimes (e.g., Stinson et al. 2008, p. 3). The motivations behind acquisitive crimes and even certain crimes of violence, by contrast, are easier for many to understand. Sex offenses are also often viewed as impulsive (although many crimes in this category are clearly not impulsive and require extensive planning – the production of child pornography or prostitution, for instance). By these lights, certain sex offenders

appear to be unable to make conscious decisions about whether they ought to commit these sorts of crimes. A consequence of this perspective is that sex offending is regularly associated with mental illness or some other immutable or biological trait (White 1999, pp. 357, 392–394).

If one begins with the assumption that sex offenders cannot appreciate the nature of their conduct or are unable to control their behavior *at all*, the economic model of crime with its focus on deterrence plays no useful role in understanding or preventing criminal behavior. Instead, the policy goals of incapacitation and rehabilitation must take center stage. In this vein, researchers have proposed many theories (not mutually exclusive) to account for sexually deviant behavior. Scholars developed these noneconomic models in part to reduce the extent of offending by accurately identifying and treating those who are at risk of committing such offenses (as opposed to deterring them). Stinson et al. (2008) summarize and integrate this body of literature, describing biological theories, cognitive theories, behavioral theories, social learning theories, personality theories, and evolutionary theories.

But in all cases, these models do in fact assume that sex offenders retain at least some ability to respond to their environment, and accordingly, the economic model of crime can be profitably applied to understanding sexually criminal behavior. Moreover, alternative causal theories are complementary to the economic model because the former aid in understanding the psychological benefits offenders pursue when offending and how they perceive (and misperceive) their environment, particularly whether the behavior the potential offender is considering is truly transgressive and, if so, whether it is likely to result in apprehension and punishment.

The Behavioral Consequences of Sex Offender Laws

For at least two reasons, the theoretical and empirical research on sex offenders as economic actors has focused to a considerable degree on the behavioral consequences of postrelease regulations. First, simply examining the effects of more conventional criminal sanctions on sex offender

behavior is at best a minor contribution, at least when viewed from within the four corners of the economic model of crime. As with criminal behavior in general, more severe sentences or more policing resources may induce changes in sex offending behavior, whether accomplished in a straightforward way or through the implementation of policies likely to have the same effect – e.g., mandatory minimum sentences, three-strikes law, and so on. In essence, the basic mechanics of the economic model of crime are fundamentally the same, whether or not the crime in question is a sex offense. Second, policies such as registration, community notification, and residency restrictions are relatively novel in modern times; these sanctions could in theory be applied to other types of released offenders, but with a few notable exceptions in a few jurisdictions (e.g., elder abuse), sex offenders remain their sole targets. Sex offender behavior thus provides a unique opportunity to study empirically the effects of these postrelease laws and their complex interactions with traditional sanctions (like incarceration), all of which may push sex offense rates in different and potentially offsetting ways (see, e.g., Teichman 2005).

Consider the framework offered by Prescott and Rockoff (2011) involving a potential sex offender i who must decide whether to commit a sex offense against victim j . This decision is a function of a victim-specific probability of punishment (p_{ij}), targeting costs (c_{ij}), the level of punishment if convicted (f_i), and victim-invariant utility (benefit) of committing the sex offense relative to lawful alternatives (u_i). In this economic environment, postrelease sex offender laws – unlike an across-the-board increase in the sentencing range for sex offenses or devoting more resources to the detection of sex crime – may have the effect of altering offending behavior in multiple distinct ways.

For instance, registration may enhance police monitoring and facilitate the apprehension of registered offenders who commit crimes against local victims, thereby increasing certain victim-specific punishment probabilities (p_{ij}). Registration may also influence the behavior of nonregistered individuals because punishment (f_i) now includes a

higher future probability of detection upon release. Notification, by contrast, effectively shames sex offenders by publicly airing their personal information and criminal history, which can result in loss of employment and social ties, poor housing, and even harassment (Levenson and Cotter 2005a). For nonregistered sex offenders, notification makes the consequences of any conviction more severe (f_i). Already registered offenders, however, would face no *increase* in punishment (f_i). The relative utility of crime (u_i) would be higher under notification for already registered offenders because engaging solely in law-abiding behavior would involve more stress, loneliness, and unemployment, but the ability of local residents and acquaintances to identify and avoid these offenders would make targeting this particular subset of potential victims (c_{ij}) more difficult. Thus, the predicted consequences of notification for crime rates are ambiguous.

Sex Offender Laws and the Geography of Victimization

Postrelease laws also have the potential to alter the geography of sex offender behavior in ways both expected and unexpected. For instance, as Prentky (1996) notes, notification may “accomplish nothing more than changing the neighborhood in which the offender looks for victims” (p. 295). Agan and Prescott (2014) develop this idea in more depth; they alter Prescott and Rockoff’s (2011) framework by considering a potential offender i who is deciding whether to commit a crime against victim j in location l . The offender now additionally weighs the cost of travel to location l (c_{il}), the cost of finding a victim j in l (c_{jl}), and a location specific probability of punishment (p_{il}). Although registration and notification laws apply only to registered (convicted) sex offenders, the information these laws provide and the responses of the police and at-risk individuals may have effects on the offense locations of *all* potential offenders, both those who are registered as well as potential first-time offenders.

Using this model, one can explore the potential consequences of more registered sex offenders living in a particular neighborhood on that

neighborhood’s crime rates. When only registration and not notification laws are in effect, enhanced monitoring of registered sex offenders presumably increases the probability of punishment for sex offenses that such offenders commit near their homes (p_{il}), possibly displacing or deterring those crimes. Registration alone, however, is in theory confidential and thus seems unlikely to affect victim behavior (c_{ijl}) or the location parameters for nonregistered potential offenders. When the contents of the sex offender registry are made publicly available via notification, however, potential victims, registered sex offenders, and even nonregistered sex offenders may all change their behavior, implying ambiguous effects on crime level predictions near registered sex offender residences. Publicizing their identities may render registered offenders even more likely to offend in other neighborhoods as they attempt to reduce the likelihood of their being caught (p_{il}). Alternatively, notification might make offending more attractive to registered sex offenders (e.g., by aggravating recidivism risk factors like unemployment). Thus, if victim precaution-taking is ineffective and/or travel is difficult, registered sex offenders may find attacking near home significantly more attractive.

One key insight from this work is that post-release laws that alter the information environment generally, like notification, may affect the behavior of potential sex offenders who have never been convicted of a sex crime. These non-registered offenders may be able to use the information provided by public notification in deciding both *whether* and *where* to offend, and the changes in their decisions may have the consequence of altering the geography of sex offense victimization risk in unexpected ways. For example, if the police tend to focus their energies on monitoring registered offenders, *nonregistered offenders* may be less likely to be caught in neighborhoods with high concentrations of publicly known offenders (p_{il}). Similarly, victims may avoid registered offenders and their homes, but lower their guard around individuals who do not appear on the registry (c_{ijl}). If these factors attract nonregistered offenders to neighborhoods where registered offenders reside, these neighborhoods

may become more rather than less dangerous after notification. Whether the effect of notification laws on sex offense rates in neighborhoods with resident registered offenders is positive or negative ultimately depends on the net effect of these factors.

Empirical Evidence on Sex Offenses and Crime-Reduction Policies

The frequency and distribution of sex offenses are not only functions of traditional criminal justice policies (such as policing levels and sentence lengths), but also of the unique post release regulations to which many sex offenders are made subject in the U.S. and elsewhere. This part of the entry will begin by briefly and selectively reviewing a few recent findings with respect to the effects of traditional policy innovations in which sex offenders (or sex offense rates) are singled out by the researchers as behaving distinctively. The section will focus more heavily, however, on a growing body of work that studies the empirical consequences of postrelease sex offender laws.

The Effects of Traditional Policy Innovations

The below economic model of criminal behavior laid out previously implies that there exist many policy levers that may influence crime rates, including sex offense rates, such as the probability that an offender is arrested conditional on committing a crime and the punishment (often the extent of incarceration) if the offender is convicted of the crime. Few research papers in this arena focus *solely* on the empirical effect of general criminal justice policies on sex offense rates. However, because the major data sources, and in particular the UCR, include sex offenses as one explicit category of crime, research that reports results disaggregated by offense type will often include and sometimes discuss separately results for rape and occasionally for other categories of sex offenses.

A conventional model of criminal offending predicts that increasing the probability of punishment (e.g., through heightened detection levels or

more resources devoted to apprehension) will decrease the likelihood that an individual will choose to commit a crime. Levitt (1998) provides classic evidence for this conclusion – increases in arrest rates are associated with decreases in commission rates for all crimes, including rape specifically. However, Levitt concludes that for rape, as opposed to other crimes, the main channel through which higher arrest rates reduce rape rates is through incapacitation – that is, incarcerating sex offenders reduces the relative frequency of rape *not* because potential rapists are deterred, but instead because sex offenders, once incapacitated, commit fewer crimes. With respect to the magnitude of the effect, Levitt finds that one additional arrest for rape is associated with a reduction in the number of rapes by one-half as a result of sex offender incapacitation.

An additional method of increasing the probability of punishment, presumably through either higher arrest rate probabilities or heightened monitoring, which in turn leads to greater deterrence, is to supplement the number of police officers in a particular locale. While boosting law enforcement personnel levels does appear generally to reduce crime incidence reports (i.e., aggregate crime levels), both Levitt (1997) and Evans and Owens (2007) – using different empirical strategies – estimate no significant effect of additional police officers on rates of reported rape incidents.

The traditional punishment meted out for serious crimes is incarceration, and over the last three decades, there has been a significant increase in the number of Americans in prison. Larger prison populations have the potential for both an incapacitation effect on those in prison as well as a deterrent effect on potential criminals who infer a greater likelihood of being incarcerated if they commit crime in the future. Johnson and Raphael (2012) explore the marginal effect of an additional prisoner on total crime rates, evaluating the combination of these two effects. They estimate that an additional prisoner reduces rapes by about 0.03, a statistically significant effect, although much smaller in magnitude than the overall elasticities they estimate for property crimes (2.3) or all violent crimes (0.12).

The Effects and Implications of Postrelease Sex Offender Laws

The application of postrelease laws (including registration, notification, and others) to those who have been convicted of committing a sex offense also has obvious potential to affect the frequency of sex offenses, although not only by reducing their numbers, as is usually assumed by policymakers. These regulations are implemented ostensibly to increase public safety and prevent known offenders from reoffending (particularly, close to where these offenders live). These laws may also influence the behavior of potential victims, the police, and even individuals considering committing a sex offense for the first time, who may fear the possibility of being subjected to such regulations if apprehended and convicted of a covered offense. Perhaps surprisingly, it is theoretically uncertain whether these laws ought to reduce sex offense rates; whether there may be collateral consequences to the passage and implementation of these laws is also an important policy consideration. Several major social science research studies have explored these empirical research questions.

Registration, Notification, and Offender Behavior

Given that these laws are primarily intended to increase public safety by reducing sex offender recidivism, a clear place to start the discussion is with recent work seeking to understand whether either registration or notification (the earliest categories of postrelease sex offender laws) are effective in this regard. In the U.S., two papers that have examined national-level data on the effects of sex offender registration and notification on recidivism. Agan (2011) uses state-by-state panel data from the FBI's Uniform Crime Reports to understand the impact of registration and notification laws on rates of reported rape incidents and arrests for other sex offenses as these provisions went into effect at different times across different states. Her analysis finds no effect of registration or notification on rape rates, and no effect of registration on sex offense arrests. However, she does find some evidence that sex offense arrests decreased after the implementation of internet notification. Using individual data on all sex offenders released from prison in 1994 and followed for 3 years, Agan

also shows that convicted sex offenders released in states that had registration laws were no less likely (and if anything, more likely) to recidivate than sex offenders released in states that did not have registration laws.

Prescott and Rockoff (2011) study similar questions, but use the previously described NIBRS data, which allows for a more nuanced look at sex offense definitions and victim-offender relationships. One of their contributions is to use the number of registered offenders at the county level to separately identify the deterrent and recidivism-reduction effects of registration and notification policies. They find no discernible deterrent effect for registration laws alone but do find evidence that these laws reduced recidivism for crimes committed by offenders known to the victim (but not for those committed against strangers). Notification, on the other hand, does appear in their data to deter potential offenders (i.e., nonregistered individuals). By contrast, Prescott and Rockoff find no evidence that public notification reduced recidivism by registrants, supposedly the primary purpose of these laws. In fact, in the states they study, notification may actually have increased recidivism rates.

While there are some glimmers of hope in these results, the overall takeaway from these studies is that no reliable evidence exists that registration or notification laws increase public safety by reducing recidivism. Several researchers have performed state-specific analyses of the effects of these policies on sex offense rates and recidivism and have reached a similar conclusion (see, e.g., Tewksbury, Jennings, and Zgoba (2012) in New Jersey; Tewksbury and Jennings (2010) in Iowa; Sandler et al. (2008) in New York, among others), although the results are in no sense uniform and many open research questions remain.

Notification and Potential Victim Behavior

Notification laws were designed to reduce offender recidivism directly by influencing the behavior of potential offenders, but also indirectly, by changing the habits and activities of potential victims – i.e., community members who receive sex offender registry information. Indeed, notification policies are premised on the

idea that newly informed neighbors and other individuals at risk will engage in effective precaution-taking in response to registry information, which would in turn make it more difficult for potential recidivists both to locate plausible victims and, if one is located, to succeed at committing a crime and evading arrest. Empirical support for this premise, however, is mixed. Anderson and Sample (2008) surveyed a representative sample of Nebraskans and found that only roughly one-third of respondents reported taking any preventative actions upon learning relevant sex offender registry information. Moreover, the most common responsive action reported by potential victims in their study was simply further communicating registry information to others.

In a small survey in Ohio, Beck and Travis (2004) evaluated the precaution-taking responsiveness of individuals who were “actively” notified by law enforcement of the identities and whereabouts of nearby registered sex offenders. They found that individuals receiving notification engaged in relatively more protective behaviors. Bandy (2011), however, conducted a similar survey of households affected by Minnesota’s active community notification program (which applies to those living within approximately three blocks of a registered sex offender), but additionally surveyed a matched sample of households that were not affected by the program. Once neighborhood demographics and other confounding characteristics are taken into account, Bandy finds no significant relationship between receiving notification information and engaging in protective behaviors. She posits that because sex offenders often live in disadvantaged neighborhoods, and because individuals in such neighborhoods already engage in extensive precaution-taking behavior in any event, notification does not change the way they perceive the safety of their surroundings.

Residency Restrictions and Offender Behavior

Residency restrictions limit where registered sex offenders can live, usually by prohibiting residency within a certain distance from areas where children might congregate (such as schools, day-care centers, and sometimes even bus stops). These restrictions are real constraints; they make

it demonstrably more difficult for registered sex offenders to find stable housing. In fact, Zandbergen and Hart (2006) show that the particularly restrictive residency restrictions in Florida implied that only five percent of housing was legally available to registered sex offenders in the state. These restrictions often force registered sex offenders to live in socially disorganized, impoverished neighborhoods (Tewksbury and Mustaine 2008). These neighborhoods are likely to be less conducive to reintegration because they offer fewer opportunities for employment and social connection.

There is little research on the effects of residency restrictions on sex offender recidivism or sex offense rates more generally. The evidence that does exist, however, does not support the idea that residency restrictions reduce recidivism. Using data from North Carolina, Kang (2012) finds that offenders subject to residency restrictions are more likely to be reconvicted for a non-sex crime (a result consistent with the idea that residency restrictions make finding employment and stable housing more difficult), but also finds no evidence that they are more (or less) likely to commit another sex offense. Kang echoes the hypothesis that the adverse effects of restrictions on offending behavior generally may have their source in the social disorganization of the neighborhoods in which registered sex offenders live. One study by Huebner et al. (2014) proceeds by comparing recidivism rates of registered sex offenders who were released from prison in Michigan and Missouri before and after residency restrictions had gone into effect in those states. Using a propensity score matching design, the authors find no evidence that residency restrictions decreased recidivism levels, and in Michigan, registered sex offenders subject to such restrictions appeared to be slightly more likely to recidivate. Unfortunately, low overall recidivism rates make it difficult for the authors to be confident in this latter finding. In Minnesota, Duwe et al. (2008) study the geography of reoffending patterns of registered sex offenders *before* Minnesota’s residency restrictions were implemented, asking hypothetically whether the restrictions ultimately adopted could possibly

have made a difference in the committed crimes. They conclude that for the 224 sex offenders they evaluate, there is little evidence that any of their subsequent sex offenses would have been prevented by Minnesota's residency restrictions.

Evidence on the Geography of Victimization

Registration and notification laws as well as residency restrictions are based on the idea that sex offenders are likely to commit crimes near their home, and so restricting them from getting too close to particular victims or allowing at-risk individuals to steer clear of potential recidivists by publicly identifying where convicted sex offenders live should increase public safety. Whether sex offenders *actually* offend near their homes is an empirical question, however. Offsetting influences on the locational offending choices of registered sex offenders as well as potential sex offenders render theoretical conclusions ambiguous.

Tewksbury et al. (2008) compare census tracts in Jefferson County, Kentucky (home of Louisville), and find no evidence that a higher concentration of registered sex offenders in a census tract is associated with more sex offenses. Agan and Prescott (2014) study the effects of registration and notification laws on the geography of sex offense victimization in Baltimore County, Maryland. They find that when registered sex offenders are only subject to registration, there are *fewer* reported sex offenses in neighborhoods where registered sex offenders live. When notification about registered offenders becomes required, however, reported sex offense rates *increase* in those neighborhoods where registered sex offenders live, although overall there are still fewer sex offenses committed in these neighborhoods. Importantly, there is variation in these effects for different types of sex offenses. In particular, the findings are reversed for the crimes of forcible rape and sex offenses against children (as opposed to sex offenses against adults or peeping-tom, pornography, and prostitution crimes). For these crimes, an additional registered sex offender in a neighborhood is associated with an increase in the frequency of sex offense reports when only registration is in place but is associated with less risk once notification begins.

Collateral Consequences of Sex Offender Laws

Research has failed to demonstrate empirically the efficacy of postrelease regulations aimed at reducing sex offense rates and recidivism by convicted sex offenders. By contrast, extensive research indicates that these laws have serious collateral consequences for many community members, not to mention for released sex offenders and their families.

For instance, postrelease sex offender laws can have far-reaching consequences for the housing market. Perhaps because people fear living near convicted sex offenders, prospective buyers appear to use information on the residential proximity of registered sex offenders when deciding whether to purchase a home and what price to pay for it. In particular, evidence suggests that a registered sex offender living very nearby greatly reduces a residential property's value. Both Linden and Rockoff (2008) and Pope (2008) show that prices for homes located near a registered sex offender are significantly lower than otherwise comparable homes located away from a known sex offender's residence: Pope's results, for example, indicate that houses located within one-tenth of a mile of a convicted sex offender's home sell for roughly 2.3% less than otherwise similar homes with no registered sex offenders living nearby.

Survey evidence also shows that notification can hamper social integration and result in greater fear of crime in a community (Zevitz 2004; Beck and Travis 2004). This is true despite the fact researchers have yet to find compelling evidence that potential victims take protective actions in response to notification information or that these policies increase public safety.

Research has also shown that postrelease sex offense laws have significant collateral consequences for convicted sex offenders and their families. Surveys have found that registered sex offenders experience significant stress, isolation, fear, shame, and embarrassment as a result of being publicly identified as a convicted sex offender. Many have lost their jobs or homes or struggle with distress or depression and some also experience threats or property damage (Levenson et al. 2007). These consequences may be risk factors that lead registered sex offenders to

become more rather than less likely to recidivate (Levenson and Cotter 2005a). A similar survey studying the collateral effects of residency restrictions suggests that these restrictions also lead to decreased stability; some of the respondents in the survey even warn that this reduced stability increases recidivism risk (Levenson and Cotter 2005b).

Concluding Thoughts

A sex offense is defined as the occurrence of a sexual act, a sexual touching, or a display of body parts considered to be sexual in nature, coupled with actual or legally imposed lack of consent on the part of the victim or society. Notwithstanding the fact that reported sex offense rates have been declining over the past two decades, and despite the varied nature of crimes that are legally considered “sex offenses,” postrelease laws that broadly and uniformly apply to offenders who have committed these crimes are proliferating, from registration to notification to residency restrictions. By and large, rigorous empirical evidence offers no support for the claim that postrelease regulations have the potential to reduce sex offense rates or recidivism, despite significant research effort and decades of experience with these laws. Indeed, these regulations may increase recidivism risk for those previously convicted of sex offenses and at the same time create a culture of fear and dilute social cohesion in communities. Open questions beckon – among them the consequences of civil commitment and GPS tracking, in addition to the potential offered by innovative, local experimentation. Yet, the conclusion remains that postrelease sex offense regulations appear thusfar to be causing fear without function.

References

- Agan AY (2011) Sex offender registries: fear without function? *J Law Econ* 54(1):207–239
- Agan AY, Prescott JJ (2014) Sex offender law and the geography of victimization. *J Empir Leg Stud* 11(4):786–828
- Anderson AL, Sample LL (2008) Public awareness and action resulting from sex offender community notification laws. *Criminal Justice Policy Rev* 19(4):371–396
- Associated Press (2005) Sex offenders kept from Storm Shelters. *New York Times*, August 8. Available <http://www.nytimes.com/2005/08/08/national/08florida.html>. Accessed 23 Mar 2015
- Bandy R (2011) Measuring the impact of sex offender notification on community adoption of protective behaviors. *Criminol Public Policy* 10(2):237–263
- Beck VS, Travis LF (2004) Sex offender notification and fear of victimization. *J Criminal Justice* 32(5):455–463
- Connolly A (2014) Public sex offender registry coming soon, says Peter MacKay. *CBC News*, February 28. Available <http://www.cbc.ca/news/canada/calgary/public-sex-offender-registry-coming-soon-says-peter-mackay-1.2556080>. Accessed 23 Mar 2015
- Duwe G, Donnay W, Tewksbury R (2008) Does residential proximity matter? A geographic analysis of sex offense recidivism. *Criminal Justice Behav* 35(4):484–504
- Evans WN, Owens EG (2007) COPS and crime. *J Public Econ* 91(1):181–201
- Huebner BM, Kras KR, Rydberg J, Bynum TS, Grommon E, Pleggenkuhle B (2014) The effect and implications of sex offender residence restrictions. *Criminol Public Policy* 13(1):139–168
- Johnson R, Raphael S (2012) How much crime reduction does the marginal prisoner buy? *J Law Econ* 55(2):275–310
- Kang S (2012) The consequences of sex offender residency restriction: evidence from North Carolina. Unpublished manuscript, Duke University Department of Economics, Durham
- Langan PA, Levin DJ (2002) Recidivism of prisoners released in 1994. *Fed Sentencing Report* 15(1):58–65
- Langan PA, Schmitt EL, Durose MR (2003) Recidivism of sex offenders released from prison in 1994. US Department of Justice, Bureau of Justice Statistics, Washington, DC
- Levenson JS, Cotter LP (2005a) The effect of Megan’s Law on sex offender reintegration. *J Contemp Criminal Justice* 21(1):49–66
- Levenson JS, Cotter LP (2005b) The impact of sex offender residence restrictions: 1,000 feet from danger or one step from absurd? *Int J Offender Ther Comp Criminol* 49(2):168–178
- Levenson JS, D’Amora DA, Hern AL (2007) Megan’s law and its impact on community re-entry for sex offenders. *Behav Sci Law* 25(4):587–602
- Levitt SD (1997) Using electoral cycles in police hiring to estimate the effect of police on crime. *Am Econ Rev* 87(3):270–290
- Levitt SD (1998) Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error? *Econ Inq* 36(3):353–372
- Linden L, Rockoff JE (2008) Estimates of the impact of crime risk on property values from Megan’s laws. *Am Econ Rev* 98(3):1103–1127
- Pope JC (2008) Fear of crime and housing prices: household reactions to sex offender registries. *J Urban Econ* 64(3):601–614
- Prentky RA (1996) Community notification and constructive risk reduction. *J Interpers Violence* 11(2):295–298

- Prescott JJ, Rockoff JE (2011) Do sex offender registration and notification laws affect criminal behavior? *J Law Econ* 54(1):161–206
- Sandler JC, Freeman NJ, Socia KM (2008) Does a watched pot boil? A time-series analysis of New York State's sex offender registration and notification law. *Psychol Public Policy Law* 14(4):284
- Smith NJ (2011) Protecting the children of the world: a proposal for tracking convicted sex offenders internationally. *San Diego Int Law J* 13:623
- Soothill K (2010) Sex offender recidivism. *Crime Justice* 39(1):145–211
- Stinson JD, Becker JV, Sales BD (2008) Self-regulation and the etiology of sexual deviance: evaluating causal theory. *Violence Vict* 23(1):35–51
- Teichman D (2005) The market for criminal justice: federalism, crime control, and jurisdictional competition. *Mich Law Rev* 103:1831–1876
- Tewksbury R, Jennings WG (2010) Assessing the impact of sex offender registration and community notification on sex-offending trajectories. *Criminal Justice Behav* 37(5):570–582
- Tewksbury R, Mustaine EE (2008) Where registered sex offenders live: community characteristics and proximity to possible victims. *Vict Offenders* 3(1): 86–98
- Tewksbury R, Mustaine EE, Stengel KM (2008) Examining rates of sexual offenses from a routine activities perspective. *Vict Offenders* 3(1):75–85
- Tewksbury R, Jennings WG, Zgoba KM (2012) A longitudinal examination of sex offender recidivism prior to and following the implementation of SORN. *Behav Sci Law* 30(3):308–328
- Thomas T (2011) *The registration and monitoring of sex offenders: a comparative study*. Routledge, New York
- Truman JL, Langton L, Planty M (2013) *Criminal victimization, 2012*. US Department of Justice, Bureau of Justice Statistics Bulletin. Bureau of Justice Statistics, Washington, DC
- Tucker G (1803) *Blackstone's commentaries*. William Young Birch & Abraham Small, Philadelphia
- United States Department of Justice, Federal Bureau of Investigation (2014) *Crime in the United States, 2013*. Available <http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2013/crime-in-the-u.s.-2013>. Accessed 23 Nov 2014
- White A (1999) Victims' rights, rule of law, and the threat to liberal jurisprudence. *Ky Law J* 87(2):357–415
- Whitting L, Day A, Powell M (2014) The impact of community notification on the management of sex offenders in the community: an Australian perspective. *Aust N Z J Criminol* 47(2):240–258
- Zandbergen PA, Hart TC (2006) Reducing housing options for convicted sex offenders: investigating the impact of residency restriction laws using GIS. *Justice Res Policy* 8(2):1–24
- Zevitz RG (2004) Sex offender placement and neighborhood social integration: the making of a scarlet letter community. *Criminal Justice Stud* 17(2): 203–222

Sexual Assault

► Sex Offenses

Sexual Offenses

► Sex Offenses

Shadow Banking

Alessio M. Paccès

Erasmus School of Law, Rotterdam Institute of Law and Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

Definition

Shadow banking is often defined by reference to what is not, namely official banking. This essay takes a different approach. Focusing on the purpose of shadow banking regulation, namely minimizing systemic risk, shadow banking is defined as leveraging on collateral to support liquidity promises.

This essay discusses the economic case for regulating shadow banking by asking three questions. First, what is shadow banking? Second, why regulate shadow banking? Third, how to regulate shadow banking efficiently? It is argued that shadow banking should be regulated because the social cost of systemic risk is not internalized by the players generating it. Moreover, because systemic risk cannot be accurately measured and priced, this essay argues that the negative externalities of shadow banking should be limited through quantity regulation – restrictions on the size of shadow banking – rather than corrective taxation.

Introduction

This essay deals with the economic rationale for regulating shadow banking. It discusses whether

the regulatory initiatives proposed by academics and policymakers are consistent with this rationale. Positing that the ultimate goal of financial regulation is to promote financial stability, shadow banking regulation is evaluated based on its ability to reduce financial instability efficiently.

Regulating shadow banking is challenging because shadow banking is often defined by reference to what it is not, namely licensed or official banking. However, such an approach does not capture the essence of the shadow banking problem. The official banking system has always implicitly or explicitly supported a significant part of what is known today as shadow banking, by way of so-called liquidity puts (Claessens and Ratnovski 2014). A liquidity put is a put option which backstops the liquidity needs of a financial institution. A liquidity put may be private, for instance from a commercial bank, or public, for instance from a central bank.

One should rather start from the observation that shadow banking is effectively banking, albeit carried out in such a way as to avoid regulatory constraints. In order for regulation to be effective in promoting financial stability, shadow banking is to be defined functionally, based on its contribution to systemic risk. The definition of systemic risk can be instrumental to measuring it (Brunnermeier and Krishnamurthy 2014). Measuring systemic risk is important. However, such measurement can never become so precise as to prevent financial crises. To define shadow banking, it is better to adopt a conceptual approach to systemic risk. In this perspective, systemic risk can be characterized as the likelihood of financial system failure, which will impair the financing of production and consumption, and thus will have consequences on the performance of the real economy.

Anchoring the definition of shadow banking to a conceptual framework of systemic risk, instead of to a list of systemically relevant activities, allows identifying shadow banking despite financial innovation and the imperfection of risk models employed by the financial industry and its regulators. One can define shadow banking by its ability to fund long-term commitments through short-term liabilities, which are considered *as*

safe as money (Mehrling et al. 2013). The implied liquidity of such liabilities makes them subject to run. A generalized run on banking institutions – both official and shadow – is a systemic risk event because it destabilizes the financial system, which becomes unable to support production and consumption in the real economy.

To preserve financial stability, regulation of shadow banking should focus on entities and transactions allowing liabilities to be accepted as a substitute for money. For the banking business to be profitable, such safe, liquid, and short-term liabilities must be invested in risky, illiquid, and long-term assets. Both official and shadow banking are to be identified in connection with maturity transformation. This is the ultimate source of systemic risk. Liquidity promises can no longer be honored when there is uncertainty on how to transform risky, long-term assets into safe, short-term ones. This situation creates panic and runs. To be sure, the maturity mismatch between assets and liabilities is not essential for shadow banking to generate systemic risk. The latter stems likewise from liquidity transformation (financing illiquid assets with liquid liabilities) and credit transformation (enhancing the credit risk of assets in order to fund them) (Adrian and Ashcraft 2016). For the purpose of the present essay, maturity transformation is to be intended broadly, as to include liquidity and credit transformation too.

The Definition of Shadow Banking in Policymaking

Defining shadow banking is challenging because it involves a variety of financial instruments, institutions, and activities, and it is changing over time. Therefore, searching for an all-encompassing definition would not take us far. Policymakers should rather define shadow banking by reference to the purpose for which it is studied. Since the shadow banking is worrisome because of its contribution to systemic risk, a useful definition needs to cover the systemic risk implications of shadow banking.

Currently, there are three approaches to defining shadow banking. Shadow banking may be

defined by the activities, the entities, or the instruments constituting shadow banking. Activity-based definitions include maturity, liquidity and credit transformation so long as they are geared towards performing credit intermediation – namely, taking savings from lenders and channeling them towards borrowers. Activity-based definitions may or may not exclude credit intermediation performed by banks, for instance because the latter engage in regulatory arbitrage. Alternatively, shadow banking may be defined by the entities carrying out credit intermediation subject to the risk of runs. Entity-based definitions typically exclude banks. One example of such entities, among many, are Money Market Mutual Funds (MMMFs). Finally, shadow banking may be defined by the instruments through which it is carried out. Take, for instance, the repurchase agreement (repo), which is functionally a contract to borrow on financial collateral. Depending on the quality of collateral, its credit enhancement and the term of the contract, a repo can generate money-like liabilities. There are several other examples of such instruments. Importantly, the list is not finite, but it includes every credit instrument that can generate safe and liquid liabilities, which are potentially subject to runs.

An instrument-based definition is a functional definition anchored to a financial instrument that may or may not yet exist. Activity-based and entity-based definitions are functional, too, because they reflect the systemic risk concerns stemming from shadow banking. However, these definitions are static due to their reliance on a list of entities and/or activities. Activity-based and entity-based definitions are useful to identify the object of regulation, whereas an instrument-based approach is more promising to adapt regulation to unknown challenges to financial stability.

Activity-based definitions are very attractive in theory. “If it looks like a duck, quacks like a duck, and acts like a duck, then it is a duck” – so the saying goes. When it comes to shadow banking, however, such definitions are very difficult to operationalize because they are either too narrow or too broad. Too narrow definitions are easy to circumvent. Too broad definitions implicitly confer discretion upon the agency that is supposed to

identify and regulate shadow banking. This may generate legitimacy concerns because identifying the financial institutions subject to regulation is a sensitive policy issue, which is usually for the legislature to decide.

The Financial Stability Board (FSB 2011) has adopted an activity-based definition, which is operationalized by reference to a broad set of nonbank entities (“credit intermediation involving entities and activities outside the regular banking system”). The International Monetary Fund (IMF 2014) has taken a completely different approach, identifying shadow banking based on issuing liabilities different from those of traditional banking (“noncore liabilities”). The IMF approach is therefore instrument-based. Both approaches support a narrow and a broad version for precautionary reasons. However, they are not comparable because the IMF definition includes banks as well. According to the FSB (2015), the narrow measure of shadow banking in 26 jurisdictions was \$36 trillion in 2014, as opposed to \$135 trillion of the official banking sector.

A Collateral-Based Definition

Choosing one approach to define shadow banking is not easy, for they all have shortcomings. Activity-based definitions are underinclusive or too vague to operationalize in a democratic regulatory system. Entity-based definitions may be sufficiently precise to operationalize, but they are ultimately uninformative about the source of systemic risk. Moreover, both activity-based and entity-based definitions are static and may fail to capture new forms of shadow banking that use financial innovation to generate systemic risk. Instruments-based definitions may fare better in connection with systemic risk, and if carefully crafted, can adapt to financial innovation. However, they tend to be overinclusive, which brings us back to the operationalization problem. The best way to define shadow banking is, therefore, a combination of the three approaches, with a special focus on the instruments.

Systemic risk stems from the promise of liquidity against long-term, risky investments. There are

not many instruments making such promises credible. These instruments can be characterized as collateral, namely assets that may be liquidated to make good on the promise if the borrower defaults. Liquidity promises are credible if the collateral backing them is liquid, too, implying that the collateral can be turned into cash immediately without losing much of its value.

Banking adds another dimension to the liquidity of collateral, which is leverage. Because banks profit from transforming debts (long into short, safe into risky), they tend to have as much debt, or leverage, as possible. Banks can leverage more than any other economic actor because they rely on various liquidity puts – from other banks, and ultimately, from the central bank. However, leverage is a multiplier of losses as well as of gains, as a result of which banks are more fragile than other economic actors. In the absence of regulation, the market liquidity of assets increases the leverage possibilities of banks (funding liquidity), until the asset price bubble bursts and both market and funding liquidity turns in the other direction. Liquidity and leverage are, therefore, the defining features of banking, both official and shadow (Brunnermeier and Pedersen 2009; Adrian and Shin 2010).

Nabilou and Paccès (2017) define banking as the business of leveraging on collateral to support liquidity promises. When this is not subject to the regulation of liquidity and leverage imposed on banks for the purpose of financial stability, it is shadow banking. For example, demand deposits are banking, but not shadow banking, because they are subject to the liquidity and leverage regulation of banks. Similarly, MMMFs are shadow banking so long as they are not subject to regulatory constraints comparable to those faced by banks. Repos are shadow banking, unless it is carried out by banks whose operation is curtailed by liquidity and leverage regulations. Derivatives were, and still are, a form of shadow banking because they support leveraged liquidity promises, although these are not obvious. A marked-to-market derivative, which is used extensively in shadow banking operations, is effectively an overnight promise to pay the margin call. Because the liquidity and the leverage of derivatives are not

regulated as such, but only indirectly when derivatives are entered into by banks or other systemically important institutions, derivatives markets are still the realm of shadow banking.

This instrument-based approach has several advantages. First, the definition includes all – past, present and future – activities giving rise to systemic risk, because it does not rely on particular markets. Rather, this definition relies on the core instruments of banking at all times: liquidity and leverage. Second, this definition is functional, but not as vague as a functional activity-based definition. Because financial regulation can embed ex-ante the instruments of liquidity and leverage, it can also confer upon a regulatory agency the mandate to regulate them across the board, without creating legitimacy concerns. Third, the definition is not linked to any particular entity. Unfortunately, this is also a limitation of this approach. Recently, a number of techniques to monitor liquidity and leverage have been identified in economics (Brunnermeier and Krishnamurthy 2014). All these overlook that effective reporting obligations can only be legally established on specific entities. You cannot regulate something you cannot measure (Pozsar 2014). Similarly, you cannot measure something you do not know of. For this reason, policymakers should rely on entity-based approaches, too, albeit guided by instrument-based criteria to identify the right entities and adapt this identification with time (Nabilou and Paccès 2017).

The Economic Rationale of Shadow Banking

To cope with systemic risk, it might be tempting to ban shadow banking altogether. This would not be a good idea. First, a prohibition will probably not be effective. Because shadow banking caters to a structural demand for safe assets, a ban would not likely be sufficient to deter the private creation of money-like liabilities; it would only marginally increase the cost of accepting them. Second, even if one could prohibit shadow banking, this would not be efficient. Arguably, a limited amount of shadow banking is welfare

increasing, which turns the policy question into how regulation can induce the optimal amount of shadow banking.

Shadow banking plays an important role in contemporary finance. As traditional banking, it channels credit to the real economy. The comparison of the assets held by shadow banks with those of the official banking sector reveals that, after a short dip in the aftermath of the global financial crisis, the shadow banking system has been growing at the expenses of traditional banking (IMF 2014). The growth of shadow banking is driven by supply-side as well as demand-side factors.

On the supply-side, shadow banking relies on financial markets for productive efficiency. Shadow banking has managed to replicate the banking functions at a lower cost (Mehrling et al. 2013). Liquidity puts and other explicit or implicit guarantees by the banks sponsoring shadow banking replaced the government safety net. Moreover, shadow banking enables regulatory arbitrage. Capital arbitrage plays a major role given the adverse impact of capital regulation on the profitability of banking. In the past, the banking industry used off-balance-sheet structures to mitigate this impact and perform shadow banking activities. The new capital adequacy requirements established after the Global Financial Crisis (GFC) have placed constraints on the off-balance-sheet operations of banks. However, by increasing the cost of doing banking within the regulatory perimeter, as opposed to shadow banking, the new rules have made regulatory arbitrage even more attractive.

On the demand side, the growth of the shadow banking system is attributable to several macroeconomic developments around the world, which created room for new markets (Pozsar 2015). The savings glut starting from 2003, especially from China and the Middle-East, lead to a sizeable demand for safe assets where those savings could be safely invested. This shift of the demand for safe assets, however, happened in a time when the supply of such assets was diminishing. In that period, particularly in the US, a considerable amount of government debt was retired. This shortage of safe assets prompted the private sector to create such assets, at a profit. In addition to the

savings glut, the rise of professional asset management (mutual funds and pension funds) also generated a demand for safe assets for the optimal management of institutional investors' cash balances. Since deposit insurance throughout the globe is capped at amounts too low for large cash pools, turning to traditional bank liabilities is not an option for institutional investors.

Banks could cater to this demand for safe assets through a very old financial instrument, the repo. Thanks to overcollateralization and the short maturity, the repo provides an efficient substitute for demand deposits. Over-collateralization is realized through a so-called haircut, or margin. The haircut is the proportion of the collateral that the borrower has to own, i.e., the equity/debt ratio (the reciprocal of leverage ratio). Haircuts are the lender's margin of safety. Because the repos' safety does not depend on the amount of the contract, they are suitable for the cash-management needs of institutional investors.

The ability of shadow banking to supply safe assets in response to the rising demand for them is, in principle, efficient for the financial system as well as for the real economy (Caballero and Farhi 2018). Firstly, a shortage of safe assets creates a macroeconomic imbalance where savings exceed investments, which ultimately leads to reduced output and unemployment. Moreover, securitization and Securities Financing Transactions (SFTs), which are at the heart of the shadow banking, are beneficial for the financial markets because they allow for a more efficient use and pricing of financial resources. In sum, by providing alternatives to bank deposits for large investors, shadow banking benefits society.

The key question is, however, the cost of shadow banking for society. This cost can be substantial because of systemic risk. The same mechanism to meet the demand for safe assets by global investors is responsible for the collapse of the financial system when the promise of safety cannot be honored. In economic terms, this means that safe assets are overproduced due to a negative externality. A negative externality means that the social cost of shadow banking is higher than the private cost. Similar to banks, the case for regulating shadow banking stems from

the circumstance that shadow-banking crises impose negative externalities on the rest of the system.

Efficient Regulation of Shadow Banking

The negative externalities of shadow banking are the cost of a financial crisis. A financial crisis occurs because there are runs. Runs are systemic when the insolvency concerns affect banking as a whole, not the idiosyncratic features of one or more institutions. A perception of systemic insolvency can happen because of the interconnectedness between the banks (credit channel of contagion) or because all banks are invested in the same classes of assets and their value suddenly becomes unclear (market channel of contagion).

Shadow banks can credibly promise to convert their short-term liabilities into money by investing their funds in financial assets, which are marketable. Under certain conditions, specified by risk models, these assets can be liquidated safely and deliver the funds necessary to honor the liabilities. One of these conditions is the haircut that intermediaries have to maintain to make good on their liquidity promise in every conceivable state of the world (Geanakoplos 2014). Importantly, when the asset is considered sufficiently liquid, the haircut can be zero – i.e., the collateral itself is sufficient to back the liquidity promise. Moreover, shadow banking often relies on a liquidity put by banks for the very unlikely situation in which the safety conditions do not hold true. Asset safety is based on statistical models that can be proved wrong. When this is the case, liabilities such as repos, as well as the assets backing them, are no longer considered safe. A run on shadow banking ensues. This can spread to the more traditional banking system simply by way of activating the above-mentioned liquidity puts.

A run on shadow banking (Gorton and Metrick 2012) seeks to recreate the safety of the assets backing the shadow bank liabilities by imposing higher haircuts for refinancing. This is problematic because, if the system does not have enough equity to meet the margin call, the price of the collateral will crash. This, in turn, will lead to

increasing haircuts making more financial intermediaries insolvent. However, this is not the only way in which a run can occur on shadow banking. A withdrawal from shadow banking can resemble a more “classic” bank run in two situations. First, the collateral may suddenly become unacceptable (i.e., haircut equal to 100%) and no refinancing be offered. Second, shadow banking liabilities are not always guaranteed by what we would call proper collateral in legal terms. For instance, like demand deposits, MMMF shares are not collateralized. In addition, banks issue other short-term wholesale liabilities, such as commercial paper, which may or may not be collateralized. All of these liabilities are susceptible to runs, too.

The law and economics debate on shadow banking rightly focuses on collateral. Collateral is a distinctive feature of banking – both traditional and shadow – if understood in economic, not legal terms. In economic terms, the assets backing a liquidity promise are collateral because they can be sold in deep markets. Whether lenders can seize them to get their money back is of secondary importance. Shadow banking promises money against securities deemed safe according to the prevailing risk models. Upon materialization of tail risk undermining this perceived safety, a collateral crisis ensues, triggering withdrawals from all sorts of lending, whether or not explicitly collateralized. In the presence of large withdrawals from shadow banking, asset prices go down because the agents with leveraged long positions lose wealth and must sell, whereas new agents cannot leverage as much as to absorb the forced sales. Haircuts are a measure of leverage on specific assets, but as we are speaking about banking, leverage implies liquidity and vice versa. When asset prices crash because of withdrawals from the shadow banking system, market liquidity and funding liquidity shrink simultaneously (Brunnermeier and Pedersen 2009).

A run on shadow banking can only be stopped by a central bank credible enough to recreate the perception of safety that was lost. Showing willingness to buy any quantity of the distressed assets at a given price, the central bank effectively

sets an upper limit on haircuts and, by stabilizing the price of the assets being levered upon, stops a financial crisis (Minsky 1986).

Unfortunately, by stopping a financial crisis, the central bank suspends market discipline, too (Pacces 2013). Assets purchased by the central bank cannot be worth less than what the central bank is willing to pay for them. They become “safe” again precisely for this reason. The expectation of such central bank intervention fuels moral hazard. With moral hazard in place, the negative externalities of shadow banking are exacerbated. The overproduction of safe assets by the private sector, which characterizes shadow banking, would be an even bigger problem if central banks guaranteed the safety of collateral at no cost.

To address this problem, commentators have proposed to charge for the prospective liquidity insurance by the central bank. To operationalize this levy, access to collateral for short-term funding should be restricted. For instance, collateralized borrowing could be reserved to a special class of financial institutions, which would be regulated, and pay for insurance, as banks do for deposit insurance (Gorton and Metrick 2010). Alternatively, the levy could be charged as a condition for collateral to enjoy the bankruptcy law privileges that make it attractive for backing short-term liabilities (Perotti 2013). Collateral in repos and marked-to-market derivatives can be seized immediately in case of default, by way of exception to the bankruptcy rules. Charging for the use of bankruptcy-exempt collateral is effectively a Pigovian tax to correct the externality, aiming to make shadow banks bear ex-ante the social cost of illiquidity.

Approaches based on liquidity insurance have the following shortcomings. First, it would be overly difficult to price this insurance accurately. The unavoidable intervention by central banks in collateral crisis makes liquidity a public good (Heremans and Pacces 2011). Setting the price of public goods is always difficult, but pricing a liquidity put can be a daunting task. To avoid moral hazard, the price of insurance should be risk-based. However, the risk of illiquidity is systemic, which is difficult to define, let alone to

measure. Despite the efforts by authoritative economists to measure systemic risk, this measurement can never be precise. Because illiquidity stems from the failure of credit risk models, there is no reason why systemic risk models, however sophisticated, should be infallible.

As second problem with charging for the central bank’s liquidity put is time inconsistency. The threat to withhold the liquidity put from those who have not paid the insurance premium is not credible. As soon as the collateral becomes systemic, the central bank’s liquidity put is the tool of last resort to avoid financial collapse.

Finally, the economic definition of collateral is much broader than the legal definition. A liquidity tax can only be imposed on entities or instruments we know of. However, how shadow banking will look like in the future is not known. In the future, shadow banking may not rely explicitly on collateral as we know it but still produce money-like liabilities in such proportions as to claim the liquidity put by the central bank during a financial crisis.

According to Nabilou and Pacces (2017), quantity regulation is preferable to a Pigovian tax to cope with the externalities of shadow banking because the social cost – the individual contribution to systemic risk by shadow banks – is difficult to estimate. Moreover, quantity regulation is relatively easy to implement via a single policy variable, which is leverage. Finally, leverage is a straightforward indicator to which regulation can be immediately linked (e.g., as minimum haircut regulation), without need to rely on a complex, thus fallacious, model of systemic risk. Leverage is a necessary condition for shadow banking, so the latter can be identified and cut down in size through the former. Following the economic literature advocating monitoring and curbing leverage on the assets that can be pledged as collateral (Geanakoplos and Pedersen 2014), shadow banking should be regulated indirectly, by capping the admissible levels of leverage on these assets. Regulation can constrain leverage at the asset level by setting up minimum requirements in terms of collateral ownership by the borrower, that is to say, minimum haircut regulations.

As argued by Nabilou and Paccès (2017), minimum haircuts should be established on every debt contract generating short-term, money-like liabilities. This is necessary but not sufficient to curb shadow banking. Short-term liabilities could be issued without explicit reference to collateral. Therefore, leverage should be monitored and restricted at the level of the entities making such promises, too. This is not as difficult as it looks. In practice, liquidity promises by shadow banks are credible only if they are backed by marketable assets or a robust liquidity put. Banks are able to make unsecured promises because they rely on a web of credit and liquidity puts from the public sector. For this reason, banks' capital is directly regulated. Such a regulation does extend to shadow banking in as much as the latter is sponsored by banks and such sponsorships count towards the leverage contains imposed on banks.

An indirect regulation of shadow banking based on restricting the channels for money creation has the advantage to cope with financial innovation. These channels are defined broadly enough to capture different ways in which shadow banking may play out in the future and in different jurisdictions. The latter circumstance is very useful for international cooperation, which is essential to regulate shadow banking effectively. At the same time, asset leverage and bank leverage restrictions are sufficiently specific concepts to embed into legislation. This overcomes the legitimacy concerns that could arise if these restrictions were to be established exclusively by regulatory agencies.

Two challenges remain, however. The first challenge is to set the minimum haircuts at the right level, bearing in mind that excessive curbs on (shadow) banking undermine efficiency. The second challenge is to adjust these regulations over time. Dealing with the first problem requires statistical models, similar to those necessary to identify the optimal restrictions for banking institutions. Asset-based models of risk assessment, however, are much simpler than the models used to measure systemic risk. Setting conservative haircuts only requires identifying the worst-case price scenario for each class of assets.

The second problem is more difficult to deal with. On the one hand, new classes of assets may be used for leverage. On the other hand, the conservative haircuts will have to be adapted to new circumstances, in one direction or another. The adaptation problem depends on the same reason why, in the end, all preventative regulations for minimizing systemic risk – including shadow banking regulation – will fail and require crisis management. This is a general problem in the pursuit of financial stability as a policy goal. Arguably, the regulation of leverage, particularly at the asset level, fares better than alternative approaches to regulate shadow banking. However, the efficient regulation of shadow banking, as of any other source of systemic risk, calls upon financial regulators to exercise some discretion in order to adapt to unforeseen circumstances.

Conclusion and Future Research

This essay discusses the economics of shadow banking and the case for its regulation. An important challenge in addressing shadow banking is to define it. All the main approaches to defining shadow banking have shortcomings. An instrument-based approach seems to be preferable to cope with systemic risk in the presence of financial innovation. However, an effective instrument-based regulation presupposes that the entities having access to the relevant instruments are subject to reporting obligations.

Shadow banking is beneficial for society because it caters to a demand for safe assets. However, because of negative externalities, shadow banking is overproduced. Quantity regulation is preferable to a Pigovian tax to cope with the externalities of shadow banking because uncertainty makes any measure of systemic risk imprecise. Regulation should limit the leverage of shadow banking by way of minimum haircuts regulation on assets being used as collateral. Moreover, shadow banking should be regulated indirectly, through the capital regulation of banks providing explicit or implicit liquidity puts.

The implementation of an optimal regulation of shadow banking faces two important challenges.

First, it is difficult to identify the optimal levels of asset and institutional leverage. Second, it is even more difficult to adapt these levels to new circumstances. These challenges, however, apply as well to banking in general. A way to address them may be to provide financial regulators with some discretion under incentive-compatible regulatory governance. This is an interesting avenue for future research.

Cross-References

- Banks
- Central Bank
- Exploring the Deterrent Impact of Financial Supervisory Liability
- Externalities
- Law and Finance

Acknowledgments This essay is partly based on Nabilou and Paccès (2017).

References

- Adrian T, Ashcraft AB (2016) Shadow banking: a review of the literature. In: Garrett J (ed) *Banking crises: perspectives from the new Palgrave dictionary*. Palgrave Macmillan, New York
- Adrian T, Shin HS (2010) The changing nature of financial intermediation and the financial crisis of 2007–2009. *Ann Rev Econom* 2:603–618
- Brunnermeier M, Krishnamurthy A (2014) *Risk topography: systemic risk and macro modeling*. University of Chicago Press, Chicago
- Brunnermeier MK, Pedersen LH (2009) Market liquidity and funding liquidity. *Rev Financ Stud* 22(6): 2201–2238
- Caballero RJ, Farhi E (2018) The safety trap. *Rev Econ Stud*. forthcoming
- Claessens S, Ratnovski L (2014) What is shadow banking? IMF working paper no. 14/25, Washington, DC
- FSB (Financial Stability Board) (2011) Shadow banking: strengthening oversight and regulation recommendations of the financial stability board. Bank for International Settlements, Basel
- FSB (Financial Stability Board) (2015) Global shadow banking monitoring report 2015. Bank for International Settlements, Basel
- Geanakoplos J (2014) Leverage, default, and forgiveness: lessons from the American and European crises. *J Macroecon* 39:313–333
- Geanakoplos J, Pedersen LH (2014) Monitoring leverage. In: Brunnermeier M, Krishnamurthy A (eds) *Risk topography: systemic risk and macro modeling*. University of Chicago Press, Chicago, pp 113–127
- Gorton GB, Metrick A (2010) Regulating the shadow banking system. *Brook Pap Econ Act* 41(2): 261–312
- Gorton GB, Metrick A (2012) Securitized banking and the run on repo. *J Financ Econ* 104(3):425–451
- Heremans D, Paccès AM (2011) Regulation of banking and financial markets. In: Van den Bergh R, Paccès AM (eds) *Encyclopedia of law and economics*, vol 9. Edward Elgar, Cheltenham, pp 558–606
- IMF (International Monetary Fund) (2014) Shadow banking around the globe: how large, and how risky? In: *Global financial stability report*, October, chapter 2, Washington, DC
- Mehrling P et al (2013) Bagehot was a shadow banker: shadow banking, central banking, and the future of global finance. <https://doi.org/10.2139/ssrn.2232016>.
- Minsky HP (1986) *Stabilizing an unstable economy*. McGraw-Hill, New York
- Nabilou H, Paccès AM (2017) The law and economics of shadow banking. ECGI law working paper no. 339/2017. <http://www.ecgi.global/working-paper/law-and-economics-shadow-banking>.
- Paccès AM (2013) The future in law and finance. Erasmus law lectures No. 32. Eleven International Publishing, Den Haag
- Perotti E (2013) The roots of shadow banking. CEPR Policy Insight No. 69
- Pozsar Z (2014) Shadow banking: The money view. Office of financial research working paper no. 14–04, Washington, DC
- Pozsar Z (2015) A macro view of shadow banking: levered betas and wholesale funding in the context of secular stagnation. <https://doi.org/10.2139/ssrn.2558945>.

Shadow Economy

Friedrich Schneider

Department of Economics, Johannes Kepler University of Linz, Linz-Auhof, Austria

Abstract

In this paper the main focus lies on the shadow economy. The most influential factors on the shadow economy are tax policies and state regulation. The size of the shadow economy was decreasing over 1999 to 2007 from 34.0% to 31.2% for 161 countries (unweighted average).

Keywords

Shadow economy; Tax morale; Tax pressure; State regulation; Labor market

Synonyms

[Black economy](#); [Informal economy](#); [Unofficial economy](#)

Definition of the Shadow Economy

Up to today, authors trying to measure the shadow economy face the difficulty of a precise definition of the shadow economy. According to one commonly used definition, it comprises all currently unregistered economic activities that contribute to the officially calculated gross national product. Do-it-yourself activities are not included. This definition is used, e.g., by Feige (1989, 1994), Schneider (1994a, 2003, 2005), and Frey and Pommerehne (1984). Smith (1994, p. 18) defines it as “market-based production of goods and services, whether legal or illegal, that escapes detection in the official estimates of GDP.” Put differently, one of the broadest definitions is: “...those economic activities and the income derived from them that circumvent or otherwise avoid government regulation, taxation or observation.” This definition is taken from Dell’Anno (2003), Dell’Anno and Schneider (2004), and Feige (1989).

In this entry the following more narrow definition of the shadow economy is used. The shadow economy includes all market-based legal production of goods and services that are deliberately concealed from public authorities for the following reasons:

1. To avoid payment of income, value added, or other taxes
2. To avoid payment of social security contributions
3. To avoid having to meet certain legal labor market standards, such as minimum wages, maximum working hours, safety standards, etc.

4. To avoid complying with certain administrative obligations, such as completing statistical questionnaires or other administrative forms

Thus, I will not deal with typically illegal underground economic activities that fit the characteristics of classical crimes like burglary, robbery, drug dealing, etc. I also exclude the informal household economy which consists of all household services and production.

Introduction

Fighting tax evasion and the shadow economy have been important policy goals in OECD countries during recent decades. In order to do this, one should have knowledge about the size and development of the shadow economy as well as the reasons why people are engaged in shadow economy activities. This is the content of this entry. Tax evasion is not considered in order to keep the subject of this entry tractable and because too many additional aspects would be involved. Also tax morale or experimental studies on tax compliance are beyond the scope of this entry.

My entry is organized as follows: section “[Some Theoretical Considerations About the Shadow Economy](#)” presents theoretical considerations about the measurement of the shadow economy and discusses also the main factors determining its size. In section “[Size of the Shadow Economies all Over the World](#)” the empirical results of the size and development of the shadow economy are discussed. Finally section “[Conclusions](#)” concludes.

Some Theoretical Considerations About the Shadow Economy**Measuring the Shadow Economy**

The definition of the shadow economy plays an important role in assessing its size. By having a clear definition, a number of ambiguities and controversies can be avoided. In general, there are two types of shadow economic activities: illicit employment and the in the household produced

goods and services mostly consumed within the household. The following analysis focuses on both types but tries to exclude illegal activities such as drug production, crime, and human trafficking. The in the household produced goods and services, e.g., schooling and childcare are not part of this analysis. Thus, it only focuses on productive economic activities that would normally be included in the national accounts but which remain underground due to tax or regulatory burdens. With this definition the problem of having classical crime activities included could be avoided, because neither the MIMIC procedure nor the currency demand approach captures these activities: e.g., drug dealing is independent of increasing taxes, especially as the included causal variables are not linked (or causal) to classical crime activities. See, e.g., Thomas (1992), Kazemir (2005a, b), and Schneider (2005). Although such legal activities contribute to the country's value added, they are not captured in the national accounts because they are produced in illicit ways (e.g., by people without proper qualification or without a master craftsman's certificate). From the economic and social perspective, soft forms of illicit employment, such as moonlighting (e.g., construction work in private homes), and its contribution to aggregate value added can be assessed rather positively.

Although the issue of the shadow economy has been investigated for a long time, the discussion regarding the "appropriate" methodology to assess its scope has not come to an end yet. For the strengths and weaknesses of the various methods see, e.g., Dell'Anno and Schneider (2009), Feige (1989), and Feld and Larsen (2005). There are three methods of assessment:

- (1) Direct procedures at a micro level that aim at determining the size of the shadow economy at one particular point in time. An example is the survey method
- (2) Indirect procedures that make use of macroeconomic indicators in order to proxy the development of the shadow economy over time
- (3) Statistical models that use statistical tools to estimate the shadow economy as an "unobserved" variable

Today in many cases the estimation of the shadow economy is based on a combination of the MIMIC procedure and on the currency demand method or the use of only the currency demand method. These methods are presented in detail in Schneider (1994a, b, c, 2005, 2011, 2014), Schneider and Williams (2013), Feld and Schneider (2010), and Schneider and Enste (2000, 2002, 2006). The MIMIC procedure assumes that the shadow economy remains an unobserved phenomenon (latent variable) which can be estimated using quantitatively measurable causes of illicit employment, e.g., tax burden and regulation intensity, and indicators reflecting illicit activities, e.g., currency demand, official GDP, and official working time. A disadvantage of the MIMIC procedure is the fact that it produces only relative estimates of the size and the development of the shadow economy. Thus, the currency demand method is used to calibrate the relative into absolute estimates (e.g., in percent of GDP) by using two or three absolute values (in percent of GDP) of the size of the shadow economy. This indirect approach is based on the assumption that cash is used to make transactions within the shadow economy. By using this method one econometrically estimates a currency demand function including independent variables like tax burden, regulation, etc. which "drive" the shadow economy. This equation is used to make simulations of the amount of money that would be necessary to generate the official GDP. This amount is then compared with the actual money demand, and the difference is treated as an indicator for the development of the shadow economy. On this basis the calculated difference is multiplied by the velocity of money of the official economy and one gets a value added figure for the shadow economy.

In addition, the size of the shadow economy is estimated by using survey methods (Feld and Larsen (2005, 2008, 2009). In order to minimize the number of respondents dishonestly replying or totally declining answers to the sensitive questions, structured interviews are undertaken (usually face-to-face) in which the respondents are slowly getting accustomed to the main purpose of the survey. Like it is done by the contingent valuation method (CVM) in environmental

economics (Kopp et al. 1997), a first part of the questionnaire aims at shaping respondents’ perception as to the issue at hand. In a second part, questions about respondents’ activities in the shadow economy are asked, and the third part contains the usual sociodemographic questions.

In addition to the studies by Merz and Wolff (1993), Feld and Larsen (2005, 2008, 2009), Haigner et al. (2013), and Enste and Schneider (2006) for Germany, the survey method has been applied in the Nordic countries and Great Britain (Isachsen and Strøm 1985; Pedersen 2003) as well as in the Netherlands (van Eck and Kazemier 1988; Kazemier 2006). While the questionnaires underlying these studies are broadly comparable in design, recent attempts by the European Union to provide survey results for all EU member states run into difficulties regarding comparability (Renooy et al. 2004; European Commission 2007); the wording of the questionnaires becomes

more and more cumbersome depending on the culture of different countries with respect to the underground economy.

To summarize: Although each method has its strength and weaknesses and biases in the estimates of the shadow economy almost certainly prevail, no better data are currently available. Clearly, there can be no exact measure of the size of the shadow economy, and estimates differ widely with an error margin of $\pm 15\%$. These days, macro estimates derived from the MIMIC model, the currency demand method, or the electricity approach are seen as upper bound estimates, while micro (survey) estimates are seen as lower bound estimates.

The Main Causes Determining the Shadow Economy

Table 1 presents an overview of the most important determinants influencing the shadow

Shadow Economy, Table 1 The main causes determining the shadow economy

Causal variable	Theoretical reasoning	References
Tax and social security contribution burdens	The distortion of the overall tax burden affects labor-leisure choices and may stimulate labor supply in the shadow economy. The bigger the difference between the total labor cost in the official economy and after-tax earnings (from work), the greater is the incentive to reduce the tax wedge and to work in the shadow economy. This tax wedge depends on social security burden/payments and the overall tax burden, making them to key determinants for the existence of the shadow economy	E.g., Thomas (1992), Johnson, Kaufmann, and Zoido-Lobaton (1998a, b), Giles (1999), Tanzi (1999), Schneider (2003, 2005), Dell’Anno (2007), Dell’Anno et al. (2007), Buehn and Schneider (2012)
Quality of institutions	The quality of public institutions is another key factor for the development of the informal sector. Especially the efficient and discretionary application of the tax code and regulations by the government plays a crucial role in the decision to work underground, even more important than the actual burden of taxes and regulations. In particular, a bureaucracy with highly corrupt government officials seems to be associated with larger unofficial activity, while a good rule of law by securing property rights and contract enforceability increases the benefits of being formal. A certain level of taxation, mostly spent in productive public services, characterizes efficient policies. In fact, the production in the formal sector benefits from a higher provision of productive public services and is negatively affected by taxation, while the shadow economy reacts in the opposite way. An informal sector developing as a consequence of the failure of political institutions in promoting an efficient market economy and entrepreneurs going underground, as there	E.g., Johnson et al. (1998a, b), Friedman et al. (2000), Dreher and Schneider (2009), Dreher et al. (2009), Schneider (2010), Buehn and Schneider (2012), Teobaldelli (2011), Teobaldelli and Schneider (2012), Amendola and Dell’Anno (2010), Losby et al. (2002), Schneider and Williams (2013)

(continued)

Shadow Economy, Table 1 (continued)

Causal variable	Theoretical reasoning	References
	is an inefficient public goods provision, may reduce if institutions can be strengthened and fiscal policy gets closer to the median voter's preferences	
Regulations	Regulations, for example, labor market regulations or trade barriers, are another important factor that reduces the freedom (of choice) for individuals in the official economy. They lead to a substantial increase in labor costs in the official economy and thus provide another incentive to work in the shadow economy; countries that are more heavily regulated tend to have a higher share of the shadow economy in total GDP. Especially the enforcement and not the overall extent of regulation – mostly not enforced – is the key factor for the burden levied on firms and individuals, making them operate in the shadow economy	E.g., Johnson, Kaufmann, and Shleifer (1997), Johnson, Kaufmann, and Zoido-Lobaton (1998b), Friedman et al. (2000), Kucera and Roncolato (2008), Schneider (2011)
Public sector services	An increase of the shadow economy may lead to fewer state revenues, which in turn reduce the quality and quantity of publicly provided goods and services. Ultimately, this may lead to increasing tax rates for firms and individuals, although the deterioration in the quality of the public goods (such as the public infrastructure) and of the administration continues. The consequence is an even stronger incentive to participate in the shadow economy. Countries with higher tax revenues achieved by lower tax rates, fewer laws and regulations, a better rule of law, and lower corruption levels should thus have smaller shadow economies	E.g., Johnson, Kaufmann, and Zoido-Lobaton (1998a, b), Feld and Schneider (2010)
Tax morale	The efficiency of the public sector also has an indirect effect on the size of the shadow economy because it affects tax morale. Tax compliance is driven by a psychological tax contract that entails rights and obligations from taxpayers and citizens on the one hand but also from the state and its tax authorities on the other hand. Taxpayers are more heavily inclined to pay their taxes honestly if they get valuable public services in exchange. However, taxpayers are honest even in cases when the benefit principle of taxation does not hold, i.e., for redistributive policies, if such political decisions follow fair procedures. The treatment of taxpayers by the tax authority also plays a role. If taxpayers are treated like partners in a (tax) contract instead of subordinates in a hierarchical relationship, taxpayers will stick to their obligations of the psychological tax contract more easily. Hence, (better) tax morale and (stronger) social norms may reduce the probability of individuals to work underground	E.g., Feld and Frey (2007), Kirchler (2007), Torgler and Schneider (2009), Feld and Larsen (2005, 2009), Feld and Schneider (2010)
Deterrence	Despite the strong focus on deterrence in policies fighting the shadow economy and the unambiguous insights of the traditional economic theory of tax noncompliance, surprisingly little is known about the effects of deterrence from empirical studies. This is due to the fact that data on the legal background and the frequency of audits are not available on an international basis; even for OECD countries such data is difficult to collect. Either is the legal background quite complicated differentiating fines and punishment according to the severity of the offense	E.g., Andreoni et al. (1998), Pedersen (2003), Feld and Larsen (2005, 2009), Feld and Schneider (2010)

(continued)

Shadow Economy, Table 1 (continued)

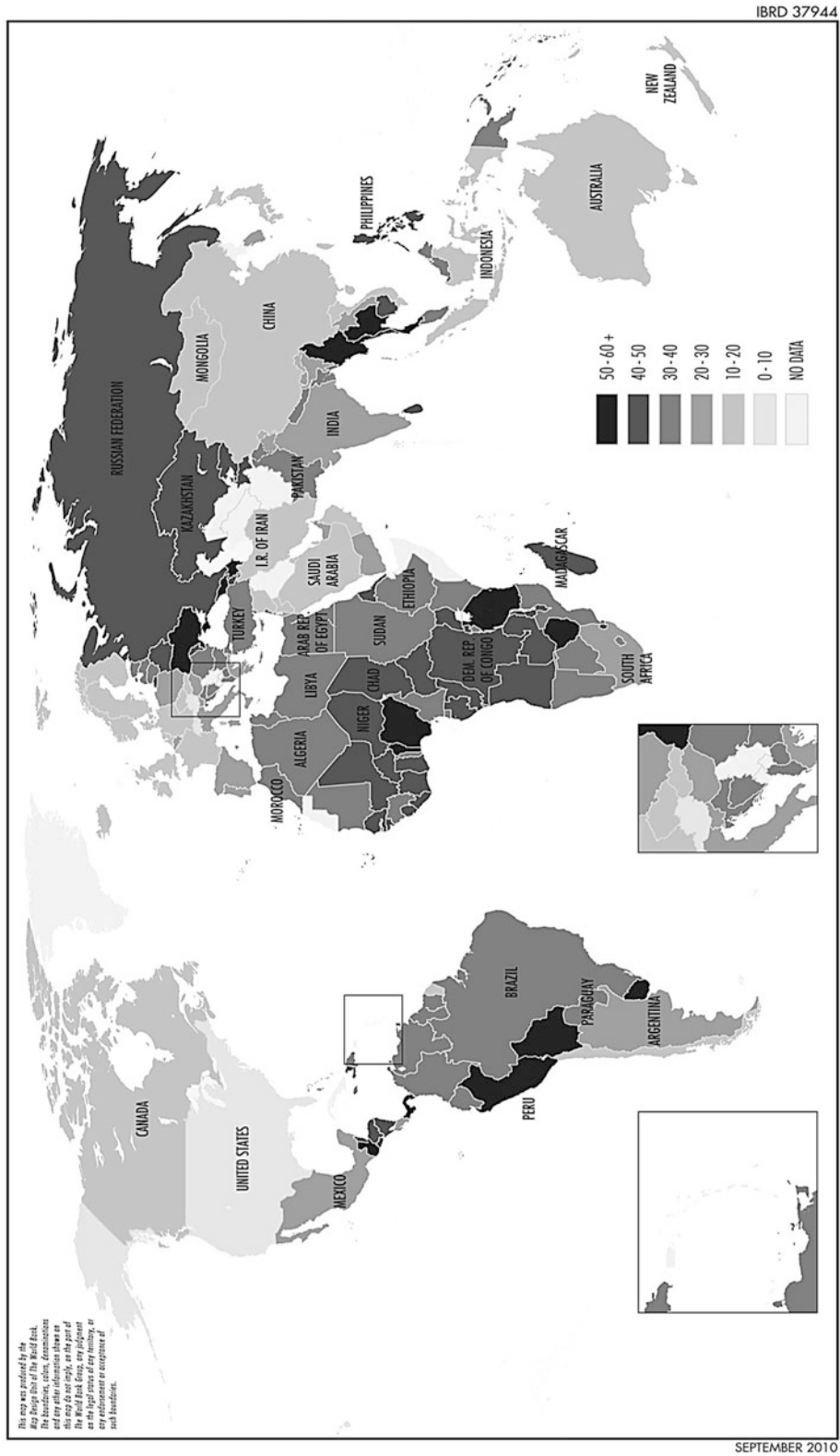
Causal variable	Theoretical reasoning	References
	and the true income of the noncomplier or tax authorities do not reveal how intensively auditing is taking place. The little empirical survey evidence available demonstrates that fines and punishment do not exert a negative influence on the shadow economy, while the subjectively perceived risk of detection does. However, the results are often weak and Granger causality tests show that the size of the shadow economy can impact deterrence instead of deterrence reducing the shadow economy	
Agricultural sector	The importance of agriculture in the economy is included, since many studies endorse the idea that informal work is concentrated in highly segmented sectors, with clear prevalence for the agricultural and related sectors. One of the most important reasons for this is the minimum enforcement capacity of governments prevalent in rural areas. The importance of agriculture is measured as the share of agriculture as percentage of GDP. The larger the agricultural sector, the larger the expected size of the shadow economy, <i>ceteris paribus</i>	E.g., Vuletin (2008), De la Roca et al. (2002), Greenidge et al. (2005), Mootoo et al. (2002), Amendola and Dell'Anno (2010), Losby et al. (2002)
Development of the official economy	The development of the official economy is another key factor of the shadow economy. The higher (lower) the unemployment quota (GDP growth), the higher is the incentive to work in the shadow economy, <i>ceteris paribus</i>	Schneider and Williams (2013) Feld and Schneider (2010)
Self-employment	The higher self-employment is, the more activities can be done in the shadow economy, <i>ceteris paribus</i>	Schneider and Williams (2013) Feld and Schneider (2010)

economy. Due to space reasons, there is no detailed discussion of the various determinants/causes of the shadow economy.

Size of the Shadow Economies all Over the World

The figures below are taken from Schneider et al. (2010). The econometric MIMIC estimation results are not shown here due to space reasons; see, e.g., Schneider et al. (2010). Figure 1 shows the average size of the shadow economy of 162 countries over 1999–2007. In Tables 2 and 3 the average informality (unweighted and weighted) in different regions is shown using the regions defined by the World Bank. The World Bank distinguishes 8 world regions which are East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa, High Income OECD, Other High

Income, South Asia, and Sub-Saharan Africa. If we consider first Table 2 where the average informality (unweighted) is shown, we see that Latin America and the Caribbean have the highest value of the shadow economies of 41.1%, followed by Sub-Saharan Africa of 40.2%, and then followed by Europe and Central Asia of 38.9%. The lowest has the High Income OECD countries with 17.1%. If we consider the average informality of the shadow economies of these regions weighted by total GDP in 2005, Sub-Saharan Africa has the highest with 37.6%, followed by Europe and Central Asia with 36.4%, and Latin America and the Caribbean with 34.7%. The lowest again has the High Income OECD with 13.4%. If one considers the world mean weighted and unweighted, one sees that if one uses the unweighted measures, the mean is 33.0% over the periods 1999–2007. If we consider the world with weighted informality measures, the shadow economy takes “only” a value of 17.1% over the period 1999–2007.



Shadow Economy, Fig. 1 Average size of the shadow economy of 162 countries over 1999–2007 (Source: Schneider et al. 2010)

Shadow Economy, Table 2 Average informality (unweighted) by World Bank's regions

	Region	Mean	Median	Min	Max	sd
EAP	East Asia and Pacific	32.3	32.4	12.7	50.6	13.3
ECA	Europe and Central Asia	38.9	39.0	18.1	65.8	10.9
LAC	Latin America and the Caribbean	41.1	38.8	19.3	66.1	12.3
MENA	Middle East and North Africa	28.0	32.5	18.3	37.2	7.8
OECD	High Income OECD	17.1	15.8	8.5	28.0	6.1
OHIE	Other High Income	23.0	25.0	12.4	33.4	7.0
SAS	South Asia	33.2	35.3	22.2	43.9	7.0
SSA	Sub-Saharan Africa	40.2	40.6	18.4	61.8	8.3
World		33.0	33.5	8.5	66.1	12.8

Source: Schneider et al. (2010)

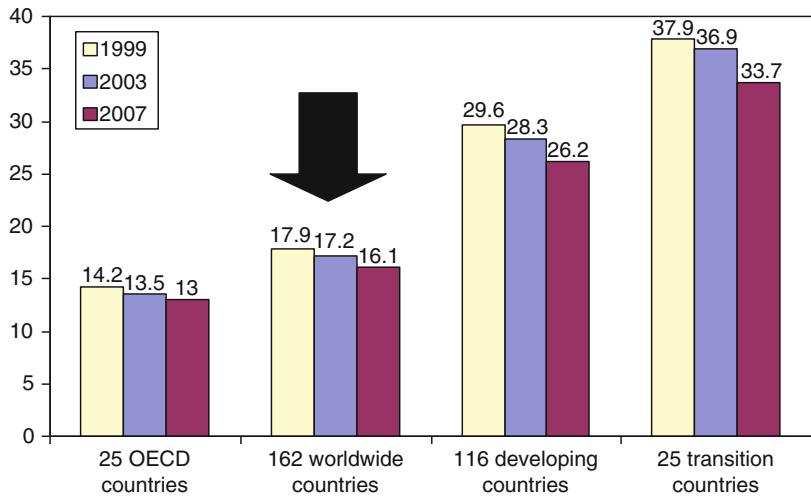
Shadow Economy, Table 3 Average informality (weighted) by total GDP in 2005

	Region	Mean	median	min	max	sd
EAP	East Asia and Pacific	17.5	12.7	12.7	50.6	10.6
ECA	Europe and Central Asia	36.4	32.6	18.1	65.8	8.4
LAC	Latin America and the Caribbean	34.7	33.8	19.3	66.1	7.9
MENA	Middle East and North Africa	27.3	32.5	18.3	37.2	7.7
OECD	High Income OECD	13.4	11.0	8.5	28.0	5.7
OHIE	Other High Income	20.8	19.4	12.4	33.4	4.9
SAS	South Asia	25.1	22.2	22.2	43.9	5.9
SSA	Sub-Saharan Africa	37.6	33.2	18.4	61.8	11.7
World		17.1	13.2	8.5	66.1	9.9

Source: Schneider et al. (2010)

Shadow Economy,

Fig. 2 Size and development of the shadow economy of various countries' groups (weighted averages (!); in percent of official total GDP of the respective Country Group) (Source: Schneider et al. 2010)



Weighting the values makes a considerable difference.

One general result of the size and development of the shadow economies worldwide is that there is an overall reduction in the size. In Fig. 2 the size

and development of the shadow economy of various countries' groups (weighted averages by the official GDP of 2005) over 1999, 2003, and 2007 are shown. One clearly realizes that for all countries' groups (25 OECD countries, 116 developing

countries, 25 transition countries) I observe a decrease in the size of the shadow economy. The average size of the shadow economies of the 162 countries was 34.0% of official GDP (unweighted measure!) in 1999 and decreased to 31.2% of official GDP in 2007. This is a decrease of almost 3.0 percentage points over 9 years. Growth of the official economy with reduced (increased) unemployment (employment) seems to be the most efficient mean to reduce the shadow economy.

Conclusions

In this entry some of the most recent developments in research on the shadow economy in highly developed OECD, developing, and transition countries are shown. The discussion of the recent literature shows that economic opportunities for employees, the overall situation on the labor market, not least unemployment are crucial for an understanding of the dynamics of the shadow economy. Individuals look for ways to improve their economic situation and thus contribute productively to aggregate income of a country. This holds regardless of their being active in the official or the unofficial economy.

A last question remaining is: What type of policy conclusions can I draw? One conclusion may be that – besides the indirect tax and personal income tax burden, which the government can directly influence by policy actions – self-employment and unemployment are two very important driving forces of the shadow economy. Unemployment may be controllable by the government through economic policy in a traditional Keynesian sense; alternatively, the government can try to improve the country's competitiveness to increase foreign demand. The impact of self-employment on the shadow economy is less or only partly controllable by the government and may be ambiguous from a welfare perspective. A government can deregulate the economy or incentivize "to be your own entrepreneur," which would make self-employment easier, potentially reducing unemployment and positively

contributing to efforts in controlling the size of the shadow economy. Such actions however need to be accompanied with a strengthening of institutions and tax morale to reduce the probability that self-employed shift reasonable proportions of their economic activities into the shadow economy, which, if it happened, made government policies incentivizing self-employment less effective. This entry clearly shows that a reduction of the shadow economy can be achieved using various channels the government can influence. The main challenge still is to bring shadow economic activities into the official economy in a way that goods and services previously produced in the shadow economy are still produced and provided but in the official economy. Only then, the government gets additional taxes and social security contributions.

Finally, if I ask what we know about the shadow economy, I clearly realize that we have some knowledge about the size and development of the shadow economy. What we do not know are the exact motives, why people work in the shadow economy, and what is their relation and feeling if a government undertakes reforms in order to bring them back into the official economy. Hence, much more micro studies are needed to obtain a more detailed knowledge about people's motivation to work either in the shadow economy and/or in the official one.

Cross-References

► [Informal Sector](#)

References

- Amendola A, Dell'Anno R (2010) Institutions and human development in the Latin America shadow economy. *Estudios en Derecho y Gobierno* 3(1):9–25
- Andreoni J, Erard B, Feinstein J (1998) Tax compliance. *J Econ Lit* 36(4):818–860
- Buehn A, Schneider F (2012) Shadow economies around the world: novel insights, accepted knowledge, and new estimates. *Int Tax Public Finan* 19:139–171
- De la Roca J, Hernandez M, Robles M, Torero M, Webber M (2002) Informal Sector Study for Jamaica, Group of Analysis for Development. Inter-American Development Bank, Washington, DC

- Dell'Anno R (2003) Estimating the shadow economy in Italy: a structural equation approach. Working paper 2003–7. Department of Economics, University of Aarhus
- Dell'Anno R (2007) The shadow economy in Portugal: an analysis with the MIMIC approach. *J Appl Econ* 10:253–277
- Dell'Anno R, Schneider F (2004) The shadow economy of Italy and other OECD countries: what do we know? *J Public Finan Public Choice* 21:223–245
- Dell'Anno R, Schneider F (2009) A complex approach to estimate shadow economy: the structural equation modelling. In: Faggnini M, Looks T (eds) *Coping with the complexity of economics*. Springer, Berlin, pp 110–130
- Dell'Anno R, Gomez-Antonio M, Alanon Pardo A (2007) Shadow economy in three different Mediterranean countries: France, Spain and Greece. A MIMIC approach. *Emp Econ* 33:51–84
- Dreher A, Schneider F (2009) Corruption and the shadow economy: an empirical analysis. *Public Choice* 144(2):215–277
- Dreher A, Kotsogiannis C, McCorriston S (2009) How do institutions affect corruption and the shadow economy? *Int Tax Public Finan* 16(4):773–796
- Eck R, Kazemier B (1988) Features of the hidden economy in the Netherlands. *Rev Income Wealth* 34(3):251–273
- Enste D, Schneider F (2006) Umfang und Entwicklung der Schattenwirtschaft in 145 Ländern. In: Schneider F, Enste D (eds) *Jahrbuch Schattenwirtschaft 2006/07. Zum Spannungsfeld von Politik und Ökonomie*. LIT Verlag, Berlin, pp 55–80
- Commission E (2007) Stepping up the fight against undeclared work, COM (2007) 628 final. European Commission, Brussels
- Feige EL (ed) (1989) *The underground economies. Tax evasion and information distortion*. Cambridge University Press, Cambridge
- Feige EL (1994) The underground economy and the currency enigma. *Suppl Public Financ* 49:119–136
- Feld LP, Frey BS (2007) Tax compliance as the result of a psychological tax contract: the role of incentives and responsive regulation. *Law Policy* 29(1):102–120
- Feld LP and Larsen C (2005) Black activities in Germany in 2001 and 2004: a comparison based on survey data, Study No. 12. The Rockwool Foundation Research Unit, Copenhagen
- Feld LP, Larsen C (2008) “Black” activities low in Germany in 2006. News from the Rockwool Foundation Research Unit, March 2008, pp 1–12
- Feld LP, Larsen C (2009) Undeclared work in Germany 2001–2007 – impact of deterrence, tax policy, and social norms: an analysis based on survey data. Springer, Berlin
- Feld LP, Schneider F (2010) Survey on the shadow economy and undeclared earnings in OECD countries. *German Econ Rev* 11(2):109–149
- Frey BS, Pommerehne W (1984) The hidden economy: state and prospect for measurement. *Rev Income Wealth* 30(1):1–23
- Friedman E, Johnson S, Kaufmann D, Zoido-Lobaton P (2000) Dodging the grabbing hand: the determinants of unofficial activity in 69 countries. *J Public Econ* 76(4):459–493
- Giles DEA (1999) Measuring the hidden economy: implications for econometric modelling. *Econ J* 109(3):370–380
- Greenidge K, Holder C, Mayers S (2005) Estimating the size of the underground economy in Barbados. Paper presented at the 26th annual review seminar. Research Department, Central Bank of Barbados, 26–29 July 2005
- Haigner S, Jenewein S, Schneider F, Wakolbinger F (2013) Driving forces of informal labour supply and demand in Germany. *Int Labour Rev* 152(3–4):507–524
- Isachsen AJ, Strøm S (1985) The size and growth of the hidden economy in Norway. *Rev Income Wealth* 31(1):21–38
- Johnson S, Kaufmann D, Shleifer A (1997) The unofficial economy in transition. *Brookings Papers on Economic Activity* No. 2, pp 159–221
- Johnson S, Kaufmann D, Zoido-Lobaton P (1998a) Regulatory discretion and the unofficial economy. *Am Econ Rev, Paper Proceed* 88(2):387–392
- Johnson S, Kaufmann D and Zoido-Lobaton P (1998b) Corruption, public finances and the unofficial economy. World Bank Policy Research Working Paper Series No. 2169. The World Bank, Washington DC
- Kazemier B (2005a) The underground economy: a survey of methods and estimates. Discussion paper. Statistics Netherlands, Voorburg
- Kazemier B (2005b) Monitoring the underground labor market: what surveys can do. Discussion paper. Statistics Netherlands, Voorburg
- Kazemier B (2006) Monitoring the underground economy: a survey of methods and estimates. In: Schneider F, Enste D (eds) *Jahrbuch Schattenwirtschaft 2006/07. Zum Spannungsfeld von Politik und Ökonomie*. LIT Verlag, Berlin, pp 11–53
- Kirchler E (2007) *The economic psychology of tax behaviour*. Cambridge University Press, Cambridge
- Kopp RJ, Pommerehne WW, Schwarz N (1997) Determining the value of non-marketed goods: economic, psychological, and policy relevant aspects of contingent valuation methods. Kluwer, Boston
- Kucera D, Roncolato L (2008) Informal employment: two contested policy issues. *Int Labor Rev* 147(3):321–348
- Losby JL, Else JF, Kingslow ME, Edgcomb EL, Malm ET and Kao V (2002) *Informal economy literature review*. The Aspen Institute, Microenterprise Fund for Innovation, Effectiveness, Learning and Dissemination, Washington, DC, and ISED Consulting and Research, Newark
- Merz J, Wolff KG (1993) The shadow economy: illicit work and household production – a microanalysis of West Germany. *Rev Income Wealth* 39(2):177–194
- Mootoo L, Sookram S, Watson PK (2002) Size and structure of the hidden economy in the Caribbean, Economic Measurement Unit, Department of Economics, University of the West Indies, St. Augustine, Trinidad & Tobago
- Pedersen S (2003) The shadow economy in Germany, Great Britain and Scandinavia: a measurement based

- on questionnaire service. Study No. 10. The Rockwool Foundation Research Unit, Copenhagen
- Renooy P, Ivarsson S, van der Wusten-Gritsai O, Meijer E (2004) Undeclared work in an enlarged union – an analysis of undeclared work: an in-depth study of specific items. European Commission, Brussels
- Schneider F (1994a) Measuring the size and development of the shadow economy. Can the causes be found and the obstacles be overcome? In: Brandstaetter H, Güth W (eds) *Essays on economic psychology*. Springer, Berlin, pp 193–212
- Schneider F (1994b) Determinanten der Steuerhinterziehung der Schwarzarbeit im internationalen Vergleich. In: Smekal C, Theurl E (eds) *Stand und Entwicklung der Finanzpsychologie*. Nomos, Baden-Baden, pp 247–288
- Schneider F (1994c) Can the shadow economy be reduced through major tax reforms? An empirical investigation for Austria. *Suppl Public Financ* 49:137–152
- Schneider F (2003) Shadow economy. In: Rowley CK, Schneider F (eds) *Encyclopedia of public choice*, vol II. Kluwer, Dordrecht, pp 286–296
- Schneider F (2005) Shadow economies around the world: what do we really know? *Eur J Polit Econ* 21(4):598–642
- Schneider F (2010) The influence of public institutions on the shadow economy: an empirical investigation for OECD countries. *Eur J Law Econ* 6(3):441–468
- Schneider F (ed) (2011) *Handbook on the shadow economy*. Edward Elgar, Cheltenham
- Schneider F (2014) The shadow economy and shadow labor force: results, problems and open questions. Discussion paper. Department of Economics, University of Linz, Linz
- Schneider F, Enste D (2000) Shadow economies: size, causes and consequences. *J Econ Lit* 38(1):73–110
- Schneider F, Enste D (2002) *The shadow economy: theoretical approaches, empirical studies, and political implications*. Cambridge University Press, Cambridge, UK
- Schneider F, Enste D (eds) (2006) *Jahrbuch Schattenwirtschaft 2006/07. Zum Spannungsfeld von Politik und Ökonomie*. LIT Verlag, Berlin
- Schneider F, Williams CC (2013) *The shadow economy*. Institute of Economic Affairs, London
- Schneider F, Buehn A, Montenegro CE (2010) New estimates for the shadow economies all over the world. *Int Econ J* 24(4):443–461
- Smith P (1994) Assessing the size of the underground economy: the Canadian statistical perspectives. *Canadian Economic Observer*, Catalogue No. 11–010, pp 16–33
- Tanzi V (1999) Uses and abuses of estimates of the underground economy. *Econ J* 109(3):338–347
- Teobaldelli D (2011) Federalism and the shadow economy. *Public Choice* 146(3):269–289
- Teobaldelli D and Schneider F (2012) Beyond the veil of ignorance: the influence of direct democracy on the shadow economy. CESifo working paper MO3749. University of Munich, Munich
- Thomas JJ (1992) *Informal economic activity*. Harvester/Weatsheaf, New York
- Torgler B, Schneider F (2009) The impact of tax morale and institutional quality on the shadow economy. *J Econ Psychol* 30(3):228–245
- Vuletin G (2008) Measuring the informal economy in Latin America and the Caribbean. International Monetary Fund, IMW working paper WP/08/102, Washington, DC

Shapley Value

André Casajus^{1,3} and Helfried Labrenz^{2,3}

¹HHL Leipzig Graduate School of Management, Leipzig, Germany

²Institut für Unternehmensrechnung, Finanzierung und Besteuerung, Wirtschaftswissenschaftliche Fakultät, Universität Leipzig, Leipzig, Germany

³LSI Leipziger Spieltheoretisches Institut, Leipzig, Germany

Definition

The Shapley value (Shapley 1953) probably is the most eminent (single-valued) solution concept for cooperative games with transferable utility (TU games)¹. A **(TU) game** is a pair (N, v) consisting of a nonempty and finite set of players N and a **coalition function** $v \in \mathbb{V}(N) := \{f : 2N \rightarrow \mathbb{R} | f(\emptyset) = 0\}$. Subsets of the player set are called **coalitions**. The coalition function assigns a **worth** $v(S)$ to any coalition S , which indicates what the players in S can achieve by cooperation. A (single-valued) **solution concept** φ on N assigns to any game (N, v) a payoff vector $\varphi(N, v) \in \mathbb{R}^N$, where $\varphi_i(N, v)$ denotes the payoff of player i in the game (N, v) .

The Shapley value on the player set N is defined as follows: A rank order on N is a bijection $\rho : N \rightarrow \{1, \dots, |N|\}$ with the interpretation that i is the $\rho(i)$ th player in ρ ; $R(N)$ denotes the set of all rank orders. The set of players up to and

¹Peleg and Sudhölter (2007) provide a good introduction to cooperative game theory.

including i in ρ is denoted by $P_i(\rho) = \{j \in N : \rho(j) \leq \rho(i)\}$. The marginal contribution of i in the game v in ρ is defined as $MC_i^v(\rho) := v(K_i(\sigma)) - v(K_i(\sigma) \setminus \{i\})$. Finally, the Shapley value, Sh , is the average marginal contribution over all rank orders, i.e.,

$$Sh_i(N, v) := \frac{1}{|R(N)|} \cdot \sum_{\rho \in R(N)} MC_i^v(\rho) \text{ for all } v \in \mathbb{V}(N) \text{ and } i \in N.$$

The marginal contribution $MC_i^v(\rho)$ measures the productivity of player i with respect to the coalition of his predecessors in ρ . By taking the average over all rank orders, the Shapley value can be viewed as an indicator of the players' productivity in a game.

Properties and Characterization

The Shapley value has a lot of properties worth mentioning, among them efficiency, symmetry, and strong monotonicity. A solution concept φ on N is called **efficient** if $\sum_{i \in N} \varphi_i(v) = v(N)$ for all $v \in \mathbb{V}(N)$. Efficiency indicates that all players from N first cooperate in order to create the worth of $v(N)$ and then this worth is distributed among them without losses (e.g., bargaining costs) or gains (e.g., subsidies). Players $i, j \in N$ are called symmetric in $v \in \mathbb{V}(N)$, if $v(S \cup \{i\}) - v(S) = v(S \cup \{j\}) - v(S)$ for all $S \subseteq N \setminus \{i, j\}$, i.e., symmetric players are equally productive. A solution concept φ on N is called **symmetric** if $\varphi_i(v) = \varphi_j(v)$ for all $v \in \mathbb{V}(N)$ and $i, j \in N$ such that i and j are symmetric in v , i.e., symmetry requires that players who are equally productive obtain the same payoffs. A solution concept φ on N is called **strongly monotonic** if $\varphi_i(v) \geq \varphi_i(w)$ for all $v, w \in \mathbb{V}(N)$ and $i \in N$ such that $v(S \cup \{i\}) - v(S) \geq w(S \cup \{i\}) - w(S)$ for all $S \subseteq N \setminus \{i\}$. Strong monotonicity requires that a player's payoff does not decrease when his productivity measured by marginal contributions does not decrease, irrespective of what happens to other players' productivity. This implies that a player's payoff only depends on his own productivity.

The Shapley value not only satisfies these three properties, but it is already characterized by them—the Shapley value is the unique solution concept that meets efficiency, symmetry, and strong monotonicity (Young 1985). This supports the view that the Shapley value is *the* measure of the players' productivities in a TU game.

Weighted Voting Games and the Shapley-Shubik (Power) Index

A game (N, v) is called **simple** if $v(S) \in \{0, 1\}$ for $S \subseteq N$; it is called **monotonic** if $v(S) \leq v(T)$ for all $S, T \subseteq N$ such that $S \subseteq T$. Important examples of simple and monotonic games are the so-called weighted voting games. A weighted voting game (N, v) is given by a quota $q > 0$ and weights of the players $w_i > 0, i \in N$ as follows. For all $S \subseteq N$, we have $v(S) = 1$ if $\sum_{i \in S} w_i > q$ and $v(S) = 0$ if $\sum_{i \in S} w_i \leq q$. The Shapley value applied to voting games is also known as the **Shapley-Shubik (power) index** (Shapley and Shubik 1954). For these games, the calculation of the Shapley value can be simplified: A coalition $S \subseteq N \setminus \{i\}$ is called a swing for player $i \in N$ in v if $v(S \cup \{i\}) = 1$ and $v(S) = 0$, i.e., if i turns S into a winning coalition. We then have

$$Sh_i(N, v) = \frac{1}{|N|!} \cdot \sum_{S \subseteq N \setminus \{i\} : S \text{ is a swing for } i} |S|! \cdot (|N| - |S| - 1)!,$$

where $|N|!$ is the number of all rank orders on N and $|S|! \cdot (|N| - |S| - 1)!$ is the number of all rank orders ρ in which player i "enters" coalition S , i.e., $P_i(\rho) \setminus \{i\} = S$.

Applications in Law and Economics

Most applications of the Shapley value in law and economics used are for cost allocation (see, e.g., Shubik 1962; Roth and Verrecchia 1979) or as a power index. For example, Newman (1981) analyzes the voting power of the Big Eight audit firms in the Accounting Principles Board (APB) and in

the Financial Accounting Standards Board (FASB). Leech (1988) and Casajus et al. (2009) use the Shapley-Shubik index in order to determine the voting power of shareholders in shareholders' meetings.

References

- Casajus A, Labrenz H, Hiller T (2009) Majority shareholder protection by variable qualified majority rules. *Eur J Law Econ* 28(1):9–18
- Leech D (1988) The relationship between shareholding concentration and shareholder voting power in British companies: a study of the application of power indices for simple games. *Manag Sci* 34(4):509–527
- Newman DP (1981) An investigation of the distribution of power in the APB and FASB. *J Account Res* 19(1):247–262
- Peleg B, Sudhölter P (2007) Introduction to the theory of cooperative games. In: *Theory and decision library C*, vol 34, 2nd edn. Springer, New York
- Roth AE, Verrecchia RE (1979) The Shapley value as applied to cost allocation: a reinterpretation. *J Account Res* 17(1):295–303
- Shapley LS (1953) A value for n -person games. In: Kuhn H, Tucker A (eds) *Contributions to the theory of games*, vol II. Princeton University Press, Princeton, pp 307–317
- Shapley LS, Shubik M (1954) A method for evaluating the distribution of power in a committee system. *Am Polit Sci Rev* 48:787–792
- Shubik M (1962) Incentives, decentralized control, the assignment of joint costs, and internal pricing. *Manag Sci* 8(3):325–343
- Young HP (1985) Monotonic solutions of cooperative games. *Int J Game Theory* 14:65–72

Spence's idea is that the education level of an applicant serves as a signal to indicate productivity level. The crucial assumption is that people with higher productivity incur lower costs (money, effort) of acquiring education than those with lower productivity. This leads to the fundamental insight that even if education did not contribute anything to work productivity, it can serve as a tool to separate high- from low-ability employees. In 2001, Michael Spence won the Nobel Prize in economic sciences, along with George A. Akerlof and Joseph E. Stiglitz, for their significant contribution on the analysis of markets with asymmetric information.

Definition

A signal is a choice that conveys information about the sender of the signal. When the communicating parties have different goals, the signal to be credible must be costly to fake. When selling a product, for example, the seller of a product can issue a warrantee to signal the quality to the buyer. Signalling takes place in situations of information asymmetry where one party has more information than the other party. In contract theory and economics, signalling models fall under the category of principal-agent models. Here, the “agent” has private information compared to the “principal.”

Signalling

Ines Lindner
Department of Econometrics and OR, VU
University Amsterdam, Amsterdam, The
Netherlands

Abstract

In his seminal paper “Spence (Q J Econ 87(3):355–374, 1973)” introduced a signalling model for the job market. In this model the employer (the principal) is not sure about the ability of the potential employee (the agent).

Signalling

The need for reliable signals results from the problems of communication between parties with different goals. The solicitor for a job, for example, has an incentive to overstate her ability. The buyer of insurance, likewise, has an incentive to misrepresent her risks. Similarly, the seller of a used car has an incentive to overstate the quality of her car, while the buyer is not keen to reveal her reservation price. This is fundamentally different to communication with common goals as, for example, the communication between bridge playing partners. Here, communication is simply

a transfer of information, and the receiver of the message does not need to question its credibility.

For a signal (message) to be credible between adversaries, the signal must be costly to fake. As an example consider a certificate from a massive open online course that was gained by taking an online test. Since taking the test is not fraud resistant and hence passing the test is easy to fake, a certificate gained in this way is usually not accepted as a credible signal for ability.

In contract theory and economics, the term information asymmetry refers to scenarios where one party holds information that the other party does not have, usually represented by principal-agent models. Here, the “agent” has private information in contrast to the “principal” who has information about general statistics. When applying for a job, for example, the applicant is the agent who knows his ability for the job, while the employer, the principle, lacks this information but can usually estimate the likelihood of a candidate being capable. Typically, the agent tries to send a signal about his private information, and the principle has to decide whether this signal is credible.

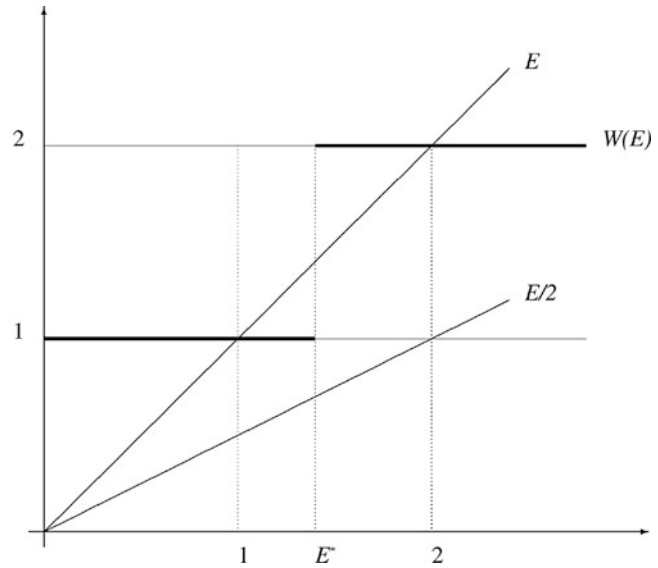
The formal tools to describe and analyze the strategic scenario of signalling are provided by game theory. The following example is taken from Spence (1973) and illustrates some basic game theoretic terminology by means of a job market model. This model, together with Akerlof’s (1970) model of The Market for Lemons, belongs to the classics of games with information asymmetry. The simplest version is as follows. There are two groups of employees, workers with low or high productivity, say, productivity level 1 (low) or 2 (high). An agent knows his productivity level and applies for a job. If the employer could assess his productivity *ex ante*, she would offer wage 1 in case it is low and 2 if it is high.

Agents in the job market can choose an education level E which is visible. There is a vast literature as to what extent education *causes* high productivity. The problem, in a nutshell, is that wage differences between employees with degrees from prestigious and nonprestigious institutions cannot be entirely attributed to the education itself.

After all, many prestigious universities screen their applicants such that they are already more productive to begin with. In order to keep the model simple, however, assume that education itself does not affect the productivity of the worker (for a more general approach, see Spence 1974, 2001). Assume that the cost of education is, *ceteris paribus*, higher for low-productivity workers, where costs can be money, time, and effort. In particular, assume the cost for low-ability worker of E years of education is E , whereas it is $E/2$ for high-ability workers. Figure 1 is reproduced from Spence (2001) where the lines with a positive slope represent the costs and the bold horizontal lines represent the wages. There are three kinds of equilibria.

1. *Separating equilibria.* The low- and high-ability workers choose different education levels such that the employer can identify the sender of this signal as either “low” or “high.” The high-ability workers choose the education level E^* with $1 < E^* < 2$. This implies a wage level of $2 - E^*/2 > 1$. The low-ability workers choose an education level of 0 with a net wage of 1. Note that the low-ability workers do not have an incentive to imitate the education level of the high-ability workers as their net wage would be $2 - E^* < 1$. In order for these choices to be part of a separating equilibrium, the strategy of the employer must be a best answer to the behavior of the employees. The employer observes the education level E and offers wage 1 if $E < E^*$ or wage 2 if $E \geq E^*$, reflecting her belief about the applicant’s productivity. This belief is confirmed by the behavior of the employees. Assuming $1 < E^* < 2$ implies that there is a continuum of separating equilibria, one for each value of E^* between 1 and 2.
2. *Pooling equilibria.* All workers choose the same education level; in particular consider the case where all workers choose an education level of zero. In the absence of a separating tool, the employer pays a weighted average wage. Pooling equilibria are sustained when the fraction of low-ability workers is sufficiently small such that the effect on average

Signalling, Fig. 1 Costs, wages, and the separating education level in the model of Spence (1973)



wage is not large enough to trigger expensive signalling for the high-ability workers. To be more precise, recall that sending the signal of education E^* is costly with a net wage $2 - E^*/2$. Assume α is the fraction of the low-productivity population. In the absence of a separating equilibrium, there is no way the employer could distinguish productivity ex ante. This implies she would be willing to contract for a wage reflecting the average productivity $\alpha \cdot 1 + (1 - \alpha) \cdot 2$ which simplifies to $2 - \alpha$. For a sufficiently small population of low-productivity workers, α , the average wage paid to the high-ability worker is therefore higher than getting wage 2 with signalling costs $E^*/2$. In particular, the inequality $2 - \alpha > 2 - E^*/2$ can be simplified to $\alpha < E^*/2$. Since E^* is bounded from below by 1, this implies that for $\alpha < 1/2$ there is a pooling equilibrium in which no employee invests in education and employers pay average productivity $2 - \alpha$ as wages. As noted by Spence (2001), this is a “better” equilibrium in the Pareto sense (everybody is better off) than the separating equilibrium. The reason is that education used merely for signalling purposes costs resources without enhancing productivity. Note that zero education is not the only pooling equilibrium. It turns out that there is an $E' > 0$ such that any education level

between 0 and E' can be sustained in a pooling equilibrium (for details, see Spence 1973).

3. *Hybrid equilibria.* In this class of equilibria, one or both types randomize. For example, it is possible to show that the following can arise as equilibrium in the job market model. Low types choose an education level of 0, while the high type randomizes between 0 and a positive education level. In equilibrium, the employer identifies a high-ability worker by observing a positive education level. When the education level is zero, however, her belief that this signal comes from a high-ability worker is a probability derived from updating $1 - \alpha$ following Bayes’ rule. It relates the frequency of observing zero education from a high-ability type to the overall frequency of observing zero education (for details, see, e.g., Fudenberg and Tirole (1991), Chapter 8).

Adverse Selection in Absence of a Separating Signal

Note that $2 - \alpha$ is a very attractive wage for the low-ability worker whose productivity is in fact only 1. It is, however, less attractive to the high-ability worker such that they could have an incentive to leave the labor market (and become, e.g., entrepreneur). This would cause α to increase

such that average wages $2 - \alpha$ would drop even further. This drop, in turn, makes staying in the labor market even less attractive for high-ability workers. In the end, all workers left in the labor market may be of low ability. This phenomenon is usually referred to as the adverse selection problem: a process which attracts the less desirable potential trading partners. Adverse selection is especially relevant to insurance markets. Here, the private information is the risk profile of the policy owner (agent) which is relevant for the insurance premium. Offering insurance without deductible based on average-risk attitude will be unattractive to the low-risk type and attractive to the high-risk type. Consequently, a premium based on average risk will attract more high-risk types leading to a higher average risk and therefore an increase of the insurance premium. This in turn repels the low-risk profiles even more. In the end, only high-risk types may participate in the insurance market. In his paper "The Market for Lemons," Akerlof (1970) used the logic behind this downward spiral to explain why almost new cars sell for so much less than brand new cars. A used car in a bad state is commonly referred to as "lemon" which is private information of the seller, the agent. Buyers cannot distinguish the quality of a used car and are therefore only willing to pay a price representing average quality of the used car market. This implies that owners of good cars have a stronger incentive to keep them which implies a decrease in the average quality of cars for sale. This decrease, in turn, lowers the reservation price of the of sellers - which kicks off the downward spiral of adverse selection processes. In conclusion, Akerlof's Lemons model explains why secondhand cars are, on average, of lower quality than cars not for sale. An almost new car for sale is therefore associated with the large pool of lemons and sells only for the average price of this pool.

Signalling versus Screening and Signals versus Indices

A screening game is closely related to a signalling game. Here, the principal (the uninformed player)

chooses a strategy that induces the agent (informed player) to reveal their type. For example, insurance companies screen customers by offering different policies at different rates. High-risk types will be less likely to favor an insurance contract with deductibles (Stiglitz and Weiss 1990) and thereby revealing their type.

It is helpful to carefully distinguish between indices and signals. According to Spence (1973, 1974, 2001), "indices" are unchangeable attributes over which an individual has no control, e.g., gender, age, race, etc. A signal, however, is a visible attempt to communicate that can be changed by a person. As an example take the Irish car insurance market where males less than 25 years pay more than females of the same age. Here, age and gender are indices that are used for statistical discrimination, i.e., the practice of the insurance company to judge the risk based on gender and age. Favoring insurance contracts with deductibles, however, is a signal of the low-risk type.

The Effect of Legal Rules on Signalling

Legal rules can sometimes help to prevent inefficient equilibria in signalling games. Baird et al (1998, p. 143) offer the example of parental leave. Before signing the contract, potential employees might refrain from bargaining for parental leave in order to avoid disadvantageous inferences that the employer could draw. For example, the employer could think that employees with children are less flexible with respect to their working time. Here, workers who are likely to become parents will be tempted to imitate the behavior of those unlikely to become parents and will not ask for a clause on leave in their work contract. Demanding by law that this clause is part of all work contracts solves this problem; however, it also leads to lower average wages as the employers will internalize the extra costs. The outcome under legal regulation is efficient if the total benefit of the workers likely to become parents outweighs the total loss of the workers unlikely to become parents.

Regulation by law could also prevent informative signals. When selling a product, for example, the seller of a product can issue a warrantee to signal the quality to the buyer. The length of the warrantee signals the quality level of the product. A legal rule that mandates a warrantee would suppress this signal and might prevent that different qualities are sold at different prices. The question remains why the producer of a low-quality product has an interest to disclose this fact by offering only limited warrantee, say, only 2 months on an electronic device. To see the logic behind this principle, assume a situation in which all producers with high-quality products disclose this fact by offering a warrantee of at least 3 months while other producers remain silent, i.e., offer no warrantee at all. Without a warrantee consumers will estimate the quality of the product by the average of the unguaranteed products, say, by 1.5 months. This underestimates a product that could be sold profitably with a 2-month warrantee, which explains why the producer has an incentive to disclose this fact. By repeatedly applying this argument, any quality will eventually be disclosed. In economics, this principle is known as the full-disclosure principle (Frank 2010). It says that individuals have an incentive to disclose even unfavorable information; otherwise, their silence is taken to mean that they have something even more unfavorable to hide. The full-disclosure principle also explains why regulations on revealing information are sometimes difficult to be enforced in practice. Consider, for example, questions about marital status, children, or sexual preference which are usually prohibited by law. A job candidate, however, might have an interest to voluntarily share this information if he thinks that this speaks in his favor which leads to the unraveling process of the full-disclosure principle.

More Examples of Signalling in Law and Economics

1. **Product quality assurance.** The quality of many products is not directly verifiable by consumers, especially if a product is new and

consumer ratings are not yet known. The question is how a firm launching a new product can signal its high quality to potential customers. Besides developing a reputation, a firm can signal high quality by extensive sunk costs, i.e., extensive investments that cannot be liquidated like heavy advertising (Klein and Leffler 1981). Such a firm has a high interest to stay in business and have returning customers. This implies that consumers can take substantial sunk costs of a firm as a signal that is too costly to fake when the product is of low quality.

2. **Social norms.** Posner (2000) uses signalling to explain the rise and preservation of social norms (for secondary literature, see McAdams 2001). In line with rational choice scholars, he identifies individual's interest in cooperative opportunities as the main source of norms. The familiar problem in social or economic interaction is that individuals fail to reap all possible gains from cooperation, because defecting from cooperative behavior is a "dominant" strategy in the following sense: it provides a higher individual payoff regardless what others do. This is commonly illustrated by the prisoner's dilemma which is a classic in game theory (see, e.g., von Neumann and Morgenstern 2007). The scenario changes if the interaction occurs more than once, i.e., if the game is played repeatedly. Here, cooperative behavior can thrive because it can induce the other side to cooperate in future rounds while defective behavior may only be punished by defection in the next rounds. That cooperative behavior can arise in repeated games is illustrated by the famous computer tournament conducted by Axelrod (1981, 1984) who elicited strategies from game theorists to compete in a setup of the repeated prisoner's dilemma game. These insights are made precise in the *folk theorems* (for an overview of this large literature, see Fudenberg and Tirole 1991).

On the other hand, if the individual cares much more about immediate benefit than the discounted future payoff, the immediate benefit from defecting will exceed the prospective

gains from future cooperation. In conclusion, the more an individual discounts the future, the more likely she is to choose the one-time benefit from defection and forego future cooperation. In Posner's model an individual is characterized by his or her discount rate which separates individuals into "good" and "bad" types for cooperative relationships. People are concerned with finding and attracting cooperative partners, which, according to Posner, boils down to finding and attracting "good" partners with low discount rates. The good ones are looking for good types in order to reap the repeated gains from cooperation. The bad ones are also trying to attract good ones, but only to profit from the one-time benefit of defecting behavior when the good type cooperates. As a result, the more convincingly a person can present herself as being of the good type, the more cooperative opportunities will be offered to her. People try to "signal" a low discount rate, true or not, by making certain observable choices. As one of the many examples, Posner mentions gift giving. A good type is, other things being equal, more likely to incur a cost now for the sake of future benefits from cooperation. The following example is taken from Posner (2000). Assume that a good type values the future payoff 10 at 10% discount, while the bad type values the same future payoff at 30% discount. This leaves the value 9 for the good type and 7 for the bad type. This means, however, that the good type can distinguish herself by buying a gift that costs 8. The bad type will not do so, because the costs (8) are higher than the prospective benefit (7). This represents a separating equilibrium. A sufficiently expensive gift can therefore be interpreted as the giver trying to signal a low discount rate.

The important point in the gift example is that the costs can be observed by the receiver. Analogously, any behavior that incurs costs can be interpreted as a signal. Posner suggests that in general any conformity move like showing manners or following a fashion trend can be read as an attempt to signal low discount rates. Good types are more likely to make this

investment than bad types. The latter have an incentive to imitate the behavior of good types in order to avoid being identified as bad types. In this sense, Posner suggests that any social norm is the outcome of a signalling game with either pooling or separating equilibria.

3. **Mediation as signal.** When two or more parties are involved in dispute, there is often some resentment in turning to mediation as an alternative dispute resolution. The main reason is that, although mediation officially has to remain strictly confidential, it is difficult to prevent that the information disclosed during a mediation process will be exploited by the other party in court in case mediation fails. Holler and Lindner (2004) argue that this is not the only informational issue at stake. They claim that the call for mediation and the reaction to that call is a signal in the sense of Spence (1973), i.e., it possibly gives away information about the sender.

Game Theoretic Classification of Signalling Models

From a game theoretic viewpoint, a scenario with asymmetric information is a game with private information. In the job market example, the private information is the ability of the employee and hence the individual cost of education. In game theoretic terms, the cost or payoff (utility) of the employee is not known to the employer. Private information can also be information about the condition of a state, say, the knowledge of a market survey that other players do not have. A third possibility is the existence of alternative strategies of a player that other players are not aware of. Games with private information are referred to as games with incomplete information. In order to analyze these games, they are usually reframed with an initial choice of "nature" (Harsanyi 1967, 1968a, b). In the job market example, for example, nature chooses randomly whether the employee is of low- or high-ability type. The probability distribution of this random choice is known to all players (the parameter α in the job market example). In game theoretic terms, a type of a player indicates possible characteristics that a player can have in the view of

all players. This type is characterized by his strategy set, his possible state of information and his payoffs. With the initial random choice by nature all possible private information is included in the model and it is no longer a game with incomplete information. Instead, it is a game with imperfect information, i.e., a game in which not all players are informed about the events that have previously occurred. In the Spence model, the employer is not informed about the choice of nature. Games that have been translated from a game with incomplete information to a game with imperfect information are also referred to as Bayesian games.

Signalling games are a special class of dynamic, Bayesian games with two players, the sender of signals and the receiver. The general formal model can be found in Cho and Kreps (1987). First, player A learns about his type t drawn from a finite set of possible types T , drawn following a probability distribution π which is known by all players. With this private information of his own type, the sender can choose to send a message m from a possible set of message $M(t)$. Player B observes this message and can choose her response from a finite set of responses $R(m)$, depending on the message. The game ends with this response, and the payoff to both players are given by a utility (payoff) function denoted by $u(t, m, r)$ for player A and $v(t, m, r)$.

The common solution concept in signalling games is that of a sequential equilibrium (Kreps and Wilson 1982) which represents a refinement of the Nash equilibrium for dynamic games. A sequential equilibrium specifies not only strategies but also beliefs of the players which justify their strategy. A belief is a probability distribution on scenarios which the uninformed player cannot distinguish. In the separating equilibrium of the job market example, a belief of the employer is that she assigns probability 1 to high ability upon observing an education of at least E^* . In the pooling equilibrium, her belief is $(1 - \alpha)$ that the candidate is of high ability. A combination of strategies and beliefs represents a Bayesian equilibrium if

1. The strategies are best answers in the sense that no player can improve his situation, given the strategies of the others (Nash equilibrium).

2. The beliefs are confirmed on the outcome path of the game using Bayes' rule.

The additional requirement of a sequential equilibrium is that (1) and (2) also holds under small perturbations when the players are forced to choose every strategy with a small probability. This extra condition ensures that the beliefs are also consistent off the equilibrium path.

The concept of a sequential equilibrium is very similar to the (trembling-hand) perfect equilibrium (Selten 1975; Baird et al. 1998). The latter concept also uses perturbations in order to force players to play every strategy with positive probability. Generically (almost always), both concepts provide the same results; however, the conditions of a perfect equilibrium are stricter such that in some special cases sequential equilibria are not perfect. In a nutshell, the difference is that perfectness requires optimal behavior for any nonzero perturbation, whereas the sequential concept only requires optimality in the limit of zero perturbation.

The motivation of refinement concepts of Nash equilibria like sequential or perfect equilibria is to cut out "unreasonable" equilibria; however, in signalling games, this usually still gives rise to a plethora of equilibria. The reason is that the sequential or perfect criterion is not restrictive enough on the possible beliefs off the equilibrium path. In-Koo Cho and David M. Kreps (1987) introduce the so-called intuitive criterion which further restricts out-of-equilibrium beliefs, thereby eliminating unintuitive equilibria.

References

- Akerlof GA (1970) The market for "Lemons": quality uncertainty and the market mechanism. *Q J Econ* 84(3):488–500
- Axelrod R (1981) The evolution of cooperation. *Science* 211(4489):1390–1396
- Axelrod R (1984) The evolution of cooperation. Basic Books, New York
- Baird DG, Gertner R, Picker R (1998) Game theory and the law. Harvard University Press, Cambridge
- Cho I-K, Kreps DM (1987) Signalling games and stable equilibria. *Q J Econ* 102(2):179–221
- Frank R (2010) Microeconomics and behavior. McGraw-Hill, New York

- Fudenberg D, Tirole J (1991) Game theory. MIT Press, Cambridge
- Harsanyi J (1967) Games with incomplete information played by bayesian players. Part I: the basic model. *Manage Sci* 14(3):159–182
- Harsanyi J (1968a) Games with incomplete information played by bayesian players. Part II: bayesian equilibrium points. *Manage Sci* 14(5):320–334
- Harsanyi J (1968b) Games with incomplete information played by bayesian players. Part III: the basic probability distribution of the game. *Manage Sci* 14(7):486–502
- Holler MJ, Lindner I (2004) Mediation as signal. *Eur J Law Econ* 17:165–173
- Klein B, Leffler K (1981) The role of market forces in assuring contractual performance. *J Pol Econ* 89(4): 615–641
- Kreps DM, Wilson R (1982) Sequential equilibria. *Econometrica* 50(4):863–894
- McAdams RH (2001) Signaling discount rates: law, norms, and economic methodology. *The Yale Law J* 110(4): 625–689
- Posner EA (2000) Law and social norms. Harvard University Press, Cambridge
- Selten R (1975) Re-examination of the perfectness concept for equilibrium points in extensive games. *Int J Game Theory* 4:25–55
- Spence AM (1973) Job market signaling. *Q J Econ* 87(3):355–374
- Spence AM (1974) Market signaling. Harvard University Press, Cambridge
- Spence AM (2001) Signaling in retrospect and the informational structure of markets, Nobel prize lecture. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2001/spencelecture.pdf
- Stiglitz J, Weiss A (1990) Sorting out the differences between screening and signalling models, NBER Technical working paper, No. 93. In Bacharach MOL, Dempster MAH, Enos JL Mathematical models in economics: Oxford mathematical economics seminar, Oxford University Press, Oxford
- von Neumann J, Morgenstern O (2007) Theory of games and economic behavior. Princeton University Press, Princeton

Signing Without Reading

Gerrit De Geest

Washington University in St. Louis, St. Louis,
MO, USA

Abstract

Most people sign standard term contracts without reading them. This gives drafters an incentive to insert one-sided, inefficient terms. This

problem can be solved directly by giving the drafter a duty to draft efficient terms or indirectly by giving the signer a duty to read (which may remove the incentive to insert one-sided terms if a sufficient number of signers do read the contract). The problem can also be solved in a more draconian way by holding all standard terms unenforceable, irrespective of whether they are efficient or one-sided (as proposed by Radin, Margaret Jane, *Boilerplate: the fine print, vanishing rights, and the rule of law*. Princeton University Press, Princeton, 2013). Finally, the problem can be solved through hybrid instruments – for instance, American law gives the signer a duty to read but intervenes when terms are unconscionable.

In this short chapter, written for Springer's forthcoming *Encyclopedia of Law and Economics* (J. Backhaus, ed.), I argue that a duty to draft efficient terms is the superior instrument. Doctrinally, this means that unread contracts are best seen as agreements to delegate the drafting task to the party that can do so at least costs, as is the case under German law. Because rational parties will never give one of them a wildcard to insert inefficient terms, standard terms should be enforced only to the extent they are efficient. Moreover, economic logic dictates that it should be upon the drafter to prove that the terms are efficient, rather than upon the signer to prove that they are inefficient.

Introduction

Contract law treatises usually assume that parties read contracts before signing them. In the real world, very few people do so. The overwhelming majority of all signed contracts on this planet are standard term contracts, and the overwhelming majority of standard term contracts are never read. This is confirmed by recent empirical studies, which find that very few people even bother to take a look at the contracts they sign, and even fewer spend enough time to understand them (e.g., Bakos et al. 2014; see also Ben-Shahar and Schneider 2014).

Not reading contracts is problematic because it gives self-interested drafters an incentive to insert one-sided, inefficient terms (as empirically confirmed by, e.g., Marotta-Wurgler 2009). For instance, drafters may insert terms that exempt them from any liability for defective products, which may lead to suboptimal incentives to deliver quality. They may also insert terms that redistribute wealth from the signer to the drafter, for instance, through hidden fees. While such terms may not seem inefficient at first glance because they do not destroy wealth but apparently only redistribute it between the parties (for instance, they create a benefit of +3 to the drafter and a cost of -3 to the signer), they may indirectly lead to inefficiency, for instance, by giving the drafter an incentive to induce breach, by discouraging the signer from breaching efficiently, or by discouraging the signer from entering into contracts in which there may be such fees in the first place.

In the absence of any legal intervention, the incentive to insert one-sided terms may be very strong. To illustrate, suppose someone orders a book at a local bookshop. The unread standard terms of the signed book order state that the book has to be picked up within a week after arrival. The buyer picks it up after 2 weeks, thus breaching this term. The unread standard terms contain a \$1 million penalty clause for any breach of contract. If the legal system enforced such terms, a contract drafter could make a million dollars simply by trapping a single consumer.

To be fair, a part of the drafter's benefit from one-sided terms may be passed on to the signer in the form of lower overall prices. Indeed, if drafters can earn rents through inserting one-sided terms, they may want to lower prices in order to attract buyers from whom they can occasionally profit. To what extent price competition will completely eliminate these rents (turning them into quasi-zents) depends on the drafter's market power. In a perfectly competitive market, all of these profits will be passed on to buyers; the result is a pure harsh-term low-price equilibrium (Goldberg, 1974). But competition is rarely perfect. Moreover, the fact that the profits are passed to the "victim" does not make the economic problem

disappear. Harsh-term low-price markets are problematic from an economic point of view for the simple reason that harsh terms are inefficient.

The problem of one-sided, inefficient terms can be solved in four ways. First, it can be solved directly by giving the drafter a duty to draft efficient terms. This is the case under German law, where the signer is presumed to have given the drafter the authority to draft reasonable terms only (see Maxeiner 2003). Second, the problem can be solved indirectly by giving the signer a duty to read; the idea is that if a sufficient number of signers read the contract and walk away when they discover one-sided terms, the drafter may no longer benefit from inserting such terms. Third, the problem can be solved in a more draconian way by holding all standard terms unenforceable, irrespective of whether they are efficient or one sided. Fourth, the problem can be solved through hybrid instruments that combine some of the previous instruments. For instance, American law gives the signer a duty to read but intervenes when the drafter has inserted unconscionable terms.

At the outset, I need to define some terms. First, I will call the two parties the "drafter" and the "signer," respectively. Strictly speaking, the drafter has to sign the document as well; therefore, the term "drafter" refers to the party that drafts and signs the contract, while the "signer" refers to the one that only signs. Second, I will use the term "standard term contract" to denote contract terms that are presented in a take-it-or-leave-it form to more than one buyer. (Note that in some of the literature, the term "adhesion contract" or the more colorful but less rhetorically neutral term "boilerplate" is used instead.) Third, I will distinguish between "efficient terms" and "one-sided terms." "Efficient terms" are given a standard definition: they are terms that increase the contractual surplus by creating obligations that generate a higher benefit to the promisee than they cost to the promisor. "One-sided terms" are given a narrow definition: they are terms that are not only pro-drafter but also inefficient in the sense that they diminish contractual surplus by costing the promisor more than they benefit the promisee. The economic question is indeed not whether the

drafter inserts terms that are to her benefit; the question is whether these terms are efficient. For instance, if the drafter inserts a term that generates a benefit of 3 to her and a cost of 1 to the signer, the drafter does something desirable from an economic viewpoint. But if the drafter inserts a term with a benefit to her of 1 and a cost to the signer of 3, there is an economic problem. Therefore, our focus is on those pro-drafter terms that are also inefficient.

Duty to Draft Efficient Terms (Solution 1) Versus Duty to Read (Solution 2)

Should the drafter be given a duty to draft efficient terms or the signer a duty to read? There are many ways to frame the problem, but they all lead to the same conclusion: a duty to draft efficient terms is a better instrument than a duty to read.

Unilateral or bilateral care problem. Should the drafter abstain from inserting inefficient terms, or should the signer read the contract, or should both do their share of the work to solve the problem? In other words, is this a unilateral care problem (in which only one party has to exert “effort” – the drafter in terms of abstaining from a profitable activity or the signer in terms of spending time to read) or a bilateral care problem (in which both parties should exert effort)? Those who defend a duty to read (a category that includes most American courts) seem to assume (at least implicitly) there is a bilateral care problem. For instance, the click-wrap rule (which states that standard terms should be easy to find rather than buried several clicks away on the website, as is the case with browse-wrap) and the plain English rule (which states that standard terms should be easy to read) seem to be based on the implicit belief that the signing-without-reading problem should be solved partly through more reading.

A more careful analysis, however, tells us that it is a unilateral care problem – though the confusion may stem from the fact that it is a bilateral care problem in individually negotiated contracts.

Reading contracts has a social cost – but what exactly is its social benefit? In individually

negotiated contracts, parties need to exchange information about their idiosyncrasies in order to discover the terms that create the most surplus. Reading contract drafts, under those circumstances, has a clear social benefit – it leads to better, more individualized contracts. Signing without reading in this case is not cooperating to this process and possibly even miscommunicating intentions.

In standard term contracts, however, there is no need to communicate individual characteristics. Standard term contracts are not adapted to individual characteristics – they are offered on a take-it-or-leave-it basis. The social benefit of reading standard contracts is therefore only to discourage drafters from inserting one-sided terms. So in standard term contracts, reading is essentially monitoring the drafter.

The problem can be analogized to theft. Should theft be prevented by giving thieves a direct incentive not to steal (by punishing them when they are caught stealing) or by giving potential victims an incentive to guard their property so that stealing attempts fail? The first-best is obviously the former. If the incentive to steal can be removed, there is no need for victims to guard their property. This differs from typical bilateral (joint) care models in the economic analysis of tort law, in which the first-best is that both parties exert some effort.

To be fair, monitoring someone else (as when potential victims monitor potential thieves) may be a second-best solution when the legal system cannot fully eliminate the first incentive problem. In a society in which theft cannot be completely eliminated because it is too costly for the police to catch a sufficient number of thieves to make theft always unprofitable, it can be desirable, as a second-best solution, to let potential victims invest in surveillance cameras and locks. But in a contractual setting, apprehension rates are naturally high – often even 100 %. If a good is stolen, we may not know who did it, but if a one-sided term is inserted into a contract, we know who did it – the drafter. Moreover, the “wrongdoer” may reveal herself to the legal system by going to court to enforce a one-sided term. In addition, it is sufficient that a single, savvy “victim” discovers the one-sided nature of a term to trigger a

collective enforcement procedure, such as a class action or a suit initiated by an agency or by a state attorney general. Therefore, it is not clear why the legal system should not try to achieve the first-best solution when it comes to the signing-without-reading problem. In addition, if the duty to read were abolished, there would still be a right to read, just like there is a right to place surveillance cameras; so second-best solutions would still be possible.

Permit ex ante monitoring only or permit both ex ante and ex post monitoring? Another way to frame the problem is to analyze it in terms of the well-known distinction between ex ante and ex post monitoring. Ex ante monitoring takes place before the “wrongdoing” may occur; ex post monitoring takes place after the “wrongdoing” occurred. For instance, vehicle safety inspections or speed controls take place before the car driver causes an accident. A tort action is brought after an accident has already happened. Similarly, an employer can prevent theft by its employees by checking them every day before they go home (ex ante) or by starting an investigation after an irregularity has been discovered (ex post).

Reading a contract before signing it is a form of ex ante monitoring. The drafter’s work is inspected before the contract is signed, that is, before the “wrongdoing” takes place. Reading a contract after a dispute has arisen (and refusing to honor one-sided terms) is a form of ex post monitoring. The drafter’s work is inspected after the “wrongdoing” has taken place – that is, after one-sided terms have been inserted into an (apparently binding) contract. The economic literature on ex ante versus ex post monitoring has identified a number of advantages of each (e.g., Shavell 1984). Ex post monitoring has the advantage that it needs to occur less often, that is, only in those cases in which something went wrong (e.g., Shavell 2013, in the context of regulation versus tort law). Ex ante monitoring may be desirable in other cases, for instance, when the harm is irreparable or the wrongdoer is judgment proof. The bottom line is that both ex ante and ex post monitoring can make sense under specific conditions, so that there seems to be no reason to forbid one of these forms of monitoring.

A pure duty-to-read rule permits only ex ante monitoring: if a one-sided term is discovered after the contract has been signed, it is too late. A pure duty to draft efficient terms permits ex post monitoring: if a one-sided term is discovered after the contract has been signed, this term will not be enforced. But a pure duty to draft efficient terms does not forbid ex ante monitoring: nothing prevents a party from reading the contract before signing it and walking away when one-sided terms have been spotted. In this sense, a duty to read permits only ex ante monitoring, while a duty to draft efficient terms permits both ex ante and ex post monitoring. The question can therefore be framed as follows: should ex post monitoring of contract terms be made impossible by the legal system? Framed this way, it is once again obvious that a duty to draft efficient terms is the superior instrument.

Should the legal system correct wrongdoing or should potential victims solely rely on self-help? A third way of framing the problem is to view a pure duty-to-read regime as an instance in which the legal system refuses to intervene when it observes socially undesirable behavior (i.e., the drafter inserting inefficient terms into a contract) but instead requires potential victims to protect themselves by walking away before the socially undesirable behavior takes place. It is not clear, however, why the legal system should tolerate socially undesirable behavior in this case – and even assist the “wrongdoer” by enforcing one-sided terms. Contract law is full of rules that attack undesirable behavior (such as overbreaching, acting opportunistically, misleading, or neglecting to mitigate). Similarly, corporate law (a specific type of contract law) is largely an attempt to prevent self-dealing in one form or another. More generally, modern legal systems consist of millions of rules, and each of these rules, at its most basic level, is designed to prevent some form of socially undesirable behavior. Why should inserting one-sided, inefficient terms receive immunity?

A duty to read is unlikely to be effective. A duty to read tries to solve the drafter’s incentive problem in an indirect way; if a sufficient number of people read the contract, the benefit of inserting

one-sided terms may disappear. The question is what percentage of potential buyers must read the terms before the mechanism becomes effective.

The question can be reframed as follows: for how many buyers is it *individually rational* to read contracts? Reading contracts costs time; understanding contracts costs even more time (Bar-Gill and Warren 2008; Ben-Shahar and Schneider 2014). Apple's iTunes licensing contract, for instance, is 55 pages long. Even a lawyer specialized in licensing contracts may need days to fully understand the significance of all the terms.

In individually negotiated contracts, the incentives to read a contract are stronger than in standard contracts because the reader is likely to internalize a significant share of the surplus created by the better terms. In standard term contracts, there is a gap between what is optimal for all buyers as a whole and what is individually rational for an individual buyer. Preventing the drafter from inserting inefficient terms is a public good for all buyers. Like all public goods, reading terms is associated with collective action problems that are likely to lead to the underprovision of the good (Katz 1990; Bar-Gill and Warren 2008).

As a matter of fact, an equilibrium in which all buyers read the contract can never be individually rational (Katz 1990). Indeed, the only reason to read a contract is to check if it contains one-sided terms. If all buyers read the contract, there will be no one-sided terms. But in that case, individual buyers no longer have an incentive to read the contract. This implies that, in equilibrium, there should be at least some contracts with one-sided terms on the market – otherwise the incentive to read and the equilibrium itself would disappear. Therefore, a duty to read can never *fully* solve the problem of one-sided terms.

What percentage of buyers should read the contract before the self-interested drafter stops inserting one-sided terms? Early economic commentators drew a rosy picture by arguing that, under some conditions, a small number of informed buyers would be enough (Schwartz and Wilde 1979; Trebilcock and Dewees 1981). While this may be true in some cases, it definitely

is not always the case; usually, the required percentage is much higher, and in some cases, the percentage may even be 100 % of the buyers (Gazal-Ayal 2007). In essence, a drafter compares the foregone profits from informed buyers with the increased profits from uninformed buyers. Depending on the magnitude of the parameters, stopping the problem may require few, many, or all reading the contract.

The results are even less rosy when drafters are able to split the informed and non-informed parties (Bar-Gill and Warren 2008). They may do so by offering the informed parties the efficient terms and the non-informed parties the one-sided terms. When such contract term discrimination is possible, one-sided terms will be offered as long as there are some non-informed parties.

Another complication is that, when too few buyers read contracts, all contracts on the marketplace may contain the same one-sided terms. This further undermines the incentive to read the contract.

In sum, a duty to read only prevents one-sided terms if a sufficient number of potential signers read the contract. But in reality, the required quorum is unlikely to be reached because of collective action problems, so that a duty to read is simply ineffective. This is another argument to conclude that a duty to draft efficient terms is the superior instrument.

A duty to read is wasteful when there are many readers. Even in the rare instances in which a duty to read would be effective, it would still be wasteful. Many consumers reading the same contract are a duplicative effort. Information is a nonrival good (that is, a good that, once produced, can be used by an infinite amount of people without being used up). Nonrival goods only have to be produced once. So if all one million consumers of a certain product could meet in a hypothetical world, they would not decide that – say – 500,000 of them would read the contract. Instead, they would pay a professional lawyer to read it for them. From this perspective, rules that encourage reading (such as disclosure rules, the plain English rule, the click-wrap rule, Kim's (2013) proposal for a "duty to draft reasonably," where "reasonably" refers to the communication process not to

the efficiency of the terms themselves; and Ayres and Schwartz's (2014) proposal to make firms put their most surprising terms in a standardized warning box) encourage wasteful behavior.

Doctrinal interpretation – delegation of the drafting task to one of the parties. From a doctrinal perspective, the question is what should be the status of signed terms that have not been read. One view is that these terms are a part of the contract, because a contract is any piece of paper that has been signed by two parties. This view conflicts with the modern view (and “modern” is defined here as the view that is dominant since late medieval times) that holds that a contract is in essence a meeting of minds. In this modern view, signed documents have only an evidentiary function; they help courts to find out what was the meeting of minds. But if courts know that these documents have not been read by one of the parties, then the documents have little evidentiary value with respect to finding the meeting of minds. So in this modern view, what is the status of signed but unread terms?

Under German law, signing without reading is seen as an act of delegation: the task of drafting terms is delegated to one of the parties. Unread terms are binding because the signer agreed to be bound by the terms drafted by the other party (Karl Llewellyn's “blanket assent” theory is a variant of this). But the drafter is presumed to have received only the authority to draft “reasonable” terms. Therefore, standard terms that are unreasonable are not binding; the signer never gave permission to draft such terms.

This view makes economic sense. Rational parties know that default rules are not always efficient (or not always adapted to their idiosyncrasies). Drafting efficient terms is costly and requires expertise. Therefore, they delegate this task to one of them. However, they will never agree to give the drafter a wild card to insert one-sided, inefficient terms because this would violate one of the fundamental principles of rational contracting – that only efficient terms are inserted into contracts.

Rational parties will not even allow one-sided, inefficient terms when the signer would be fully compensated for these terms through a lower price. Indeed, inefficient terms decrease the

surplus, and therefore, it is not possible to find a price adjustment that will make both parties better off. To illustrate, if a one-sided term generates a benefit of 1 for the seller but a cost of 3 to the buyer, the buyer would only agree to such a term if the overall price was decreased by at least 3. But the seller would never forego 3 to obtain a benefit of 1. Therefore, the argument that consumers may not be willing to pay for less one-sided terms (made, e.g., in Ben-Shahar 2013) must assume either that these one-sided terms are efficient or that consumers are uninformed or irrational.

Solution 3: Standard Term Contracts Are Never Binding

This third, more draconian solution has been proposed by Slawson (1971) and recently by Radin (2013). Under this theory, standard terms are not part of the contract because they have not been agreed to. Radin (2013) has argued that inserting egregious, rights-deleting terms should even become a tort.

Extreme as it may sound, this third solution follows in a sense logically from the previously mentioned meeting of minds (or “subjective”) theory. This theory, which is not a minority view but the dominant contract doctrine since late medieval times, holds that a contract is not a ritual nor a piece of paper but two human beings agreeing to be bound by the same terms, that is, having a “meeting of minds.” The document that is signed is thus not the contract but just evidence of the contract. But if courts know that one of the parties signed without reading, the piece of paper loses that evidentiary function. As a consequence, it is not binding – the terms were never agreed to.

From an economic viewpoint, however, this solution throws the baby away with the bath water because it invalidates all unread terms, even the efficient ones. Invalidating all standard terms in practice means that the legal default rules are always binding – as a matter of fact, all default rules then become mandatory rules in contracts that are not individually negotiated. This would be fine if default rules were always efficient and

perfectly adapted to individual situations. But it is hard to see why this would be the case.

Another way of framing the issue is as follows. The drafter has to do a job (draft the contract) and can do a good job or a bad job in this respect. Under the third solution (never enforcing standard terms), the drafter's work is always thrown away, not only when she did a bad job but also when she did a good job. But why would rational parties want to do that? Why would rational parties want to get rid of efficient terms and replace them with less efficient default rules?

Such an extreme remedy is only justifiable when (1) courts can never determine at all whether terms are efficient or inefficient and (2) courts know that standard terms are on average more inefficient than default rules. But if condition (1) is fulfilled, how could condition (2) be fulfilled – that courts know that their own gap-filling rules are better than the invalidated terms they replace? Indeed, if courts (or legislators) have no clue which terms are efficient, they will choose no better terms than those who simply flip a coin. Drafters who chose only pro-business terms may also get it right in 50 % of the cases (at least if the set of efficient terms contains 50 % pro-business and 50 % pro-consumer terms; that is not necessarily the case, but it may be the most rational assumption in the absence of more specific empirical information). Also, courts can read scholarship (or consult experts), and therefore, condition (1) comes down to the condition that scholars can never figure out which *types* of terms are efficient. This is not very plausible – even those who are skeptical about the practical value of scholarship tend to believe that scholars can at least get some things right. Moreover, if scholars could never figure out whether terms are efficient, how could courts once again know that condition (2) is satisfied that standard terms are on average more inefficient than default rules?

Solution 4: A Duty to Read with an Unconscionability Exception

This is the current solution in American common law. Unread contract terms are legally binding,

unless they are unconscionable. “Unconscionability” is a notoriously vague concept that typically requires that terms be clearly unreasonable, “egregious,” or, as one judge famously put it, “shock the consciousness.” (At least, this is the definition of substantive unconscionability; most American states also require procedural unconscionability, but this condition is easily met in practice for adhesion contracts). Although translating such vague criteria into precise economic jargon is always speculative, the criterion seems to be that terms are only invalid when they are clearly inefficient.

Still, at the end of the day, the question can be asked whether the difference between solution 1 (the German regime) and solution 4 (the American regime) is not chiefly a matter of evidentiary burdens. Indeed, solution 1 (a duty to draft efficient terms) holds that terms are binding *only if* they are efficient. Solution 4 (a duty to read terms except for when they are unconscionable) holds that terms are binding *unless* they are clearly inefficient. And this raises the next question: who should bear the burden of proof of (in)efficiency?

There are two main considerations here – the relative evidence cost to the parties and the likelihood of the hypothesis; see, e.g., Hay and Spier (1997) – and they both lead to the conclusion that the drafter should have the burden of proof. First, the drafter has more information on why terms are efficient or inefficient. After all, the drafter is the expert, who also has economies of scale, and drafting is at its core determining which terms are efficient. Accordingly, when the drafter drafted in good faith, she only deviated from default rules when she had reasons to believe that other terms created more joint surplus; since she must already know these reasons, she can reveal them at little cost. In sum, the drafter is the least-cost evidence gatherer.

Second, to the extent that evidence is inconclusive in either direction, the question is whether we have reasons to believe that markets will usually get it right or wrong. (Indeed, from a Bayesian perspective, the “prior” will dominate when the “posterior” is weak). And here, there is a difference between individually negotiated and adhesion contracts. For individually negotiated

contracts, there are theoretical reasons to believe that markets will in principle lead to efficient terms; after all, markets tend to work well when parties are well informed. For adhesion contracts, however, the theoretical reasons point in a different direction. Indeed, markets tend to fail when one of the parties is completely uninformed. Since signers almost never read the contract, the natural incentive for the drafter is to insert only terms that are beneficial to the drafter. If these terms also turn out to be efficient (that is, if the cost to the signer turns out to be lower than the benefit to the drafter), that is pure coincidence. So there is no *a priori* reason to presume that unread terms will be efficient. Therefore, there is no reason to enforce them when the drafter cannot prove their efficiency.

Acknowledgments Charles F. Nagel Professor of International and Comparative Law and Director of the Center on Law, Innovation & Economic Growth, Washington University in St. Louis, School of Law. E-mail: degeest@wustl.edu. I thank Adam Badawi, Omri Ben-Shahar, Michael Greenfield, and Andrew Tuch for discussions that helped crystalize the main points of this paper. I also thank Philip Lenertz, Melissa Thevenot, and Amy Xu for helpful research assistance.

References

- Ayres I, Schwartz A (2014) The no reading problem in consumer contract law. *Stanford Law Rev* 66:545–610
- Bakos Y, Marotta-Wurgler F, Trossen DR (2014) Does anyone read the fine print? Testing a law and economics approach to standard form contracts. *J Leg Stud* 43:1–35
- Bar-Gill O, Warren E (2008) Making credit safer. *Univ Pa Law Rev* 157:1–101
- Ben-Shahar O (2013) Regulation through boilerplate: an apologia. *Mich Law Rev* 112:833–904
- Ben-Shahar O, Schneider CE (2014) More than you wanted to know: the failure of mandated disclosure. Princeton University Press, Princeton
- Gazal-Ayal O (2007) Economic analysis of standard form contracts: the monopoly case. *Eur J Law Econ* 24:119–136
- Goldberg VP (1974) Institutional change and the Quasi-Invisible hand. *J Law Econ* 17:461–492
- Hay BL, Spier KE (1997) Burdens of production in civil litigation: an economic perspective. *J Leg Stud* 26:413–432
- Katz A (1990) The strategic structure of offer and acceptance: game theory and the law of contract formation. *Mich Law Rev* 89:215–295
- Kim NS (2013) *Wrap contracts: foundations and ramifications*. Oxford University Press, Oxford
- Marotta-Wurgler F (2009) Are “pay now, terms later” contracts worse for buyers? Evidence from software license agreements. *J Leg Stud* 38:309–343
- Maxeiner JR (2003) Standard-terms contracting in the global electronic age: European alternatives. *Yale J Int Law* 28:109–182
- Radin MJ (2013) *Boilerplate: the fine print, vanishing rights, and the rule of law*. Princeton University Press, Princeton
- Schwartz A, Wilde LL (1979) Intervening in markets on the basis of imperfect information: a legal and economic analysis? *Univ Pa Law Rev* 127:630–682
- Shavell S (1984) Liability for harm versus regulation for safety. *J Leg Stud* 13:357–374
- Shavell S (2013) A fundamental enforcement cost advantage of the negligence rule over regulation. *J Leg Stud* 42:275–302
- Slawson WD (1971) Standard form contracts and democratic control of lawmaking power. *Harv Law Rev* 84:529–566
- Trebilcock MJ, Dewees DN (1981) Judicial control of standard form contracts. In: Burrows P, Veljanovski CG (eds) *The economic approach to law*. Butterworths, London, pp 93–119

Simple Majority

Silvia Fedeli

Department of Economics and Law, Sapienza - University of Rome, Rome, Italy

Definition

Simple majority (SM) is a voting system whereby the highest number of votes for one alternative within those under decision designates the winner. It may refer to a voting requirement of *half* of either all ballots cast or those voting on the given alternative *plus one* and also to the highest number of votes cast for any one alternative, while not constituting a majority.

Whenever a collective decision has to be taken, SM seems the natural way to reach a solution.

Various reasons for SM's wide use have been given in the literature. May (1952), studying the issue of the relation between group choice and individual preferences in the presence of two alternatives, showed that SM is always: *decisive* (each

set of individual preferences leads to a defined/unique group choice), *egalitarian* (interchanging any two preferences of the voters, the result is unchanged, which implies anonymity/equality of voters), *neutral* (it does not favor either alternative), and *positively responsive* (if the group decision is indifferent or favorable to x and, *ceteris paribus*, a single individual changes in favor of x , then the group decision becomes favorable to x).

A further justification for the choice under SM rule has been proposed by Rae (1967) and Taylor (1967) who show that, under restrictive assumptions, SM maximizes the expected utility under the veil of ignorance.

When there are more than two alternatives, SM rule must be redefined. This, however, turns out to be tricky. The most direct method consists in confronting/voting all possible combinations in individual pairs. Each individual when confronted with any pair either prefers one or the other or is indifferent. The alternative with the highest number of preferences/votes in all pairings prevails and is called the *Condorcet winner* (Condorcet 1785). The problem is that (unlike the case of a single pair of alternatives) a Condorcet winner might not exist, and collective preferences can be cyclic (not transitive), even if the preferences of individual voters are not (*voting/Condorcet's paradox*).

Several variants of the above method differ on how they resolve such ambiguities to determine a winner. One strategy consists in arranging the alternatives in a type of elimination tournament: alternatives are paired off against each and the one chosen by SM advances to a subsequent round, while the loser is eliminated. This system can produce a group favorite and, by repeatedly invoking the system on the remaining alternatives, also a group ranking. However, the outcomes are susceptible of *strategic agenda setting*.

A different strategy (*Borda count*) tries to produce a group ranking directly from the individual rankings: each alternative is assigned a *weight* based on its positions in all the individual rankings, and the alternatives are then ordered according to their total weight (Borda 1781). Like other variants, these *positional voting systems* can originate a complete, transitive ranking

for a set of alternatives. But they suffer if competition for top spots in group ranking critically depends on the rankings of alternatives that are further down in the list. This, in turn, might induce *strategic misreporting of preferences*.

Do All These Variants Aiming to Solve the Condorcet's Paradox Mean that We Cannot Properly Decide Under SM?

In this respect, it is important to know both the probability of the paradox actually taking place and the restrictions on individual preferences that assure a Condorcet winner.

As for the probability of a cycle in collective preference, some scholars calculated that it is increasing with the number of voters and of alternatives under vote (for a survey, see Sen 1970). With three alternatives the probability is less than 10% for any number of voters, whereas with 50 alternatives and many voters, the probability of a cycle is about 90%.

This makes the second issue quite relevant: is it possible to find restrictions on individual preferences that exclude cycles?

The most famous condition is known as single peakedness. Black (1948) and Arrow (1951) showed that if all individual rankings are single peaked, then SM applied to all pairs of alternatives produces a group preference relation that is complete and transitive, so that we could produce a group ranking from it. This observation led to the *median voter theorem*, which is the basis of Downs's (1957) economic theory of democracy: with single-peaked rankings, an odd number of voters, the median individual favorite defeats every other alternative in a pairwise majority vote.

Thus, at least in this case, a decision can be taken, **but what is the quality of the decision taken by most voters? What gives most voters the right to decide for all? Moreover, SM rule presumes that all persons are equal, but what about equal ability/capability?**

Sed hoc pluribus visum est. Numerantur enim sententiae, non ponderantur; nec aliud in publico consilio potest fieri, in quo nihil est tam inaequale quam aequalitas ipsa. Nam cum sit impar prudentia, par omnium ius est. (Plinio 2 ep.12)
(But the majority gave their assent; for votes are

counted, their value is not weighed, and no other method is possible in a public assembly. Yet this strict equality results in something very different from equity, so long as men have the same right to judge but not the same ability to judge wisely. (THE LETTERS OF PLINY, Translated by Betty Radice, Loeb Classical Library 55, p. 121))

In other words, if the basis of SM is merely quantitative and not qualitative, then the resulting choice is not necessarily the best a community can make, and it is based on the reason of the strongest. Moreover, if equality depends on quantity of equal individuals, what follows is the sovereignty of the multitude and the oppression of minorities (Aristotele, *Politica* VI, 1, 6). Finally, SM rule might give ambiguous outcomes in case of “absences” and “votes to abstain” (this issue shall not be considered here, but see Freixas and Zwicker 2003, 2009; Zwicker 2009).

Nevertheless, any society has got elementary needs including that of collective decision making.

The first organized collective choices probably date back to ancient Greece, where the assembly acclaims and a leader judges prevalent favor depending on the shriek’s intensity. This means majority voting without vote count/scrutiny, with Sparta improving the method for the election of Gerontes, through the guarantee of an impartial judge (Aristotele, *Politica* II, VI, 16; on Sparta also, see Tucidide, I, 87).

In Rome the majority principle was precluded in most cases, with the prevailing rule for collective choices being the *ius prohibendi* = right of veto. Nevertheless, Romans classified the majority rule into the juridical system by means of a legal *fictio*. The decision of the majority is valid as it would be approved by all: “*quod maior pars curiae efficit, pro eo habetur ac si omnes egerint*” (“The decision taken by the majority of the senators is valid as it would be approved by all.”) (Scevola); “*refertur ad universos quod publico fit per maiorem partem*” (Ulpian). But the law recognizes only majority and assigns no right to minority.

Although the fortune of this legal *fictio* was enormous in the Middle Ages, for collective decision the primitive Church superimposed God’s will to human will, thus recovering the qualitative criterion. This precluded SM, implying that if the

majority represents God’s will, the minority denies it, which is, per se, against the dogma that the Christian community represents the body of Christ. Therefore, the Church stuck to unanimity for a long time, in spite of some problems during the Councils solved via excommunication/persuasion/conversion. The First Council of Nicaea (AD 325) produced the “Nicene Creed”: only two bishops voted against it and were excommunicated. In the Council of Chalcedon (AD 451), a minority of Egyptian bishops (12 over 600) were thought suitable for conversion.

When conversion was not possible, the Church resorted to the theory of *sanioritas* accompanying SM: the electoral bodies in the Church have the right to vote and the moral duty/obligation to vote right/well, i.e., according God’s will as recognized by a hierarchical superior. The Third Council of the Lateran (1179) accepted the Roman *fictio* only if driven by the qualitative principle of *sanior pars*. Only for the election of the Pope a majority of two third is required, with the quantitative principle merely invoked because of the absence of a superior among voters.

Without *sanioritas*, the quantitative principle simply means that force prevails (the louder *vox* or *voice* or *voix* and later *vote*, often accompanied by the din of arms) and the defeated minority has no rights, which is exactly the system used by some primitive societies:

“Si displicuit sententia, fremitu aspernantur; sin placuit, frameas concutunt. Honoratissimum adsensus genus est armis laudare” (“If his sentiments displease them, they reject them with murmurs; if they are satisfied, they brandish their spears. The most complimentary form of assent is to express approbation with their spears.”). (Tacitus, *Germania* c.11); “*Conclamat omnis multitudo et suo more armis concrepat*” “The whole multitude raise a shout and clash their arms, according to their custom” Julius Caesar. Caesar’s Gallic War. Translator. W. A. McDevitte. Translator. W. S. Bohn. 1st Edition. New York. Harper & Brothers. 1869. Harper’s New Classical Library.)

For this reason, only after the twelfth century in the Italian Medieval communes/municipalities, we find the first examples of the use of SM, whereas even the English Parliament of the origins preferred unanimity/acclamation as ideal procedure, and

only in 1554 the Parliament approved the first law by majority rule (Ruffini 1927).

It seems striking that the theory supporting majority rule only dates back to Jusuaturalism. This is mainly developed in Germany and basically assumed that the *first social contract* ought to be based on unanimity, with SM operating afterward, and being valid over the whole community, only because so established by the fundamental social contract. Grotius, first, formulated this theory; the last was Kant (notably relevant contributions are due to Pufendorf, Wolff, Hobbes, Locke, Ickstatt, Cocceius, Meyer, Rousseau, and Fichte).

But Still the Question Remains on Why a Quantity Gives a Right? And What Is the Value of the Biggest Number of Votes?

Unexpectedly, a partial answer and a rational interpretation of the majority will as the outcome closer to the truth are a mathematical one and come from Condorcet (1785): the probability for a majority outcome to be close to the truth increases with the number of voters and depends on the probability (higher than one half) that the individual voter correctly votes. This led to the *Condorcet jury theorem*. If x is the best alternative between two and each voter chooses the best alternative with some probability $p > \frac{1}{2}$, then the probability of the majority reaching a correct decision converges to one as the number of voters increases. In this sense, the theorem is an explicit formulation of the “wisdom of crowds”: aggregating the estimates of many people can lead to a decision of higher quality than that of any individual expert.

Unfortunately, the theorem does not directly apply to decisions with more than two alternatives. Moreover if $p < \frac{1}{2}$, then adding voters makes things worse: the optimal jury consists of a single voter.

Thus, we are back to our original problem, with no decisive answer on whether a rational reason, other than simply necessity, does exist as to why the minority ought to follow the will of the majority that the former does not want (the latter issue is also related to two other problems, not considered here: “tyranny of SM” (Forte 1985) and “logrolling” (Tullock 1959)).

If we have to abandon the hope to find an ex ante legitimation of the majority principle, valid for any type of society/institution, let us look at whether it is possible to find a moral foundation.

But Here We Come Back the Question About the Value of the Opinion of the Multitude

The answer would simply be:

“Was ist die Mehrheit? Mehrheit ist der Unsinn, Verstand ist stets bei wen’gen nur gewesen. (...) der Staat muß untergehn, früh oder spät, wo Mehrheit siegt und Unverstand entscheidet“. (“Majority? The majority is madness;

Reason has still ranked only with the few.

(...)

‘Twere meet that voices should be weighed, not counted.

Sooner or later must the state be wrecked, Where numbers sway and ignorance decides.”

Friedrich Schiller, Demetrius, The Perfect Library (p.18), 2010.)

or else

“La justice est sujette à dispute. La force est très reconnaissable et sans dispute. (...)Et ainsi, ne pouvant faire que ce qui est juste fût fort, on a fait que ce qui est fort fût juste”. (“Justice is subject to dispute, power is easily recognised and cannot be disputed. Thus we cannot give power to justice, because power has arraigned justice, saying that justice is unjust, and she herself truly just; so since we are unable to bring about that what is just should be strong, we have made the strong just.” (Pascal, p.67, THE THOUGHTS OF BLAISE PASCAL, translated from text of M.A. Molinier by C. Kegan Paul, London, George Bell And Sons, 1901.)

If SM is recognized to be irrational, neither just nor good, but only based on the reason of the strong, should it be avoided for social choices? Most scholars would answer *no* at least on the basis of Descartes’ “*une morale par provision*” or because, as in Bentham, SM although far from unanimity is closer to it than the opposite vote.

Justifying the majority rule is probably difficult simply because it is not a legal institute; it is rather a legal formula, which means that it does not have its own *raison d’être*, but it may get one, depending on how and where it is applied (Ruffini 1927).

Cross-References

► [Constitutional Evolution in Ancient Athens](#)

References

- Arrow KJ (1951) Social choice and individual values. Wiley, New York. 1963
- Black D (1948) The theory of committees and elections. Cambridge University Press, Cambridge
- Borda JC (1781) Mémoires sur les élections au scrutin. In Histoire de l'Académie Royale des Sciences, Paris
- de Condorcet M-J-A-N (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. De l'Imprimerie royale, Paris
- Downs A (1957) An economic theory of democracy. Harper, New York
- Forte F (1985) Competitive democracy and fiscal constitution. *Atl Econ J* 13:1–11
- Freixas J, Zwicker WS (2003) Weighted voting, abstention and multiple levels of approval. *Soc Choice Welf* 21:399–431
- Freixas J, Zwicker WS (2009) Anonymous yes-no voting with abstention and multiple levels of approval. *Games Econ Behav* 67:428–444
- May KO (1952) A set of independent, necessary and sufficient conditions for simple majority decisions. *Econometrica* XX:680–684
- Rae DW (1967) The political consequences of electoral laws. Yale University Press, New Haven
- Ruffini E (1976) Il principio maggioritario: profilo storico (1927). Adelphi, Milano
- Sen AK (1970) Collective choice and social welfare. Holden-Day, San Francisco
- Taylor MJ (1967) Proof of a theorem on majority rule. *Behav Sci* XIV:228–231
- Tullock G (1959) Some problems of majority voting. *J Polit Econ* LXVII:571–579
- Zwicker WS (2009) Anonymous voting rules with abstention: weighted voting. In: Brams S, Gehrlein W, Roberts F (eds) The mathematics of preference, choice, and order: essays in honor of Peter C. Fishburn. Springer, Heidelberg, pp 239–258

Slavery

Jenny Bourne

Carleton College, Northfield, MN, USA

Abstract

Slavery is fundamentally an economic phenomenon; it occurs when the net benefits to free society are positive. From antiquity until the mid-twentieth century, slavery was primarily a profitable way of organizing the labor force, at least to those other than the slaves

themselves. Slave societies passed laws to perpetuate the peculiar institution, often acknowledging both the property value and the personhood of slaves.

Much scholarly research focuses on the enslavement of Africans on several continents starting in the mid-1400s, the subsequent four centuries of the transatlantic slave trade, and the plantation economies of the New World. But slavery has thrived in many parts of the world for most of human history, with slaves representing productive investments and tradable wealth for their owners. More recent forms of slavery depart from this pecuniary model – enslaved political enemies may not yield monetary rewards but instead serve as symbols of dominance for tyrannical governments. Nonetheless, slavery and the legal and political institutions that support it have largely been driven by economic factors.

Overview

Throughout history, slavery has existed where it has been economically worthwhile to those in power. Certain conditions enhance the probability that any given society might hold slaves – cheaply obtained supplies of foreigners with distinctive features, division of production processes into series of simple and easily monitored tasks, and well-developed commodity markets – but slavery has flourished in many places, regardless of religion, climate, or cultural attainments.

Slavery has features in common with other forms of coerced labor. Yet slavery is a species of servitude set apart: slaves typically have been aliens with minimal freedom of action, permanently deprived of the title to their human capital and, if female, to their children, and often with no higher authority than their masters.

Central to the economic success of slavery are political and legal institutions that validate the ownership of other persons. Some slave societies have considered slavery part of the natural order; others have viewed slavery as established only by positive law. Regardless of philosophy, nations have crafted finely nuanced legal rules that acknowledge the

peculiar nature of slaves as both property and persons. Most scholarship on slave law focuses on the US South, although scattered information about other jurisdictions is also available.

The law has generally considered slaves as personal chattels (movable wealth), although late Roman and Louisiana law sometimes classified slaves as chattels real, precluding their sale separate from land. Like other chattels, slaves and their offspring could typically be bought, sold, hired, exchanged, given, bequeathed, seized for debts, and put up as collateral. What few protections slaves have had typically stemmed from a desire to grant rights to their owners. Still, when slaves stole, rioted, set fires, or killed free persons, the law usually subverted the property rights of masters so as to preserve slavery as a social institution (Bourne 1998).

Early Slavery

Slavery likely developed after civilization turned from hunting to pastoral societies: agriculture brought specialization of tasks and opportunities for exploiting another's labor. Ancient Sumerian, Phoenician, Hebrew, Babylonian, and Egyptian societies used slaves to tend flocks, work fields and mines, and help with domestic chores (Finley 1961, 1987; Drescher and Engerman 1998).

Slavery flourished in the classical period, predominantly in cities open to commercial exchange. Greek slaves worked in mines, quarries, fields, handicrafts, and domestic, police, and secretarial work; skilled slaves often lived apart from their masters. A unique feature of Greek society was benefit clubs that lent slaves money to buy their freedom (Finley 1961).

Rome inherited the slave trade and expanded it by allowing field commanders to dispose of prisoners as they wished. Consequently, slave dealers traveled with the armies and conducted on-the-spot auctions. Slaves comprised about 30% of the population in Rome's heyday. Roman slaves enjoyed limited freedoms – they could acquire property and schooling and, unlike Greek slaves, citizenship upon manumission (Finley 1961, 1987).

Slavery declined – but did not end – with the decline of Rome. The Germanic tribes held slaves, sometimes using long hair as an identifying mark. Although the church tried to prohibit the slave trade in the early ninth century, it failed: the trade was one of the West's few sources of the foreign exchange necessary to purchase Eastern goods. Charlemagne's efforts to reunify Western Europe finally sparked the transition to serfdom there (Phillips 1985).

Elsewhere, slavery persisted. With the emergence of Islam along the Arabian peninsula, religious wars generated large slave populations, mostly used for domestic and military service. William the Conqueror ended the export of English slaves, but serfdom did not replace domestic slavery in England until about 1200. Scandinavians held and traded thralls until the Swedish king declared all offspring of Christian slaves free in 1335. Slavery continued unabated in Italy and the Iberian peninsula for decades, however. Genoa and Venice had active slave markets in the thirteenth century, Seville and Lisbon two centuries later. Captains of Prince Henry the Navigator brought the first black slaves to Portugal in the 1440s (Phillips 1985; Meltzer 1993; Drescher and Engerman 1998).

Out of Africa

From 1500 to 1900, an estimated 12 million Africans were forcibly taken west to the New World, with about two million dying en route. Nearly three million departed between 1811 and 1867 alone. An additional six million African slaves were sent from sub-Saharan Africa eastward to Arab, Asian, and even some European countries; another eight million remained enslaved on their own soil (Lovejoy 1983).

African slaves came with Europeans to the New World from the very beginning. Slaves accompanied Ponce de Leon to Florida in 1513, for example. The first African bondsmen in what became British North America arrived in Virginia in 1619 on a Dutch ship (Fogel 1989).

But institutional trading arrangements soon replaced small-scale trafficking. Although early

Spanish settlers at first enslaved native Americans, the indigenous population perished in droves from smallpox and punishingly hard work. Via licensing agreements, Portugal and Spain began importing African slaves into their colonies in the early 1500s to replace the dying natives. In 1517, Charles I began to import Africans to the West Indies and the Spanish-controlled mainland. Brazil commenced legal importation of black slaves in 1549. Traders could acquire slaves in specific African zones, then sell them in America – up to 120 per year per Brazilian planter, for example (Blackburn 1997).

Britain eventually replaced Spain and Portugal as the major slave-trading nation. Elizabethan-era captain Sir John Hawkins had transported slaves, but England did not figure large in the African trade until the seventeenth-century Royal African Company became the biggest slaver in the world (Galenson 1982).

The Atlantic slave trade created an infamous “triangle.” European captains plied the West African chieftains with liquor, guns, cotton goods, and trinkets. In exchange, Africans supplied slaves to suffer through the arduous “middle passage.” Slaves worked on sugar plantations in the West Indies, grew coffee and mined metals in Brazil, and raised tobacco, sugar, rice, and cotton in North America. These goods, along with molasses and rum, went back to England (Solow and Engerman 1987).

Many interests benefited from the Atlantic slave trade, including ship captains, European financiers, African dealers, and American industrialists. The Royal African Company was hardly a monopoly – individual ship captains earned upwards of a 10% return on slave cargo (Eltis 1987; Drescher 1999). European banks and merchant houses profited from lucrative credit and insurance arrangements with New World plantation owners (Engerman 1972; Menard 2006). African tribes sold slaves for horses and guns, then used these items to obtain more slaves (Lovejoy 1983). New England shipbuilders and textile producers also had strong connections to the slave trade and slavery (Bailey 1990).

Few imported slaves actually ended up in the United States. When the transatlantic trade to the

United States officially ended in 1808, only 6% of African slaves had come to North America. The remainder went further south, with the majority landing in the Caribbean and the rest primarily in Brazil (Fogel 1989; Blackburn 1997; Manning 2006).

Importation of slaves remained key for decades in keeping Caribbean plantations viable. In Jamaica, for example, the black population did not become self-sustaining until after the mid-1830s (Fogel and Engerman 1974, Menard 2006). The total Caribbean slave population increased from 150 thousand in 1700 to 1.2 million in 1790, with slaves comprising about 90% of the population in the late eighteenth century (Menard 2006).

Slavery in the US South was unique in that the rate of natural increase among enslaved blacks was nearly as large as among free whites. Compared with elsewhere in the New World, slave mortality rates in the South were lower due to a milder climate and less grueling tasks, and the ratio of females to males was far less skewed (Engerman and Genovese 1975; Solow and Engerman 1987). Among those who supported the closing of the transatlantic slave trade, in fact, were Southern slaveowners, who expected that reduced supply would drive up slave prices and hence their personal wealth. Because of the high fertility and low mortality rates among Southern slaves, masters did not fear a long-term shortage. Slaves comprised less than a tenth of the total Southern population in 1680 but grew to a third by 1790. After the American Revolution, the Southern slave population exploded, reaching about 1.1 million in 1810 and over 3.9 million in 1860. By 1825, 36% of the slaves in the Western hemisphere lived in the United States (Fogel 1989).

Slave Productivity

Slave trading was profitable because slavery was profitable. As Moses Finley (1961) puts it, Greek and Roman slaveowners went on for centuries believing they were making profits – and spending them. Robert Fogel and Stanley Engerman,

along with several of their students, have amassed empirical evidence that has convinced most scholars that slaves in the US South (particularly in agriculture) were far more productive than free Northern laborers. Slavery also worked well in industry and in urban areas (Goldin 1976; Dew 1991); in fact, workers at the Tredegar Iron Works in Richmond, Virginia, went out on their first strike in 1847 to protest the use of slave labor (Bourne 1998).

A large proportion of the reward to owning slaves resulted from innovative labor practices among growers of cash crops (Fogel and Engerman 1974). The use of the “gang” system in agriculture contributed to profits in Roman times as well as in the antebellum period. In the West Indies, for instance, sugar slaves worked in three separate gangs: one harvesting the crop, a second cleaning up after the first, and a third bringing water and food to the others (Dunn 1972; Menard 2006). Cotton planters in the US South also used the gang system to their advantage (Fogel and Engerman 1974).

Integral to many gang-oriented operations across the Americas were plantation overseers. The vigorous protests against drafting overseers into military service during the US Civil War confirm their significance. Yet law and custom alike circumscribed overseers’ abilities to administer correction, because slaves represented so much wealth to their masters (Bourne 1998). This balance between authority and accountability worked well enough to turn handsome profits for many planters, particularly those who entrusted reliable slave drivers as liaisons. By the mid-1820s, black slave managers actually took over much of the day-to-day operations in the West Indies (Dunn 1972).

Slaveowners experimented with other methods to increase productivity. Southern masters developed an elaborate system of “hand ratings” to improve the match between slave worker and the job. Slaves sometimes earned bonuses or quit early if they finished tasks quickly. Slaves – in contrast to free workers – often had Sundays off (Fogel and Engerman 1974, Fogel 1989). Some masters allowed slaves to keep part of the harvest or to work their own small plots. In places, slaves

could even sell their own crops and produce. To prevent stealing, however, many masters limited the crops that slaves could raise and sell, confining them to corn or brown cotton, for example. Masters capitalized on the native intelligence of slaves by using them as agents to receive goods, keep books, and the like (Bourne 1998). In some societies – for example, ancient Rome and her offspring, antebellum Louisiana – slaves even controlled a sum of money called a *peculium* (Schafer 1994). This served as a sort of working capital, enabling slaves to establish thriving businesses that often benefited their masters as well.

Slave Sales and Prices

Slaves benefited their owners via their marketability as well as their productive labor. Healthy male slaves were sold for a date plantation grove in Sumeria in 2000 BC. In early Athens, slaves cost about a year’s keep; prices soon rose significantly as slaves went to work as managers, secretaries, and civil servants. Eight oxen bought a slave among Anglo-Saxon tribes in the mid-400s (Finley 1961, 1987; Meltzer 1993). Prime field hands went for about five hundred dollars in the United States in 1800, about fourteen hundred dollars in 1850, and up to three thousand dollars on the eve of the Civil War (Fogel and Engerman 1974).

Several early societies set up thriving slave sale markets. Greek slaves were sold with other commodities in the agora. Judging from the distribution of Italian wine jars in Gaul, wine was evidently exchanged for Gallic slaves in the first and second centuries BC. Hernando Cortés described the large number of slaves brought to the great marketplace near Tenochtitlan (present-day Mexico City) (Phillips 1985; Watson 1987; Meltzer 1993).

Slave markets also existed across the antebellum US South. Even today, one can find stone markers like the one next to the Antietam battlefield, which reads: “From 1800 to 1865 This Stone Was Used as a Slave Auction Block. It has been a famous landmark at this original location for over 150 years.”

Professional traders facilitated exchange. Established dealers like Franklin and Armfield in Virginia and Nathan Bedford Forrest (Confederate cavalry officer and later Ku Klux Klan leader) in Tennessee prospered alongside itinerant traders who operated in a few counties, buying slaves and moving them in chained coffles to the lower South (Bancroft 1931).

Prices served as signals of market conditions. The price of Egyptian slaves in the fourth century dropped dramatically, for example, when the sale of babies became legal (Finley 1961). Conversely, the price of US slaves rose upon the close of the international slave trade, as did European slave prices after the ravages of the Black Death. Slave prices fell with cotton prices in the 1840s but rebounded in the 1850s as the demand for cotton and tobacco grew (Fogel and Engerman 1974, Fogel 1989). Slavery in the US South remained a thriving business on the eve of the Civil War: even controlling for inflation, slave prices rose significantly in the six decades before secession. By some estimates, average slave sale prices by 1890 would have increased more than 50% over their 1860 levels had war and emancipation not occurred (Fogel 1989).

Studies reveal that the prices of slaves also varied by their sex, age, skill levels, and physical condition (Fogel and Engerman 1974; Phillips 1985; Fogel 1989). Prices generally rose as slaves acquired strength and experience when they approached puberty; prices declined after slaves reached their mid-twenties (Kotlikoff 1979; Engerman and Genovese 1975; Fogel 1989).

One pricing variant appeared in the West Indian “scramble.” Here, owners and agents devised a fixed-price system, dividing slaves into four categories, penning them up accordingly, and assigning a single price per pen. Potential buyers then jumped into the pens, attempting to pick off the best prospects for the price (Dunn 1972; Menard 2006).

An interesting influence on prices has to do with childbearing. Fertile female slaves were sold for a premium in the US South because slaves’ expected life span was long enough that infant slaves had positive sale value (Fogel and Engerman 1974). In contrast, men who

impregnated female slaves in medieval Italy had to compensate their masters – because childbirth deaths were so common. Genoan men could even buy indemnification insurance against this possibility (Meltzer 1993).

Along with US slave sale markets came sophisticated methods for coping with risk, such as explicit – and even implicit – warranties of title, fitness, and merchantability (Bourne 1998). Legally, slave sellers were responsible for their representations, required to disclose known defects in their wares, and often liable for unknown defects, as well as bound by explicit contractual language. These rules stand in stark contrast to the *caveat emptor* doctrine commonly applied to antebellum commodity sales. South Carolina and Louisiana were particularly pro-buyer in slave sale cases.

One feature concerning slave sale laws deserves mention. If a slave was seized and sold unlawfully by someone other than his master, Southern courts often decided that the slaveowner was entitled to the return of his slave rather than mere money damages because, as a North Carolina court stated in *Williams v. Howard* (1819): “[F]or a faithful or family slave, endeared by a long course of service or early association, no damages can compensate; for there is no standard by which the price of affection can be adjusted, and no scale to graduate the feelings of the heart.”

Slave Hiring

Robust slave-hiring markets existed as well as slave sale markets. In the United States, slave hiring took root by the Revolutionary Period and thrived throughout the antebellum period (Martin 2004). Hiring mattered both in industry and agriculture; it was frequently used to keep slaves occupied during slack growing seasons and after an owner died but before the estate was settled (Goldin 1976; Dew 1991). Annual hire rates for young adult males approached \$150 during the 1850s.

Bond and free workers both faced a legal burden to behave responsibly on the job. Yet the law of the workplace differed significantly for the two (Bourne 1998). When free laborers suffered

injuries and brought suit, antebellum judges nearly always sided with employers. Under the doctrine of *Baker v. Bolton* (1808), English and American courts held that torts died with victims. In contrast, slaveowners often recovered damages for injuries to hired slaves. And, because the “victim” in a slave-hiring case was the master, courts did not apply the *Baker* rule.

People had duties to treat hired slaves the same as slaves they owned. Hirers retained authority to direct and discipline slaves; otherwise, they might not have elicited work. Masters feared slave exploitation, however, and often specified tasks and work locations in hiring contracts (Bourne 1998). Hirers paid when they violated such clauses, even if slaves had been disobedient, suicidal, drunk, or careless. Negligent hirers also paid for escaped slaves. But Southern hiring law recognized the personhood of slaves as well. If hired slaves were hurt or killed through their own carelessness, the loss fell on their owners as long as employers had complied with contracts. Courts also relieved employers of liability when slaveowners knew about risks of jobs their slaves were to perform.

Laws Governing Those Other than Slaveowners

Society at large shared in maintaining the machinery of slavery. In the United States, for example, all Southern states save Delaware passed laws to establish citizen slave patrols. Northern citizens often worked hand-in-hand with their Southern counterparts, returning fugitive slaves to masters either with or without the prompting of national and state law. Yet not everyone was civic-minded. As a result, the profession of “slave catching” evolved – often highly risky (enough so that insurance companies denied such men life insurance coverage) and just as often highly lucrative (Bourne 1998).

Slaves, like other chattels, constituted valuable cargo on common carriers. Unlike inanimate goods, slaves could escape from danger. But cargo they nonetheless were – unlike free passengers, slaves embodied large and easily calculated property values. Consequently, US Southern law

regarding slaves injured in transport struck a balance between the strict-liability standard governing goods and the no-liability standard for passengers (Bourne 1998).

Slaves also suffered wounds or death in accidental collisions, particularly with trains. Although some states held railroad companies strictly liable for injuries to livestock, legal authorities refused to do the same when slaves were hurt. In fact, early court opinions favored transportation interests in railroad mishaps involving any person, slave or free. But, over time, Southern courts modified liability rules: common-carrier defendants that had failed to offer slaves – even negligent slaves – a last clear chance to avoid accidents ended up paying damages to slaveowners. Slaveowner plaintiffs won several cases in the decade before the Civil War when engineers failed to warn slaves off the tracks. This “last-clear-chance” rule did not generally apply to free victims until the twentieth century (Bourne 1998).

US Southern law also shaped interactions among slaves and strangers, awarding damages more often for injuries to slaves than injuries to other property or persons, shielding slaves more than free persons from brutality, and generating convictions more frequently in slave-stealing cases than in other criminal cases (Bourne 1998). The law also recognized more offenses against slaveowners than against other property owners because slaves, unlike other property, succumbed to influence.

Despite some restrictions, strangers could discipline slaves in many circumstances and could defend themselves from slaves. Still, they faced civil and criminal penalties for wanton assaults. By comparison, free persons in the US South virtually never recovered civil damages nor saw criminal sanctions applied in assault cases.

Just as assaults of slaves generated civil damages and criminal penalties, so did stealing a slave to sell him or help him escape to freedom. Many Southerners considered slave stealing worse than killing fellow citizens. The counterpart to helping slaves escape – picking up fugitives – also created laws. States offered rewards to defray the costs of capture or passed statutes requiring owners to pay fees to slave catchers.

Among the South's concerns were activities that weakened masters' authority. Most states prohibited the sale of spirits to slaves and passed statutes to outlaw slave assemblies. Many states also forbade the sale or delivery of poison to slaves and imposed fines on people who bought goods from slaves.

The Master-Slave Relationship

In spite of the necessary subjugation entailed by slavery, the law – particularly in the 30 years before the US Civil War – limited owners somewhat (Bourne 1998). Southerners feared that unchecked slave abuse could lead to unpleasant scenes of pilfering, public beatings, and insurrection.

Many states required masters to give adequate food and clothing to slaves, just as states later required livestock owners to tend to their animals and – much later – made household heads provide for their families. Most states stopped short of permitting people to kill slaves; North Carolina, Alabama, Mississippi, and Virginia were willing to convict masters who had murdered their own slaves. Colonial South Carolina and Georgia restricted nonfatal abuse as well as murder, although Virginia, North Carolina, Maryland, and Delaware did not. Later on, the codes of ten Southern states instituted fines or imprisonment for cruel masters. Still, prosecuting masters was extremely difficult, because often the only witnesses were slaves or wives, neither of whom could testify against male heads of household.

Southern law encouraged benevolence, at least if it tended to supplement the lash and shackle. Court opinions indicate the belief that good treatment of slaves could enhance labor productivity, increase plantation profits, and reinforce sentimental ties. Courts also permitted slaves small diversions, such as Christmas parties and quilting bees, despite statutes that barred slave assemblies.

The master's ultimate kindness was to bestow freedom. Yet allowing masters to manumit slaves at will would create perverse incentives to free unproductive slaves. Consequently, the community at large would bear the costs of old, disabled, and rebellious former slaves. Roman emperor

Augustus worried considerably about this adverse selection problem and eventually enacted restrictions on the age at which slaves could be free, the number freed by any one master, and the number manumitted by last will (Watson 1987).

By the 1830s, most Southern states had begun to restrict manumission, often forcing newly freed slaves to leave the state. Some required former masters to file indemnifying bonds with state treasurers so governments would not have to support indigent former slaves. Some instead required former owners to contribute to ex-slaves' upkeep. Many states limited manumissions to slaves of a certain age who were capable of earning a living. By 1860, most Southern states had banned in-state and postmortem manumissions, and some had enacted procedures by which free blacks could voluntarily become slaves (Bourne 1998).

In addition to constraints on manumission, laws restricted other actions of masters and, by extension, slaves (Bourne 1998). Masters generally had to maintain a certain ratio of white to black residents upon plantations. Some laws barred slaves from owning musical instruments or bearing firearms. All states refused to allow slaves to make contracts or testify in court against whites. About half of Southern states prohibited masters from teaching slaves to read and write. Although masters could use slaves for various tasks and responsibilities, they typically could not order slaves to compel payment, beat white men, or sample cotton. Nor could slaves hire themselves out to others, although such prohibitions were often ignored by masters, slaves, hirers, and public officials. Owners faced fines and civil damages if their agent slaves caused injuries. In some jurisdictions, masters even paid for losses arising from willful acts (particularly theft) committed by their slaves. In this last respect, the law placed greater responsibility on owners of slaves than on masters of servants.

Slave Criminals

Slaves, like their free brethren, committed a host of crimes ranging from arson to theft to homicide. Other slave crimes in the US South included

violating curfew, attending religious meetings without a master's consent, and running away (Bourne 1998).

Southern states erected numerous punishments for slave crimes, including prison terms, banishment, whipping, castration, and execution (Bourne 1998). In most states, the criminal law for slaves (and free blacks) was noticeably harsher than for free whites; in others, slave law as practiced resembled that governing poorer white citizens. Particularly harsh punishments applied to slaves who had killed their masters or committed rebellious acts. Southern colonies considered these acts of treason and resorted to burning, drawing and quartering, and hanging.

Originally, many jurisdictions tried to spirit criminal slaves out of state. But Southern states quickly caught on and began to ban slave imports or check for criminal backgrounds of imported slaves. Transportation soon gave way to capital punishment for many slave crimes. Most states simultaneously implemented policies of compensating slaveowners for part of the value of executed slaves. These policies reduced incentives to conceal slave crimes or sell criminals to unsuspecting buyers. Yet lawmakers did not want people to abuse the state's generosity, so owners could not recover full value nor coerce slaves to confess.

Besides prohibiting coerced confessions, Southern states adhered to various other criminal procedures for slaves. Slave defendants had attorneys, in part because the alternative would have been self-representation. Slaves generally had jury trials, though obviously not of their peers. And slaves could testify against other slaves, although supportive testimony by one slave in defense of another carried little weight.

Modern Slavery

Rising British sentiment against slavery culminated in the *Somerset* case, which outlawed slavery in England in 1772. Britain and the United States abolished the international slave trade in 1808; Britain freed slaves in its colonies (except India) in 1833, with full emancipation in 1838.

Slavery in the United States officially disappeared by the end of the American Civil War in 1865.

Abolition came later elsewhere, often accompanied by an increase in other forms of servitude. Eastern Europe and Russia kept slavery alive into the late nineteenth century. In 1890, all major European countries, the United States, Turkey, Persia, and Zanzibar, signed the General Act of Brussels in an attempt to suppress slavery. Forty years later, an international labor convention acted to outlaw forced labor in the former Ottoman and German colonies. In 1948, the UN General Assembly declared that all forms of slavery and servitude should be abolished (Bush 1996).

Yet slavery in Southeast Asia, the Arabian peninsula, and Africa continued well into the twentieth century (Watson 1980; Lovejoy 1983; Klein 1993). Perhaps the saddest legacy of American slavery is that the system established to supply the New World with slaves also shaped society at home. Slavery within Africa burgeoned along with the Atlantic trade – in 1600 Africa had a minority of the world's slaves, in 1800 the overwhelming majority. The Great Scramble for Africa spread slavery further. Regrettably, the tragedy continues: Angolan slaves fought bloodily for freedom in 1961; Mauritania kept slavery legal until 1981; Nigeria still had slave concubinage in the late 1980s; and numerous African regions practice slavery today.

Some of the harshest forms of slavery have arrived only recently. Modern weaponry, increased population density, and mass communication and transportation technology have made it easy to capture and move purported enemies. The classic mechanism of modern slavery, patterned after the practices of Nazi Germany and the Soviet Union, is this: officials in power arrest suspected opponents of the current political regime, or those considered racially or nationally unfit, and throw them into forced-labor camps to work under terrible conditions. Unlike slaves in earlier societies, the unfortunates who landed in Nazi and Soviet concentration camps were not privately owned and traded in open markets. Rather, they served as property of the state, sometimes to be rented out to private interests.

Modern slaves consequently represent something much different than their historic counterparts. From preclassical times through the nineteenth century, masters typically viewed their slaves as productive investments, as book-keeping entries in their wealth portfolios, and as forms of valuable capital. Slaves in these circumstances could often count on minimal food, shelter, clothing, and time for rest and sleep. Not so for the “publicly owned” slaves of the twentieth century. Because these people were “acquired” at very low cost with public dollars and served primarily as political symbols, their masters had little incentive to care for them as assets.

To be sure, when Nazi Germany needed labor to fuel production of its war machinery, the country turned to the inmates of concentration camps. Likewise, the Soviets rounded up peasants to work on public projects and mineral extraction. Various regions across Africa, Asia, America, and Europe have done the same. Yet these sorts of “slaves” are often worth more dead than alive. Killing one’s political adversaries makes the state that much easier to run. Exterminating those labeled as unfit “cleanses” society – in a truly twisted sense of the word – and binds together the “chosen.” Accordingly, modern forms of mass slavery seem far different institutions than those of earlier times.

No type of slavery is benign, of course – despite the relatively better treatment of slaves held as productive investments, any form of slavery is at its core coercive. The behavior of American ex-slaves after abolition – who spent more time in leisurely pursuits and refused to work in gangs – clearly reveals that they valued their non-work time more highly than masters had and cared strongly about the manner of their work (Fogel and Engerman 1974; Fogel 1989). In effect, slavery is an economic phenomenon only because it neglects the costs to the slaves themselves.

References

- Bailey R (1990) The slave(ry) trade and the development of capitalism in the United States: the textile industry in New England. *Soc Sci History* 14:373–414
- Bancroft F (1931) *Slave trading in the old South*. Ungar, New York
- Blackburn R (1997) *The making of New World slavery: from the baroque to the modern*. Verso, London/New York
- Bourne (Wahl) J (1998) *The bondsman’s burden: an economic analysis of the common law of southern slavery*. Cambridge University, New York
- Bush M (ed) (1996) *Serfdom and slavery: studies in legal bondage*. Addison Wesley Longman, New York
- Dew C (1991) *Slavery in the antebellum southern industries*. University Publications of America, Bethesda
- Drescher S (1999) *From slavery to freedom: comparative studies in the rise and fall of Atlantic slavery*. Macmillan, London
- Drescher S, Engerman S (eds) (1998) *A historical guide to world slavery*. Oxford University Press, New York/Oxford
- Dunn R (1972) *Sugar and slaves: the rise of the planter class in the English West Indies, 1624–1713*. University of North Carolina, Chapel Hill
- Eltis D (1987) *Economic growth and the ending of the transatlantic slave trade*. Oxford University, New York
- Engerman S (1972) The slave trade and British capital formation in the 18th century: a comment on the Williams thesis. *Bus His Rev* 46:430–443
- Engerman S, Genovese E (1975) *Race and slavery in the western hemisphere: quantitative studies*. Princeton University, Princeton
- Finley M (1961) *Slavery in classical antiquity: views and controversies*. Heffer, London
- Finley M (ed) (1987) *Classical slavery*. Cass, London
- Fogel R (1989) *Without consent or contract*. Norton, New York
- Fogel R, Engerman S (1974) *Time on the cross: the economics of American Negro slavery*. Little Brown, Boston
- Galenson D (1982) The Atlantic slave trade and the Barbados market, 1673–1723. *J Econ His* 42:491–511
- Goldin C (1976) *Urban slavery in the American South, 1820–1860: a quantitative history*. University of Chicago, Chicago
- Klein M (1993) *Breaking the chains: slavery, bondage, and emancipation in modern Africa and Asia*. University of Wisconsin, Madison
- Kotlikoff L (1979) The structure of slave prices in New Orleans, 1804–1862. *Econ Inq* 17:496–518
- Lovejoy P (1983) *Transformations in slavery: a history of slavery in Africa*. Cambridge University, New York
- Manning P (2006) *Slavery and African life: occidental, oriental, and African slave trades*. Cambridge University, New York
- Martin J (2004) *Divided mastery: slave hiring in the American South*. Harvard University, Cambridge
- Meltzer M (ed) (1993) *Slavery: a world history*. Da Capo, New York
- Menard R (2006) *Sweet negotiations*. University of Virginia, Charlottesville
- Phillips W (1985) *Slavery from Roman times to the early transatlantic trade*. Manchester University, Manchester

- Schafer J (1994) Slavery, the civil law, and the supreme court of Louisiana. Louisiana State University, Baton Rouge
- Solow B, Engerman S (1987) Capitalism and Caribbean slavery: the legacy of Eric Williams. Cambridge University, New York
- Watson J (ed) (1980) Asian and African systems of slavery. University of California, Berkeley
- Watson A (1987) Roman slave law. Johns Hopkins University, Baltimore

Small Claims Litigation

► Class Action and Aggregate Litigations

Smith, Adam

Manfred J. Holler

Center of Conflict Resolution (CCR), University of Hamburg, Institute of SocioEconomics (ISE), Hamburg, Germany

Abstract

This entry discusses Adam Smith's theory of justice and its application to the law. Smith's theory of justice combines natural and acquired rights with a focus on commutative justice and economic reasoning. Core concepts are resentments, sympathy, the impartial spectator, and liberty. A possible conflict between liberty and economic expedience is illustrated by Smith's discussion of usury and bank regulation.

The Missing Book

In the closing lines of *The Theory of Moral Sentiments* (*TMS*), Adam Smith (1723–1790) states: “I shall in another discourse endeavour to give an account of the general principles of law and government, and of the different revolutions they have undergone in the different ages and periods of society, not only what concerns justice, but in what concerns police, revenue, and arms, and whatever else is the object of law” (*TMS* 1982/

1759/1790, p. 342). However, Smith has never written this book that, given his deep insights into the working of the economy, could have become the pioneer work in what has become Law and Economics.

In the “Advertisement,” which introduces the sixth edition of *The Theory of Moral Sentiments* published in 1790, the year of his death, Smith refers explicitly to the above quote commenting: “In the *Enquiry concerning the Nature and Causes of the Wealth of Nations*, I have partly executed this promise; at least so far as concerns police, revenue, and arms. What remains, the theory of jurisprudence, which I have long projected, I have hitherto been hindered from executing, by the same occupations which had till now prevented me from revising the present work. Though my advanced age leaves me, I acknowledge, very little expectation of ever being able to execute this great work to my satisfaction; yet, as I have not altogether abandoned the design, and as I wish still to continue under the obligation of doing what I can, I have allowed the paragraph to remain as it was published more than thirty years ago, when I entertained no doubt of being able to execute every thing which it announced” (*TMS*: no page number). Thus, in order to explore Smith's concept of law and its economic analysis, we have either to rely on the references in his two published volumes, i.e., *The Theory of Moral Sentiments*, first published in 1759, and *An Inquiry into the Nature and Causes of the Wealth of Nations* (*WN*), first published in 1776/1777, or on the *Lectures on Jurisprudence* (*LJ*). The latter were compiled from notes of students who attended his lectures at Glasgow University during the sessions of 1762–1763 and 1763–1764. (This was in the years before Smith resigned from his chair of Moral Philosophy at Glasgow University and went on the grand tour of France as a tutor of Henry Scott, the young Duke of Buccleuch in 1764.) The close correspondence of parts of these lecture notes with material published in *TMS* and *WN* suggests that they are quite reliable reports of what Smith presented to his audience – as confirmed by Edwin Cannan who, in 1985, discovered “Lectures on Justice, Police, Revenue and Arms,” the first of the two

manuscripts which are edited as *LJ* by Meek, Raphael, and Stein (see *LJ* 1982/1762–1766, p. 5f). McCormick (1981, p. 244) confirms that in Division II of Part II of *LJ* “we find much of the theorizing of *The Wealth of Nations* already present in embryo.”

In the following, I will focus on the material of these three books. Sections “[Sympathy and the Impartial Spectator](#)” and “[Promises and Subsistence](#)” summarize Smith’s theory of justice and outline its building blocks: resentments, sympathy, and the impartial spectator. Section “[Selected Cases and Prominent Issues](#)” comments on Smith’s discussion of usury and bank regulation and illustrates a possible conflict between *natural liberty* and economic expediency. Some general remarks on Smith’s perspective on human nature conclude the paper. First, however, I will briefly clarify the concepts of jurisprudence, law, and justice and their relationship in Smith’s work.

Jurisprudence, Law, and Justice

Smith defines “jurisprudence” as “the theory of the rules by which civil governments ought to be directed” (*LJ*, p. 5) and, alternatively, “the theory of the general principles of law and government” (*LJ*, p. 398). “The four great objects of law are Justice, Police, Revenue, and Arms” (*LJ*, p. 398). “The object of Justice is security from injury, and it is the foundation of civil government” (*LJ*, p. 398). A “man” may be injured “as a man” and “as a member of a family” and “of a state” (*LJ*, p. 399). “As a man, he may be injured in his body, reputation, or estate” (*LJ*, p. 399). Smith considers the prevention of the body and the reputation from injury as natural rights, while he classifies the rights of estate to be acquired rights. The latter “are of two kinds, real and personal” (*LJ*, p. 399). Property falls into the category of real rights. Note that John Locke (1632–1704) claimed that man’s natural rights are life, liberty, and property. Inasmuch as acquired rights are defined by the society, Smith’s acknowledgment of property is weaker as it is not embedded in the human nature.

It is this combination of natural and acquired rights, and to justice, that distinguishes Adam

Smith from utilitarians like Jeremy Bentham and John Stuart Mill and from the core of modern scholars of Law and Economics who focus on efficiency to draw their conclusions. Smith’s reference to justice concurs with the observation, elaborated in Holler and Lerach (2010), that people are more concerned with justice than with efficiency. Of course, Smith accepts that the government can also intervene to achieve efficiency. His support for public education – “those for the education of youth, and those for the instruction of people of all ages” (*WN* 1981/1776/77, p. 723) –, the abolishment of slavery (because of inefficiency and not because of injustice), and the rigorous regulation of the banking sector are prominent examples. He is quite aware that government intervention is necessary to overcome the collective action problem (due to potential free-riding). The duty of “the sovereign or commonwealth is that of erecting and maintaining those public institutions and those public works, which, though they may be in the highest degree advantageous to a great society, are, however, of such a nature that the profit could never repay the expence to any individual or small number of individuals, and which it therefore cannot be expected that any individual or small number of individuals should erect or maintain” (*WN*, p. 723).

In addition, violations of justice are a strong argument to justify public action. Smith’s theory of justice builds on injuries and resentments. The corresponding feelings indicate violations. (“Resentment seems to have been given us by nature for defence, and for defence only. It is the safeguard of justice and the security of innocence” (*TMS*, p. 79).) As a consequence, in Smith’s theory “justice is necessarily conceived of as being corrective rather than distributive” (McCormick 1981, p. 247), i.e., the focus is on commutative justice (with a potential threat of punishment). However, whether a particular injury can prevail and triggers commonly shared resentments depends on the organization of the society. As Smith pointed out, if property does not exist or is of little importance, like in the society of hunters, then an unauthorized acquisition cannot result in injuries and resentments. But whether there is an injury and whether this injury triggers resentments

and whether these resentments are acknowledged as justified depends on the social value system as captured by the impartial spectator. That is, justice depends on the social development of the society reflected in its economic and moral conditions. Thus, “laws and legal institutions are an inherent part of the economy of a society and must be understood and explained as such” (MacCormick 1981, p. 254).

Sympathy and the Impartial Spectator

Smith’s impartial spectator “can be regarded as a ‘tool’ helping people to understand which actions, emotions and passions are appropriate. In order to gain approval and sympathy of the impartial spectator humans must exercise self-command to ensure that we are not overwhelmed by our passions” (Huber 2014, p. 11). It results from our potential to put ourselves in the position of others to see things from their point of view. In *TMS*, this potential is labeled sympathy, but today we would call it empathy. “Pity and compassion are words appropriated to signify our fellow-feeling with the sorrow of others. Sympathy, though its meaning was, perhaps, originally the same, may now, however, without much impropriety, be made use of to denote our fellow-feeling with any passion whatever” (*TMS*, p. 10).

It is not the passion of others which puts our sympathy in motion, but our *hypothetical* experience of being in the other person’s position. “Sympathy, therefore, does not arise so much from the view of the passion, as from that of the situation which excites it. We sometimes feel for another, a passion of which he himself seems to be altogether incapable; because, we put ourselves in his case, that passion arises in our breast from the imagination, though it does not in his from the reality. We blush for the impudence and rudeness of another, though he himself appears to have no sense of the impropriety of his own behaviour; because we cannot help feeling what confusion ourselves should be covered, had we behaved in so absurd a manner” (*TMS*, p. 12). Moreover, “as we have no immediate experience of what other men feel, we can form no idea of the manner in

which they are affected, but by conceiving what we ourselves should feel in the like situation” (*TMS*, p. 9). More drastically, Adam Smith observes: “We sympathize even with the dead, and overlooking what is of real importance in their situation, that awful futurity which awaits them, we are chiefly affected by those circumstances which strike our senses, but can have no influence upon their happiness” (*TMS*, p. 12).

Smith’s notion of sympathy differs substantially from David Hume’s notion as the following quotation shows: “When I see the effect of passion in the voice and gesture of any person, my mind immediately passes from these effects to their causes, and forms such a lively idea of the passion, as is presently converted into the passion itself” (Hume 1978/1739, p. 576). In Hume, it is the other’s *experience* and *passion* which triggers the fellow feelings, while Smith’s fellow feelings are due to the other’s *situation* or *action*.

There are immediate consequences of Smith’s notion of sympathy: the potential of self-evaluation and the pleasure of mutual sympathy. In the latter, it seems that Smith borrows from the everyday notion of sympathy. However, let us first look at the potential of self-evaluation. Smith (*TMS*, p. 109) observes that the “principle by which we naturally either approve or disapprove our own conduct, seems to be altogether the same with that by which we exercise the like judgements concerning the conduct of the other people. We either approve or disapprove of the conduct of another man according as we feel that, when we bring his case home to ourselves, we either can or cannot entirely sympathize with the sentiments and motives which directed it. And in the same manner, we either approve or disapprove of our own conduct, according as we feel that, when we place ourselves in the situation of another man, and view it, as it were, with his eyes and from his station, we either can or cannot entirely enter into and sympathize with the sentiments and motives which influenced it (*TMS*, p. 109).” Smith argues: “We can never survey our own sentiments and motives, we can never form any judgement concerning them; unless we remove ourselves, as it were, from our own natural station, and endeavour to view them as at a certain distance

from us. But we can do this in no other way than be endeavouring to view them with the eyes of other people, or as other people are likely to view them.” Smith concludes: “We endeavour to examine our own conduct as we imagine any fair and impartial spectator would examine it” (*TMS*, p. 109).

We could argue that people follow this pattern and that is why sympathy allows for self-evaluation and the evaluation of others. However, Smith gives a further argument: “In this...as well as in every other emotion, passion, and habit, the degree that is most agreeable to the impartial spectator is likewise most agreeable to the person himself...” (*TMS*, p. 262). As we are social beings, we do not ignore the judgment of others. On the contrary, we strive for their appreciation and are therefore willing to change our habits and our behavior such that others can agree with them. From the conduct of sympathy, we derive an impartial spectator, “a man within the breast,” that helps us to evaluate, to control, and to restructure our behavior. It influences our conduct without identifying the evaluators, i.e., their judgments or, more specifically, their preferences or value systems.

However, the impartial spectator is a rather complex creature when applied to other persons. There is a “first-person-plural perspective I share with all others to whom my judgment is implicitly addressed” (Darwall 1999, p. 160) and, secondly, I put myself into the shoes of the person being judged by me. “When I make a moral assessment of someone’s motive or feeling, according to Smith, I express a sympathy with it that I expect any one (of us) to share. I impartially project myself into that person’s standpoint, not as myself but as any of us, and (attempt to) judge what any of us would be moved to do or feel if in that person’s shoes” (Darwall 1999, p. 160).

The construct of an impartial spectator and the self-evaluation by means of sympathy presuppose that there is some common platform of experience in this society. Not friendly fellow feelings trigger the “pleasure of mutual sympathy” (*TMS*, p. 262) but the fact that sympathy can work because of the common experience which the members of a society share. The common experience creates the familiarity, paired with universality, on which

the impartial spectator functions (Witztum 1997; Leroch 2008). “Man, say they, conscious of his own weakness, and of need which he has for assistance of others, rejoices whenever he observes that they adopt his own passions because he is then assured of that assistance; and grieves whenever he observes the contrary, because he is then assured of their opposition” (*TMS*, p. 13f). If our passions (and thereby our evaluations and judgments) are confirmed by others, then we feel good, although the passions as such can result from sorrows and pain.

Smith’s moral sentiments require a “moral community of common understanding and experience” (Darwall 1999, p. 160). Given such a community, it seems that, at least in the short run, we have no control of the pleasure of sympathy and the pain of opposition as they “are always felt so instantaneously, and often upon such frivolous occasions, that it seems evident that neither of them can be derived from such self-interested consideration...” (*TMS*, p. 14). Obviously, the range of sympathizing is limited and so is the reach of moral communities. (See Holler and Leroch (2008) for a discussion.) As a consequence, “an irreducible plurality of virtues and corresponding norms” results (Sturm 2010, p. 269). It is a function of the law to propose general (obligatory and universal) standards for a society and to implement them.

Promises and Subsistence

The domain of the impartial spectator and its generalization also decides on whether we observe an injury, or not. In *LJ*, Smith argues that a contract implies a promise and “the promise produces an obligation, and the breach of it is an injury” (*LJ*, p. 472). However, the “breach of a contract is naturally the slightest of all injuries...in rude ages crimes of all kinds, except those that disturb the public peace, are slightly punished, and society is far advanced before a contract can sustain action or the...breach of it to be redressed” (*LJ*, p. 472). According to Smith the probability of injury involved in the breaching of contracts is a result of the economic

environment, the social interaction, and the rationale of the agents: “When a person makes perhaps twenty contracts in a day, he cannot gain so much by endeavouring to impose on his neighbour, as the very appearance of a cheat would make him lose. When people seldom deal with one another, we find that they are somewhat disposed to cheat, because they gain more by a smart trick than they can lose by the injury which it does their character” (*LJ*, p. 538f). Note that this argument concurs with the Folk Theorem, one of the major results of modern game theory for iterated prisoners’ dilemma situations with multiple players. Tullock (1985) applied this model to corroborate Adam Smith rationale.

Does injury extend to failures to achieve subsistence and result in sentiments that trigger redistribution? Witztum (1997, p. 255) argues that Smith “would agree with considering the failure of distributing subsistence as a violation of justice in its commutative sense (i.e., requiring positive reprisal)” as it “will create harm and injury, not only through labourer’s frustrated expectations, but through physical and mental hardship as well” (p.256). In his paper “Adam Smith on justice and distribution in commercial societies,” Salter (1994) further elaborates this argument.

Ziegler (2014, p. 41) argues “that, for Smith, distributive justice exceeds the instrument of transferring benefits – he considers it to be purely complementary and not a necessity. The system of natural freedom and freedom of trade that he propagates also provides a certain state of distributive justice that is accompanied by an increase in wealth for all individuals.” Smith’s definition of justice implies abstaining from taking what others already rightfully and legally possess. However, what is meant by rightfully and legally is subject to the impartial spectator. Given the violation of subsistence and the sentiments which it arouses, Witztum (1997, p. 248) concludes that “both familiarity and universality of such sentiments in the age of commercial society would have guaranteed the impartial spectator’s approval of their resentment and would demand an enforced recompense.” Witztum and Young (2006) claim that the right to subsistence is included in Smith’s extended notion of justice.

One should add that Adam Smith was quite aware that subsistence by appropriating the fruits of labor is not guaranteed even in societies of some surplus with the consequence of “destroying a great part of the children which their fruitful marriages produce” (*WN*, p. 98). Consequently, Witztum (1997, p. 259) concludes, “. . . such an organisation of society will clearly be unjust.” One might add, especially “when masters combine together in order to reduce the wages of their workman. . . Were the workmen to enter into a contrary combination. . . the law would punish them very severely” (*WN*, p. 158). In Adam Smith’s writings we do not find a recipe of how to overcome this injustice (if not by economic growth which makes labor scarce and pushes wages above subsistence) – although such violations of impartiality “earn the resentments of the impartial spectator, and so fail the test of justice” (Buckle 2013, p. 100).

An implication of Witztum’s reasoning is that a possible redistribution can be derived from commutative justice (and not distributive justice) if the impartial spectator is universal enough to support such a policy. However, in an earlier paper (Holler 2006), I argued that the impartial spectator is of local character, as it is based on the reference to family members, friends, members of the community, etc. In general, the various groups of a society have quite distinct impartial spectators, implying different modes of behavior (while the organization of the market economy is rather general, even transgresses national borders). This raises the question of consensus: Is there an impartial spectator that considers society as an agent and the failure to achieve subsistence as its responsibility? The social security systems of Western Europe no longer accept unemployment as individual fortune. Moreover, there is a rather lively discussion of basic income which could be considered as the consequence of an expansion of the impartial spectator.

Selected Cases and Prominent Issues

In many societies usury seems to be a challenge to the impartial spectator as well as economic reasoning (and liberty). In many cultures and legal

systems, the interest on credits has been prohibited by civil, penal, or canonical law. However, as Smith (*WN*, p. 356) observes, this “regulation, instead of preventing, has been found from experience to increase the evil of usury; the debtor being obliged to pay, not only for the use of the money, but for the risk which his creditor runs by accepting a compensation for that use. He is obliged, if one may say so, to insure his creditor from the penalties of usury.” Smith favors a regulation that fixes an upper limit to the interest and thus bans usury. He argues that this rate “ought always to be somewhat above the lowest market price, or the price which is commonly paid for the use of money by those who can give the most undoubted security.” If such a limit does not exist, a “great part of the capital of the country would thus be kept out of the hands which were most likely to make a profitable and advantageous use of it, and thrown into those which were most likely to waste and destroy it. Where the legal rate of interest, on the contrary, is fixed but a very little above the lowest market rate, sober people are universally preferred, as borrowers, to prodigals and projectors. The person who lends money gets nearly as much interest from the former as he dares to take from the latter, and his money is much safer in the hands of the one set of people, than in those of the other. A great part of the capital of the country is thus thrown into the hands in which it is most likely to be employed with advantage” (*WN*, p. 357). Obviously, the consequences of usury constitute a version of adverse selection, i.e., of market failure: Without a legal upper limit, the interest rate could reach such a level that only those will borrow money who do not intend to pay back the loan, or in the end will be unable to pay the interest while “sober people” will be crowded out – of course, with negative effects to overall economic performance.

Note that Smith does not talk about justice, when arguing against usury. It is a matter of economic reasoning, i.e., taking efficiency into account, that he uses to suggest its banning. Economic reasoning is also his justification for tight regulation of the banking sector, even at the cost of a violation of natural liberty. “To restrain private people, it may be said, from receiving in

payment the promissory notes of a banker, for any sum whether great or small, when they themselves are willing to receive them, or to restrain a banker from issuing such notes, when all his neighbours are willing to accept of them, is a manifest violation of that natural liberty which it is the proper business of law not to infringe, but to support. Such regulations may, no doubt, be considered as in some respects a violation of natural liberty” (*WN*, p. 324). However, the analysis of the banking sector told him that it needs strict regulations in order to limit the circulation of bank notes and to avoid crises that are disastrous to the economy and thus for the country. He concludes: “But those exertions of the natural liberty of a few individuals, which might endanger the security of the whole society, are, and ought to be, restrained by the laws of all governments, of the most free as well as of the most despotical. The obligation of building party walls, in order to prevent the communication of fire, is a violation of natural liberty exactly of the same kind with the regulations of the banking trade which are here proposed” (*WN*, p. 324).

Smith “defends the system of free trade, but he does so because of the benefits it delivers, not because of the inviolability of individual self-interest” (Buckle 2013, p.100). If natural liberty is considered a subject of natural rights, then, in the case of regulating banking, Smith proposes a violation of natural right in favor of economic efficiency and the “wealth of a nation.” The regulation of the banking sector, as proposed by Smith, is an illustration of *second-best liberalism*. “Creating and maintaining the conditions for the system of natural liberty tends to imply a *constant challenge*, including a set of complex tasks, institutions and practices” (Sturm 2010, p. 276). This challenge may necessitate putting a limit to liberty.

However, Smith also observes the economic forces of self-interest at work when it comes to the implementation of the law. Here is an example. “The landlord and tenant, for example, might jointly be obliged to record their lease to a public register. Proper penalties might be enacted against concealing or misrepresenting any of the conditions; and if part of those penalties was to be paid to either of the two parties who informed against

and convicted the other of such concealment or misrepresentation, it would effectually deter them from combining together in order to defraud the publick revenue. All the conditions of the lease might be sufficiently known from such a record" (*WN*, p. 831).

On the other hand, Smith is quite aware that the law does not always contribute to the common good, and in some cases, it simply fails as it cannot be executed. "People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the publick, or in some contrivance to raise prices. It is impossible indeed to prevent such meetings, by any law which either could be executed, or would be consistent with liberty and justice. But though the law cannot hinder people of the same trade from sometimes assembling together, it ought to do nothing to facilitate such assemblies; much less to render them necessary" (*WN*, p. 145). In Smith's time assemblies were rendered necessary because of a regulation "which enables those of the same trade to tax themselves in order provide for their poor, their sick, their widows and orphans" (*WN*, p. 145). Today, these obligations do no longer exist, but it seems that those of the same trade still meet.

On Human Nature

There is hardly a more controversial writer in the history of economics and the social sciences than Adam Smith inasmuch as very different "social religions" identify their roots with his work. By many he is hailed as a prophet of free markets while others point to his elaborate moral theory and the fundamental social dimension of his thought (and life). One might argue that this is due to the fact that one group studied *TMS*, while the other had his *WN* at their bedside. It is very likely that only a very small percentage of those who refer to Adam Smith to sell their ideas about how to improve the world have actually read either of these books; and certainly very few have read both – let alone having read his correspondence or *LJ*.

However, it might be superficial to blame the potential readership for not adequately accomplishing its duty of reading when it comes to Adam Smith. If there is an obvious shortcoming, then there could also be a reason for this. A possible answer to this puzzle could be that Adam Smith proposed at least two "models of man:" the self-interested individual that forms the core of his economic theory, the dominant character in the *WN*, and the moral agent controlled by the sympathy, sentiments, and impartial spectator, introduced in his *TMS*. It has been argued that the latter does perhaps not reach far enough to evaluate the outcome which corresponds to his economic theory from a *social* point of view. (See the issue of subsistence, discussed above.) Human nature could be the source of this shortcoming: Sympathizing seems to be biased. "It is because mankind are disposed to sympathize more entirely with our joy than with our sorrow, that we make parade of our riches, and conceal our poverty. Nothing is so mortifying as to be obliged to expose our distress to the view of the public, and to feel, that though our situation is open to the eyes of all mankind, no mortal conceives for us the half of what we suffer. Nay, it is chiefly from this regard to the sentiments of mankind, that we pursue riches and avoid poverty" (*TMS*, p. 50).

But there is also "the desire of bettering our condition, a desire which, though generally calm and dispassionate, comes with us from the womb, and never leaves us till we go into the grave. . . there is scarce perhaps a single instant in which any man is so perfectly and completely satisfied with his situation, as to be without any wish of alteration or improvement, of any kind" (*WN*, p. 341).

Smith refers to this desire to explain the motivation for the accumulation of capital by means of saving. It represents a disposition of "man" which follows him throughout his life, "though generally calm and dispassionate." However, there is another motivation for the accumulation of capital: to become *rich*. "In Smith's view. . . man is inspired in commercial society to pursue the rewards of sympathy and the approbation through the means of acquiring wealth and property at the

expense of seeking sympathy through virtuous behaviour. The pursuit of wealth thus tends to crowd out virtue as a means to be taken notice with sympathy” (Verburg 2000, p. 38). The argument here is that there is no theory of social justice which permits us to evaluate the overall economic outcome (see Huber 2014). The implications and consequences of the economic system narrow down the application of the individualistic moral theory embedded in the impartial spectator and corrupt its standards. It seems that there is no “Adam Smith problem,” as was widely discussed in the German nineteenth-century literature (see Sturn 2010), but a problem in the human nature: being very general and broad when it comes to economic efficiency and rather specific and narrow when thinking in terms of justice.

Acknowledgments I would like to thank Heinz Kurz, Barbara Klose-Ullmann, Martin Leroch and Rigmor Osterkamp for their helpful comments.

References

- Buckle S (2013) Hume and Smith on justice. In: Routledge companion to social and political philosophy. Routledge, New York/Milton Park, pp 92–102
- Darwall S (1999) Sympathetic liberalism: recent work on Adam Smith. *Philos Public Aff* 28:140–163
- Holler MJ (2006) Adam Smith’s model of man and some of its consequences. *Homo Oecon* 23:467–488
- Holler MJ, Leroch M (2008) Impartial spectator, moral community and some legal consequences. *J Hist Econ Thought* 30:297–316
- Holler MJ, Leroch M (2010) Efficiency and justice revisited. *Eur J Polit Econ* 26:311–319
- Huber K (2014) Adam Smith’s model of man. In: Rummelshagen S, Stobbe M (eds) Adam Smith meets Walras and Machiavelli. Accedo, Munich, pp 7–26
- Hume D (1978/1739) In: Selby-Bigge LA (ed) A treatise of human nature (rev Nidditch P), 2nd edn. Clarendon Press, Oxford
- Leroch MA (2008) Adam Smith’s intuition pump: the impartial spectator. *Homo Oecon* 25:5–26; republished in: Rummelshagen S, Stobbe M (eds) Adam Smith meets Walras and Machiavelli. Accedo, Munich, pp 99–122
- MacCormick N (1981) Adam Smith on law. *Valpo Univ Law Rev* 15(2):243–263
- Salter J (1994) Adam Smith on justice and distribution in commercial societies. *Scott J Polit Econ* 41:299–313
- Smith A (1981/1776/77) An inquiry into the nature and causes of the wealth of nations. In: Campbell RH, Skinner AS (eds) The Glasgow edition of the works and correspondence of Adam Smith, vol II. Liberty Press, Indianapolis, quoted as WN
- Smith A (1982/1759/1790) The theory of moral sentiments. In: Raphael DD, Macfie AL (eds) The Glasgow edition of the works and correspondence of Adam Smith, vol I. Liberty Press, Indianapolis, quoted as TMS
- Smith A (1982/1762–1766) Lectures on jurisprudence. In: Meek RL, Raphael DD, Stein PG (eds) The Glasgow edition of the works and correspondence of Adam Smith, vol. V. Liberty Press, Indianapolis, quoted as LJ
- Sturn R (2010) On making full sense of Adam Smith. *Homo Oecon* 27(3):263–287
- Tullock G (1985) Adam Smith and the Prisoners’ dilemma. *Q J Econ* 100(Issue Supplement):1073–1081
- Verburg R (2000) Adam Smith’s growing concern on the issue of distributive justice. *Eur J Hist Econ Thought* 7:23–44
- Witztum A (1997) Distributive considerations in Smith’s conception of economic justice. *Econ Philos* 13:241–259
- Witztum A, Young JT (2006) The neglected agent: justice, power, and distribution in Adam Smith. *Hist Polit Econ* 38:437–471
- Ziegler M (2014) Adam Smith on distributive justice. In: Adam Smith meets Walras and Machiavelli. Accedo, Munich, pp 7–26

Social Damage

Stephan Marette

UMR Economie Publique INRA, Paris, France

Definition

The social damage is a damage suffered by the society as a whole because of corruption, bad use of public money, or wrongdoing. The social damage includes the stolen assets, the wasted money that could be alternatively spent by the society, and also ethical dimensions, like moral damages or violation of trust regarding public values.

A Nascent Concept

The social damage is a damage suffered by the society as a whole and differs from the private

damage that is incurred by a private entity. The possibility to pay a compensation to the victim who is the “society as a whole” recently emerged.

The social damage was directly mentioned in the settlement agreement related to corruption between Alcatel and the Costa Rica. In 2010, this French firm agrees to compensate for the damage coming from bribes and caused to citizens. Alcatel paid US\$10 million for compensating this social damage. This settlement does not impede additional compensations to other victims, like private compensation to ousted competitors (Olaya et al. 2010).

A lot of attention was given to this settlement agreement in Costa Rica by people or organizations who search to fight corruption (see Olaya et al. 2010 for details). The social damage was compensated in Costa Rica, because of an explicit legislation mentioning this precise topic, which is a very important point for implementing legal decision and compensation. In a 2010 workshop (in page 2 in IACC 2010), Mr. Eckert, a German judge, mentioned “that he always aimed at compensating social damages. However, he also made it clear that Germany had no definition of social damage in its penal code. Judge Eckert concluded by stating that there is a need for rules on social damage.”

The absence of explicit legislations defining social damages in many countries limits the possibility to effectively compensate society for this damage. However, the *United Nations Convention Against Corruption* signed by many countries in 2004 emphasizes the right to restitution of stolen assets to society (see United Nations 2004). This is a clear stage toward the recognition of the social damage, even if this concept is not explicitly mentioned in the convention. This nascent concept could be used for deterring firms and individuals of wrongdoing related to public goods.

An Economic Interpretation

As this concept is nascent regarding the victims’ compensation and not fully defined in many legislations, we turn to the economic approach to

help clarify this concept. The social damage can be accurately detailed by using the concept of public goods.

While private goods are traded in markets with prices reflecting quality levels, public goods are non-excludable and non-rival in consumption. As a consequence, it is often optimal for agents to let others contribute and benefit from this public good without contributing to its production or financing. The payoff structure reflects a prisoner dilemma trade-off coming from game theory, such that the dominant strategy is to invest nothing (or cheat) in the public good if they expect enough other participants decided to invest (or being honest). Public goods are rife and they include all social values related to social damages such as trust, honest administration, legitimacy of institutions, and fair competition.

The concept of public good may help to value the economic loss coming from a social damage, even if this is relatively difficult. For the corruption case, the economic loss can be estimated by comparing the collective utility for a situation with the social damage compared to the collective utility without the social damage. A so-called equivalent variation for maintaining a collective utility without social damage compared to a case with corruption could be estimated. New techniques like lab experiments focusing on corruption could give a social value relative to the absence of corruption, by extrapolating the elicited value of a small group of people attending the experiment to the whole population (see Bobkova and Egbert 2012). However, the empirical estimate of such a loss is difficult because of the nonmarket nature of public goods with no observable prices that help determine social values.

The other contribution of the economic approach consists in evaluating various legal and regulatory measures, for limiting the social damage and understanding agents’ incentives to avoid behaviors leading to social damages. This particularly raises the question of the optimal effort and sanction for deterring potential wrongdoers. When the social damage is not recognized by a legislation and de facto not compensated, there is an incentive to lean toward bad behaviors and an underinvestment in virtuous practices.

Eventually, an overlooked aspect of having private damages and social damages to be compensated for a same wrongdoing such as corruption is the risk of accumulating reparations. This accumulation of compensations may reinforce the judgment-proof problem, for which the potential defendant at the origin of damages is financially insolvent and unable to fully cover compensations. The judgment-proof problem modifies incentives to have virtuous behaviors compared to the socially optimal level maximizing the welfare of the whole economy (see Shavell 2005). In other words, the judgment-proof problem may lead some firms to adopt corruption behavior, while the society would be better off without any corruption behavior.

The concept of social damage is important for preventing cheating and bribes. This is also a fragile concept because the social damage is relatively hard to measure.

References

- Bobkova N, Egbert H (2012) Corruption investigated in the lab: a survey of the experimental literature. MPRA paper 38163, University Library of Munich, Germany
- IACC (2010) Finding the real cost of corruption: how to use the concept of social damage for the anti-corruption struggle. In: 14th international anti-corruption conference, long workshop report form, Thailand, 12 Nov 2010
- Olaya J, Attiso K, Roth A (2010) Repairing social damage out of corruption cases: opportunities and challenges as illustrated in the Alcatel case in Costa Rica, 6 Dec 2010. Available at SSRN: <http://ssrn.com/abstract=1779834>
- Shavell S (2005) Minimum asset requirements and compulsory liability insurances as solutions to the judgment-proof problem. *Rand J Econ* 36(1):63–77
- United Nations (2004) United Nations convention against corruption. United Nations, Office on Drugs and Crime, Vienna. Available at http://www.unodc.org/documents/treaties/UNCAC/Publications/Convention/08-50026_E.pdf

Social Preferences

- [Intrinsic and Extrinsic Motivation](#)

Spousal Support

- [Alimony](#)

Standard Cost Model

Wolfgang Weigel

Faculty of Business, Economics and Statistics,
Department of Economics, Joseph von
Sonnenfels Center for the Study of Public Law
and Economics, Vienna, Austria

Definition

The Standard Cost Model is a tool for the reduction of administrative burdens of businesses. It rests on a fairly simple technique of identifying such activities which are found unnecessary by experts and thus have the potential of savings on expenses. The economic idea behind its application is that means are unleashed which can be used more efficient and effective, thus fostering growth. However, the tool also has a couple of shortcomings, which presently give rise to the development of a new generation of SCM (SCM 2.0).

Introduction

The Standard Cost Model (SCM) is a widely used remedy against certain kinds of red tape. It belongs to the family of devices for less and better regulation, nowadays handled under the more handy label “smart regulation” (see: <http://www.oecd.org/regreform/policyconference/46528683.pdf>.) This model aims at identifying and measuring the administrative burdens of businesses. Thus, it is an indispensable prerequisite for steps towards the reduction of administrative burdens. This is

Wolfgang Weigel was retired.

partly accomplished by simplifying obligations to comply with regulatory demands of governments and partly by checking the possibilities of abolishing them altogether. One rationale behind this measure is the gain of resources, which can be spent more productive, as suggested with respect to the annual World Bank report “doing business,” where objective measures are provided for local firms regarding the intensity of regulation (see: The World Bank (<http://www.doingbusiness.org>)).

Originally developed in the Netherlands as “MISTRAL” (Measuring Instrument Administrative Burden, where burden in Dutch is Last) between 1992 and 1994, it is by now used under its more popular label in more than 20 states. It forms part of a set of four instrumental tools for scrutiny of legal rules, where the other tools are cost benefit analysis, cost effectiveness analysis, and multicriteria analysis (asking what kind of action produces the largest expected utility).

This note provides an overview of the methodology, primarily focusing on businesses but also pointing out the attempts to apply the SCM to households. It starts with a review of the model, addresses problems and shortcomings and the way towards a SCM 2.0, followed by a short outlook.

Review of the Standard Cost Model (SCM)

The SCM aims at identifying, measuring, and assessing the cost accruing to businesses when complying with administrative demands. Such costs typically arise from obligations to provide information to a central authority (the federal government, say) or the public (customers, for example).

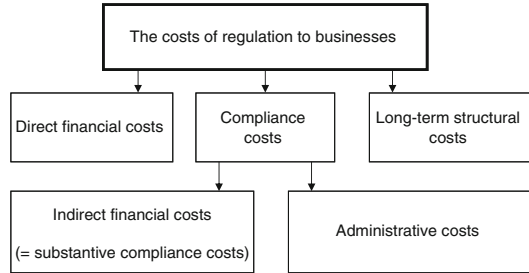
At the beginning, the outline of the model will be reproduced following the International Standard Cost Model Manual (<http://www.administrative-burdens.com/default.asp?page=122>, January 24th, 2008, 11.27).

In order to accomplish this task, a couple of charts will be reproduced. The rationale behind this is to give an impression of the distinction

between such costs, which have to be borne due to indispensable institutional regulations (the most prominent examples being taxes and fees) and those costs which are “administrative burdens” (the said obligations to provide information).

The first chart provides a general overview.

Chart 1: The different costs of regulation to businesses



The content of the entries is as follows.

Direct financial costs comprise all obligations for financial flows to government. The outstanding examples are taxes. However, businesses may also be obliged to fees for receiving certain permits, which are also seen as direct financial charges according to the International Standard Cost Model Manual (p. 6).

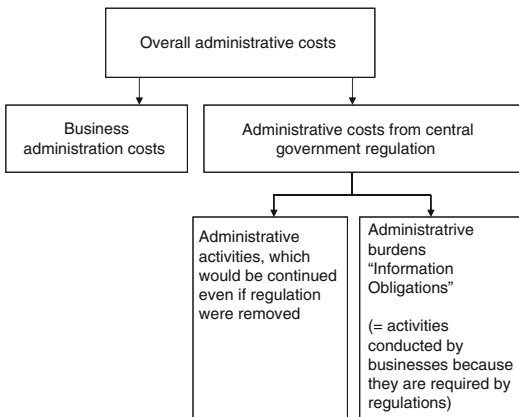
Compliance costs are all other costs with the exception of those arising from long-term structural consequences. Compliance costs can be broken down into:

- Substantive compliance costs, which contain, for example, the installation of filters or physical facilities in order to meet workplace regulations such as safety devices.
- Administrative costs, which are in the focus here. They accrue from the obligation to report about safety installations, say, or annual reports on the accomplishment of environmental standards (and many more, of course).

The latter have to be examined more closely. More specifically, administrative burdens have to be distinguished as the subset of administrative costs, which is relevant here (see Weigel 2008).

Again, the gist can be captured in a chart.

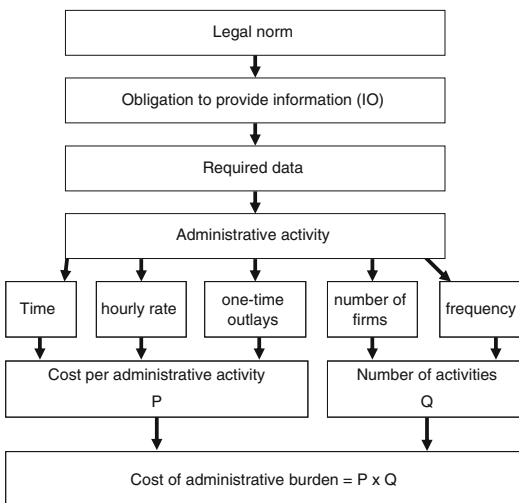
Chart 2:



While the distinction of the contents of the boxes should be evident, there is one caveat. It may be the case that some of the “administrative burdens” could be continued as deliberate administrative activities, even if the obligation were removed (such as information, which ultimately is part of a marketing strategy). However, it is very hard and laborious to make that distinction *ex ante*. It therefore is neglected.

In the next step, again making use of a chart, the way is outlined, how administrative burdens are calculated.

Chart 3:



In preparing the basis for the calculation of the cost of administrative burdens, a sophisticated

procedure is elaborated, which includes the solution of problems such as the assignment of some Information Obligations (IO) to the source, which can be European law, national law, or local regulations. Only norms of EU and the national level are taken into account (See International Standard Cost Model Manual, p. 27). Another detail worth mentioning here is whether an IO was met by the firm or had been outsourced (See International Standard Cost Model Manual, p. 35).

The whole procedure is described in detail in the International Standard Cost Model Manual (pp. 20 *passim*), and the interested reader is referred to this source.

The most important and controversial step in the course of investigations is that of determining the time and money cost at going wage rate for the cost per administrative activity. They are not raised firm by firm but rather standardized assuming a “normally efficient” firm.

The idea behind this notion is that in lieu of the time consumed for IOs as collected for each firm in the sample a representative value is chosen, which then is valued at the going wage rate (if performed by an employee, but disregarded if outsourced!) in order to arrive at the cost P according to chart 3. The core idea here is that outliers are not included in the calculation. To illustrate: some IO might take 10 min time in firms A, B, and C, whereas it takes 30 min in firm D. Then, the representative value is 10. If there was an additional firm in the sample with a performance of 2 min, this would also be excluded. However, in cases where the spread between best and lowest value becomes very large the reasons might be scrutinized by further interviews. Still the findings will not enter the outcome, since this will be taken from the average, but neither from the high nor the low performer!

Firms, which are close to the representative figure, are consequently labelled as “normally efficient.” Note that the further calculations rest only on data for “normally efficient” firms.

Surveys are carried out by experts. Following the procedure in Austria as an illustrative example experts are supported (and controlled) by special teams which are installed for each of the

ministries, for which information requirements have been identified from the law.

When the calculation was first made in Austria, for example, 561 legal provisions were identified containing 5687 information obligations. The total administrative burden was calculated at 4.6 billion € or 1.6% of GDP!

The target then was set at bringing the burden down by 25% by identifying and removing appropriate IO. Based on the insights from the World Bank's "Doing Business," the measure is supposed to push the growth of the economy considerably.

Properties, Shortcomings, and the Path Towards SCM 2.0

The following features of the SCM need to be stressed. They partly give rise to criticism and subsequently to the development of SCM 2.0, which is under way.

The first and foremost problem with the model is the implicit assumption that all legally mandated information obligations are met, both to the required extent and quality. To the contrary, large parts of the economics of property law, tort law, contracts, governance, environmental law, the economic theory of bureaucracy, and specifically the theory of corruption are based on the assumption that the agents seek their advantage under given constraints, boiling down to the neglect of regulatory demands. From the viewpoint of the regulators, no enforcement and hence no enforcement costs are taken into account. Thus, the actual administrative burdens may be lower than those calculated.

It must not be overlooked that the information obligations under scrutiny are not the only cost accruing to businesses from regulatory demands. They might face considerable compliance costs, which are rooted in overly complicated regulations, "transaction costs" of meeting the requirements through outsourced means such as expertise to appropriately understand demands and economize on them. Such considerations imply that the exercise of reducing administrative

burdens falls short of the objective. One reason could be the difficult wording or ambiguities in laws.

Another important issue relates to the rationale for regulation. In theory at least, it is rooted in some perceived market failure. Of course, regulatory requests can at the same time be a consequence of rent-seeking or simply an outflow of power. So the crucial step of deciding which IO are to be removed must not be carried out without scrutinizing the reasons for the existence of the IO at hand.

There are also problems with the "normally efficient firm". Since the extreme values of time consumed in performing IOs (both highs and lows) are ruled out by definition, distortions of the findings caused, e.g., by the presence of substantial X-inefficiency (Leibenstein 1966) within firms are not taken into account. But cases, which could serve as best practice, are also disregarded. Moreover, the way in which firms handle their IOs are also of interest. Do they employ special staff or is it done as additional business, which in turn implies that the person capable of such additional business may have comparatively high qualification.

Both points are relevant for the overall purpose of the entire exercise: to create a boost for the economy. With X-inefficiency, the relief will not become effective, since the firms by definition are not on their production possibility schedules; they may not even know them.

Likewise, firms upon relief from IOs may save on staff and not necessarily use the saved means for expansion. So the intended effects of the whole undertaking are quite uncertain.

More general objections are the following:

- In order to effectuate the objective of the SCM, it is not just necessary to enable firms to save on administrative costs but to keep them (or bring them back to be) competitive. Without such focus, the tool will create windfall profits for firms only.
- So it is not just the relief from burdens but rather the market structure, which is at stake. This in turn requires a more general focus, which includes the quest for regulation, an

issue, which is not covered by SCM, as has been mentioned earlier.

- A more general view such as that taken by Regulatory Impact Assessment (RIA) would head for maximizing the net benefits from regulation, where benefits are plotted against regulatory costs at large. SCM instead concentrates (and proofs its use) on areas of high inefficiency due to said administrative burdens only.

For the sake of completeness, it ought to be pointed out that after the first round of applications considerable effort was spent on finding out weaknesses and shortcomings in the handling of the procedure. That is to say, the institutional framework and the organization of handling the SCM was also under scrutiny (see Nijsen et al. 2009; Torriti 2007).

Thus, proponents of the SCM are fully aware of the shortcomings in what is now labelled “SCM 1.0” and therefore move towards SCM 2.0.

It is worth noting that attempts are under way to extend the SCM to private households. However, the calculus of cost per activity times frequency is hardly applicable. So research concentrates on effort in hours as well as direct costs. Moreover, qualitative aspects are considered, presently in a pilot study for Austria, for example, with the top 100 information obligations for households.

Outlook

The Standard Cost Model is a comparatively straightforward tool for the reduction of certain administrative burdens, which are consequence of governmental regulatory obligations to businesses. Given its relative simplicity, it has quickly become quite popular among governments. Said administrative burdens are quite high according to calculations in a large number of OECD member-states. To illustrate: the total burden for Austria was calculated to be 4.3 billion € (1.66% of nominal GDP). This burden is equal to those found in the UK (1.5%) and Denmark (1.9%), lower than that in the Netherlands (3.7%) but higher than that in the USA (about 1%),

according to a Review of the World Bank Group. To make these means accessible for productive purposes contains a substantial potential for economic growth (according to “Doing Business,” actual year).

The problem with what nowadays is labelled SCM 1.0 is that it may not provide accurate figures, since it rests on assumptions and procedures, which very likely lead to an overestimate of the gains. Moreover, it does not question regulation at large and obviates cost-benefit analyses, which are at the core of the more comprehensive approach of regulatory impact assessment (see Wegrich 2009).

Nowadays, these shortcomings are well understood and a SCM 2.0 is under way, which is to overcome the main sources of dissatisfaction with the first generation of the model.

Cross-References

- [Information Disclosure](#)
- [Public Interest](#)
- [Regulatory Impact Assessment](#)

References

- <http://www.oecd.org/gov/regulatory-policy/34227698.pdf>. 7 Dec 2017
- <http://ec.europa.eu/eurostat/documents/64157/4374310/11-STANDARD-COST-MODEL-DK-SE-NO-BE-UK-NL-2004-EN-1.pdf/e703a6d8-42b8-48c8-bdd9-572ab4484dd3>. 7 Dec 2017
- Leibenstein H (1966) Allocative efficiency vs. “X-efficiency”. *Am Econ Rev* 56(3):392–415
- Nijsen A et al (eds) (2009) *Business regulation and public policy*. Springer, New York
- Torriti J (2007) The Standard Cost Model: When better regulation fights against red-tape (March, 18 2009). In: Weatherill S (ed) *Better regulation*. Hart, Oxford. Available at SSRN: <https://ssrn.com/abstract=1363986>
- Wegrich K (2009) The administrative burden reduction policy boom in Europe: comparing mechanisms of policy diffusion. CARR discussion papers, DP 52. Centre for Analysis of Risk and Regulation, London School of Economics and Political Science, London. ISBN 9780853281443
- Weigel W (2008) The Standard Cost Model – a critical appraisal (September 17, 2008). 25th annual conference of the European Association of Law and Economics. Available at SSRN: <https://ssrn.com/abstract=1295861> or <https://doi.org/10.2139/ssrn.1295861>

State Aids and Subsidies

Régis Lanneau

Law School, CRDP, Université de Paris Nanterre,
Nanterre, France

Abstract

The power to give money is apprehended, in legal terms, through the notions of subsidies and state aids. In a conventional approach of state aids and subsidies control, it is because they have a potential effect on the competitive position of economic and create some trans-frontiers spillovers that these “spending” are or should be controlled at a supranational level. This entry will inquire into the rationale and limit of state aids control.

“The state – the machinery and power of the state – is a potential resource or threat to every industry in the society. With its power to prohibit or compel, to take or give money, the state can and does selectively help or hurt a vast number of industries” (Stigler 1971). The power to give money is apprehended, in legal terms, through the notions of subsidies (WTO, SCM agreement) or state aids (EU law), even if this latter notion is broader than direct subsidies (or financial contribution in WTO’s meaning even if this wording remains unclear, Wilcox 1999). Indeed, article 107 TFUE is characterizing state aids as “any aid granted by a Member State or through State resources in any form whatsoever which distorts or threatens to distort competition by favouring certain undertakings or the production of certain goods [...], in so far as it affects trade between Member States.” Case law is thus not restricting state aids to subsidies narrowly understood and includes loans, direct investment, tax reliefs, preferential tariffs, tax remission, free allocation of pollution rights, etc. Moreover, and in a conventional approach of state aids and subsidies control, it is because they have a potential effect on the competitive position of economic agents (because they are providing a benefit or an advantage) and create some trans-frontiers spillovers

(affect trade between member states) that these “spending” are or should be controlled at a supranational level. To strengthen this logic, both legal definitions are also including the idea of “specificity” (WTO) or “selectivity” (EU law) to distinguish between general measures (which are legal) and these legal categories.

In the remaining of this article, I will only focus on state aids, even if the framework which will be developed could largely be transposed to WTO’s subsidies (see Sykes 2003). Moreover, I will not provide a legal and economic analysis of each of the elements of qualification (advantage, using state resources, selectivity, affecting trade between member states) (for more details, see Röller et al. 2008). They are no doubt prone to interpretation (as ECJ case law would testify, see also European Commission 2008), and an economic analysis could provide some sound rationale at this level; nevertheless, a general framework for understanding state aids is a prerequisite for such analysis. This framework revolves around two questions. First, why do we need to control them at a supranational level? Second, how to design state aids control to enhance efficiency?

Why Should We Control State Aids at a Supranational Level?

If governments were only intervening in the economy to enhance efficiency (through the reduction of transaction cost and the reduction of market failures) and if that intervention would not lead to negative spillovers, controlling for state aids at a supranational level would not make sense since it will not allow for efficiency gains (quite the contrary actually because of the administrative costs it incurs). This does not mean that government failures are, by themselves, a sufficient reason to control for state aids at a supranational level (A) (even if, in some model, it would make sense; Dewatripont and Seabright 2006). The best rationale for this control seems to be trans-frontier spillovers (B) (Besley and Seabright 1999). Nevertheless, even in that case, only some spillovers should be targeted. Of course, both rationale could be and are combined to justify the actual system.

Limiting Rent-Seeking Rationale?

It is true that without any mechanism for controlling state aids, nothing, except maybe voters, would prevent governments to waste their resources to favor some successful interest groups, and there are few doubts that government spending decisions are not always promoting efficiency (Neven and Roller 2000). Knowing that it is possible to extract a rent through lobbying, interest groups will try to earn some “aids” and might even adapt their behavior even before any aid is granted. This rent-seeking approach of state intervention is quite classic (Stigler 1971; Posner 1974; Peltzman 1976) and largely documented (Persson and Tabellini 2000).

The mere prohibition of state subsidies or aids through, for example, a constitutional provision, would of course be inefficient since they are sometimes desirable; but a control at the level of each state – providing the possibility to use or design an efficient institution to achieve such end – could have made sense if the purpose is only to reduce rent-seeking. A control at a supranational level would not then seem to be required to limit the inefficiency effects of rent-seeking; after all, why would a supranational level institution be less prone to failure than national institutions? It is true that a supranational control would, in theory, insulate granting aids from political cycles (and, e.g., overspending before elections) but this rationale does not seem to be sufficiently strong to justify for a control. Indeed, in that case, the controlling institution will be perceived as “blocking” national will and might then be criticized for being technocratic – this could be privately welcomed and publicly criticized by national politicians; it is only if it appears that this institution is blocking more bad than good projects that it could earn some legitimacy (even if nationalist would, even in that case, criticize the institution as being anti-democratic). Moreover, it would be “too easy” to require another level of control to overcome known political failures.

It is true that if a supranational institution enjoys some expertise in this domain, using it could reduce administrative costs, but, once again, this logic is not sufficiently strong to justify for a systematic control of state aids.

Trans-frontier Negative Spillovers?

To justify the intervention of a supranational level, the effect of the state aids should not be strictly national (a parallel could be drawn using the subsidiarity model provided by Cooter 2000). If the aid is creating positive spillovers, it is unlikely to be criticized internationally; only perceived (but not always “real”) negative spillovers will then be targeted by litigants and institutions in charge for the control of state aids. Moreover, the potentially affected parties should agree to cooperate (which is the case for EU or WTO members; quite strangely, no control of state aids exists in the USA); otherwise, no supranational entity could exist.

There is little doubt that aids could be strategically used by states to attract companies and all their associated benefits in terms of taxes, employment, and more broadly economic dynamism. It would then be logical that, in the absence of such control, governments could compete using state aids (but not only). If governments were spending their resources efficiently, such competition would enhance welfare: the mere fact that a government is ready to spend more reveals that national spillovers will be higher than in other locations. This situation of course assumes that the national spillovers are varying across jurisdictions and that using the taxation system is not too costly. Note also that if the use of the tax system were zero, state aids would only have distributional effects and no efficiency effects (Spector 2009). Nevertheless, the use of the taxation system is costly so that transferring money to a firm costs more than the benefits the firm is receiving. If state aids were controlled, states would not waste resources to attract firms and avoid the prisoner’s dilemma game of subsidy races. Moreover, the distortions produced by an increase in taxes to insure this subsidy race are undesirable (Collie 2000).

State aid control is often justified by distortion of competition. Nevertheless, if the products are sufficiently homogenous and the market competitive, it is difficult to understand how its costs for competition would be higher than its benefits for consumers since it is unlikely to affect the market price in the long run. In an oligopolistic

structure, things could be different since the aid could lead to predatory pricing or end up with a merger between the recipient of the aid and one of its competitor. If products are differentiated, it is also unlikely that, on average, state aids will lead to inefficiencies: being differentiated, the aid will not have a tremendous impact on competitors but will benefit consumers.

How to Design State Aid Control(s) to Enhance Welfare?

To enhance welfare, it is first required to identify when a state aid is likely to be beneficial. These rationales could then be used in court to obtain their compatibility with EU law. Moreover, since a state aid control is costly, a special attention is required at its design. In here we will simply focus on state control at a supranational level. At a national level, the competitive neutrality principle could lead to substantial reduction of wasteful public spending. This principle could also be used at a supranational level, but it would then appear that the reason for the control is more paternalistic than really supranational.

Classical Justifications for Providing an “Aid”: Efficiency and Equity

Like other state intervention in the economy, state aids could enhance welfare when they are used to correct market failures. Compensation for the charges spent to fulfill public services obligations is probably the most known reason. In that case, if the money given to the firm is merely compensating the charges, it would not produce any advantage and would then be without any effect on competition or efficiency in the private sector. A state aid would also be beneficial if it is compensating for a positive externality produced by some entities (meaning that the entity will not have the possibility to capture all the benefits they are creating). This would be the case, for example, regarding research and development through knowledge spillovers and more generally through creating a company in a geographical area suffering, for example, of unemployment. Other rationale entails asymmetric information (and

especially some aids to risk capital to overcome the equity gap) or coordination failure (Neven and Verouden 2008).

EU law (article 107, (3) TFEU) also provides for the possibility to intervene for social or cultural reason. It, for example, refers to “aid to promote the economic development of areas where the standard of living is abnormally low or where there is serious underemployment” or “aid to promote culture and heritage conservation where such aid does not affect trading conditions and competition in the Union to an extent that is contrary to the common interest.” From a purely economic point of view, these reasons should find their justification in the positive externalities they are creating, and the efficient mean to achieve these ends should be considered (e.g., direct income transfers instead of aids to firms).

Designing an Efficient Framework

If economics could help to understand the logic of state aids, it reveals itself very limited at a practical level since it cannot provide clear-cut categories. It could nevertheless identify some key elements and trade-offs for aids controlled at a supranational level.

First, since enforcement is costly, it would make sense to allow for exemptions per categories or some per se condemnation of some subsidies or aids – these categories would then evolve with new theories and empirics. The WTO is doing that for aids regarding exports. EU law has developed “general block exemption regulation” (Council Regulation No 994/98 of 7 May 1998, amended by Council Regulation No 733/2013 of 22 July 2013) but no per se condemnation especially since Commission’s emphasis of the potential usefulness of state aid in improving the market outcome (European Commission, 2005, State Aid Action Plan, less and better targeted State aids). Moreover, for some exemptions (such as the one provided by 107 (3) TFEU), a balancing test is still required and the burden of the proof simply shifted. More interestingly, the standard of proof required is also playing a role. For example, in Philip Morris case (C-730/79), the ECJ showed a tendency to equate selectivity with distortion of competition. More recently, the necessity for an

empirical proof of the aid was questioned (T-479/11 and T-157/12). In any case, legal certainty is required to lower the administrative costs of state aid control (even though it will lead to a higher number of some false-positive and false-negative cases).

Second, a state aid control could be designed to work *ex ante* (through notification) or *ex post* (through litigation). The efficiency of one over the other remains unclear, and empirical data are lacking. Of course, the efficiency will depend, among other things, on the possibility to access courts, the time required for a decision to occur, and the efficiency of the *ex ante* screening.

Third, economic reasoning would recommend a differential depth of analysis depending on cases. For example, state aids in oligopolistic markets are, on average, more dangerous than in markets where products are heavily differentiated; requiring the same administrative burden to settle these situations is expected to be inefficient.

Conclusion

The economic analysis of state aids is a relatively new field with very few papers before 2000. Regarding the amount of researches in competition policy, this situation could be surprising. The fact that such a system does not exist in the USA could provide some information. Nevertheless, the fact that it is difficult to identify a simple set of core questions or core mechanisms, and the inherent complexity of state aid mechanisms requiring the use of a vast array of fields in economic theories, is probably delivering stronger explanation for the relative scarcity of researches. Considering its political importance in the EU context, this situation is bound to evolve.

Cross-References

- [Government](#)
- [Market Failure: History](#)
- [Public Choice: The Virginia School](#)
- [Regulatory Impact Assessment](#)
- [Rent Seeking](#)

References

- Besley T, Seabright P (1999) The effects and policy implications of state aids to industry. *Econ Policy* 34:13–53
- Collie D (2000) State aid in the European Union: the prohibition of subsidies in an integrated market. *Int J Ind Organ* 18:867–884
- Cooter R (2000) *The strategic constitution*. Princeton University Press, Princeton
- Dewatripont M, Seabright P (2006) ‘Wasterful’ public spending and state aid control, *J Eur Econ Assoc* 4(2/3), Papers and proceedings of the twentieth annual congress of the European economic association (Apr–May 2006), pp 513–522
- European Commission (2008) *Vademecum on community law on state aids*. Available online: http://ec.europa.eu/competition/state_aid/studies_reports/vademecum_on_rules_09_2008_en.pdf
- Neven D, Roller L (eds) (2000) *The political economy of industrial policy in Europe and the member states*. Edition Sigma, Berlin
- Neven D, Verouden V (2008) Towards a more refined economic approach in state aid control. In: Mederer W, Pesaresi N, Van Hoof M (eds) *EU competition law – volume IV: state aid*. Claeys & Casteels, Deventer
- Peltzman S (1976) Toward a more general theory of regulation. *J Law Econ* 19(2):109–148
- Persson T, Tabellini G (2000) *Political economics: explaining economic policy*. MIT Press, Cambridge
- Posner R (1974) The social costs of monopoly and regulation. *J Polit Econ* 83(4):807–828
- Röller L, Friederiszick H, Verouden V (2008) European state aid control: an economic framework. In: Bucirossi P (ed) *Handbook of antitrust economics*. The MIT Press, Cambridge
- Spector D (2009) State aids: economic analysis and practice in the European Union. In: Vives X (ed) *Competition policy in the EU*. Oxford University Press, Oxford
- Stigler G (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 2:3–21
- Sykes A (2003) *The economics of WTO rules on subsidies and countervailing measures*, Chicago John M. Olin Law & Economics working paper no. 186. Available online: http://www.law.uchicago.edu/files/files/186_aos_subsidies.pdf
- Wilcox W (1999) GATT-based protectionism and the definition of a subsidy. *Boston Univ Int Law J* 16:129

State Intervention in the Market

- [Public Investments: Broadband](#)

Stateless Law

► Customary Law

Statelessness

► Anarchy

State-Owned Enterprises

Michael Trebilcock

University of Toronto, Toronto, ON, Canada

Abstract

This entry reviews the performance of state-owned enterprises (SOEs) in both developed and developing countries, comparing performance with private sector counterparts in the same or other jurisdictions and their own performance before and after privatization. It then reviews alternative privatization modalities, including voucher privatizations, management buyouts (MBOs), privatization through the sale of state property by direct sale or public offerings, and partial privatization. It then reviews alternative reform strategies, including management contracts, performance contracts, policies to increase competition in sectors dominated by state-owned enterprises, and public-private partnerships (PPPs).

Trends in the Role of State-Owned Enterprises in Developed and Developing Country Economies

State-owned enterprises are sometimes referred to as “structural heretics” (Hodgetts 1971). On the one hand, they typically engage in commercial activities often associated with private enterprises. On the other hand, while owned by the government, they typically have some degree

of operational autonomy from the government, although subject to the government’s ability to appoint members of the Board of Directors and often senior management and approve major capital expenditures and sometimes long-term business plans.

SOEs were a relatively minor phenomenon in developed economies until World War II. While the share of mixed economies accounted for by state-owned enterprises had been on the rise since the early 1930s, the economic scope of the SOE sector at that time was fairly limited as such firms were concentrated mainly in a narrow range of industries or sectors, especially monopolistic public utility-type industries such as port authorities, railroads, and water and electricity distribution systems (Vernon and Aharoni 1981 at 8). However, the period following World War II witnessed a widespread expansion of the SOE sector in many developed countries, for a variety of reasons. SOEs were established in order to maintain employment in declining industries, to merge weak local firms, to prevent the acquisition of local firms by foreign enterprises, and to invest in industries for which private capital was unavailable. Prior to the onset of the privatization wave of the 1980s and 1990s, reflecting “supply-side” policies adopted throughout much of the world in the Thatcher-Reagan (Washington Consensus) era, the contribution of the SOE sector to GNP across the industrialized world was often in excess of 10 %: the share of GNP accounted for by SOEs was 22 % in Italy, 25 % in Austria, and more than 11 % in the UK. SOEs in developed countries dominated a number of strategic sectors of the economy, including energy, public transportation, communications, and in some countries forestry, iron, steel, and financial institutions (Aharoni 1986 at 1–2).

Similarly, the state sector in developing countries played an equally significant and in many cases a larger role in the economic life of these nations prior to the privatization movement of the 1980s and 1990s. Data from the 1970s reveals that the GDP share of nonfinancial SOEs fell within the 7–15 % range for most developing countries, although some countries in North Africa and Eastern Europe had state shares of GDP that approached 65 %. In several sub-Saharan African

countries, the share of wage employment attributable to SOEs ranged between 40 % and 74 %. Finally, in some developing countries, the industrial sector was almost synonymous with state-owned enterprise (*ibid.* at 17–25).

Despite extensive privatization of SOEs in both developed and developing countries in the 1980s and 1990s, they still play a significant role in many economies. A 2011 OECD working paper (Christiansen 2011) based on questionnaire responses from 27 of the 34 member countries covering the years 2008 and 2009 finds that while state ownership of enterprises has declined in recent decades, employment in state-owned organizations across the OECD still exceeds six million individuals, and the value of all SOEs combined is close to US \$2 trillion. In addition to the direct government ownership of these enterprises, in many countries the state holds minority stakes in listed enterprises that are significant enough to confer effective control on it. These enterprises, of which 54 were listed at the end of 2009, employ an additional three million people and are valued at close to US \$1 trillion. State-owned enterprises are increasingly concentrated in a few strategic sectors. Around half in value terms of all state-owned enterprises in OECD countries are located in the utilities sectors, such as transportation, power generation, and other forms of energy production and distribution, and a further quarter are accounted for by financial institutions (For the Canadian experience, see Iacobucci and Trebilcock (2012)). Among developed countries, the Thatcher Government in the UK in the 1980s initiated the most thoroughgoing privatization program, privatizing cable and wireless providers, coal, steel, gas, electricity, water, and airline and airport industries, automakers, railways, and shipbuilding (as well as public housing).

As a consequence of the privatization wave of the 1980s and 1990s, promoted by the World Bank, IMF, and other development agencies as an element of “the Washington Consensus,” developing countries have also witnessed a sharp reduction in most cases in the extent of the states’ involvement in these countries’ economies. For instance, as of 2003, 22 former communist

countries located in Europe and Central Asia possessed a private sector share of GDP that exceeded 50 % – up from only 9 % in 1994. Similarly, in China, the state share of GDP has dropped to less than 20 %, after accounting for almost 80 % in 1978 (Sunita Kikeri and Aishetu Kolo, “State Enterprises: What Remains?” World Bank. Public Policy for the Private Sector, note no. 304). However, data gathered by the World Bank suggests that on average, SOEs still account for between 20 % and 40 % of output in a number of many developing countries (Shapiro and Globerman 2012). In most sub-Saharan African countries, for example, state-owned enterprises still operate in many sectors of the economy and in aggregate account for more than 15 % of GDP, which is only a slight decrease since 1991. Similarly, state-owned enterprises continue to constitute a significant economic presence in many of the countries in the Middle East and North Africa, accounting for more than 50 % of GDP. In India, the state-owned sector produces 95 % of India’s coal, 66 % of its refined oil, 83 % of its natural gas, and approximately one-third of its finished steel and aluminum. In China, state-owned enterprises still employ almost half of the country’s 750 million workers, control 57 % of its industrial assets, and control a number of key industries, including electricity generation and distribution, financial services, and telecommunications (Kikeri and Kolo, *supra* note 7 at 2.).

Rationales for State-Owned Enterprises

While there is immense heterogeneity of experience in the developed world with respect to the creation of SOEs or the nationalization of previously private activities, Toninelli contends that the motives for such activity can be grouped into three broad categories (Toninelli 2000 at 5–9). First, there are political or ideological reasons for nationalization. In particular, historically, ideological motivations not only were fundamental to the nationalization policies of fascist and communist countries but were influential in the expansion of state involvement in countries such as France, Austria, the Netherlands, and the UK following

World War II. Nationalization programs instituted during this period were based on the belief that enlarging the public sphere could facilitate a fundamental redistribution of power within society, engendering a new equilibrium in which the power of labor was enhanced, and corporations would be responsible to the whole community for their decisions. Second, there were social motives for nationalization. These included the desire to guarantee full employment, to offer better working conditions to the labor force, and to pursue nation-building or reconstruction objectives. Finally, there were a number of economic reasons underlying the initiation of state activity, including addressing market failure in cases of natural monopoly, rescuing sectors of the economy in financial distress, and promoting economic growth in underdeveloped regions or sectors.

SOEs emerged in the developing world after World War II for a rather similar mix of political, ideological, and economic reasons. They were often formed (and subsequently expanded) to marshal support for new and fragile governments in developing countries. Developing country governments were also often influenced by the spread of socialist ideology following World War II, which dictated that the state should control the commanding heights of its economy (see, e.g., Haririan 1989 at 10; Muir and Saba 1995 at 11; Gillis 1980 at 263). Nationalism, often in the form of sovereignty concerns following decolonization from the mid-1950s onward, similarly fueled the desire for developing countries' governments to control the strategic sectors of their economies. SOEs were also advocated as a workable method of income redistribution – for example, by reducing prices of goods primarily consumed by the poor – especially in developing countries with large informal sectors in which income redistribution could not practically be accomplished through a tax-and-transfer system. The primary economic motivation behind SOEs was increased capital investment designed to address low domestic savings rates in many developing countries, which also often lacked the institutions, especially effective financial intermediaries, necessary to facilitate efficient capital allocations. Moreover, it was widely believed that developing

countries suffered from particular forms of market failures, such as low-level investment traps and a lack of basic infrastructure (see Smith and Trebilcock 2001 at 218, 219).

Evaluating the Performance of SOEs

A number of factors complicate the evaluation of the performance of SOEs. First, it is often difficult to determine or specify the objectives according to which performance should be measured. SOEs are often expected to achieve multiple and even conflicting economic, social, and political goals. The success of a firm could be assessed according to its internal efficiency or its contribution to social welfare, economic growth, or employment creation, to name a few examples. Economic theory would suggest that SOEs in general are likely to be less efficient, on conventional economic performance criteria, than private enterprises due to (1) the ownership effect and (2) the competition effect.

In the modern large corporation, there is a division between principals (the owners or shareholders) of the organization and the agents (the management and employees). The governance structure of the corporation such as the election of directors by shareholders (voice) or the ability to sell shares (exit) if disaffected with performance is designed to ensure that the principals (shareholders) have effective control over their agents' actions, and the interests of the agents are aligned with those of the principals so that agency costs are minimized. In addition, the market creates further managerial incentives to improve the organization's performance, as it rewards proven management skills and punishes inadequate performance through product market responses, markets for corporate control, markets for managerial talent, and bankruptcy laws (see Vickers and Yarrow 1998). In the context of SOEs, however, often both the governance structure of SOEs and the market have a much more attenuated impact on managerial performance. This is because SOE managerial appointments are often made on the basis of politics rather than merit and because SOEs are not subject to

traditional market risks such as takeovers or bankruptcy. In the private sector, shareholders often monitor the activities of a corporation, because they expect returns from their investments in monitoring. In the case of SOEs, the owner is the state, and thus, in some sense, each citizen can be considered a shareholder and will realize few personal gains from investing in monitoring efforts and cannot exit if disaffected with performance, while citizens as a whole face major collective action problems in coordinating their monitoring efforts. Further, because SOEs often face multiple economic and noneconomic objectives, it is difficult for SOE monitors to properly assess the performance of managers, who can blame poor enterprise financial performance on noncommercial objectives.

Another possible explanation for the inefficiencies of SOEs is the fact that, in the absence of natural monopoly, private enterprise encourages competition, whereas government ownership tends to encourage monopoly. Competition engenders allocative efficiency by removing the ability to set prices from the monopolist and placing it in the market. In theory, competition also engenders productive efficiency by encouraging producers to minimize costs. Where cost reductions permit producers to earn economic profits, new producers will enter the market, once again driving prices down to the marginal cost of supply. In these circumstances, constant attention to productive efficiency becomes a matter of an organization's very survival. In contrast, state-protected monopolists are not faced with the threat of being priced out of the market place and so do not have the same incentives to increase productive efficiency. However, the empirical evidence for both developed and developing countries does not conclusively support these theoretical predictions.

Mary Shirley and Patrick Walsh provide an extensive review of a number of these empirical studies (Shirley and Walsh 2000). Indeed, the 52 studies that are reviewed span the period from 1971 to 2001, use a variety of performance measures, and are balanced with respect to developing/industrialized nations. Additionally, a variety of competitive environments are

considered, from statutory monopoly to perfect competition. Of the 52 papers, 32 conclude that the performance of private firms is significantly superior to that of public firms, 15 find either that there is no significant relationship between ownership and performance or that the relationship is ambiguous, and five studies conclude that publicly owned firms perform better than private firms. When these studies are broken down by market environment, however, several important implications emerge in regard to the impact of ownership on enterprise performance. In particular, in both industrialized and developed nations, private enterprise superiority is evident in fully competitive markets. In fact, no study found that public ownership is superior to private ownership in such a case. However, this advantage is far less pronounced in monopolistic markets, where six studies find private superiority, five find neutral results, and five find public superiority (*ibid* at 51–52). These empirical findings are thus consistent with the conclusion drawn by Vickers and Yarrow that where firms face little product market competition and are extensively regulated, there is no generally decisive evidence in favor of either public or private ownership (Vickers and Yarrow 1991). Smith and Trebilcock provide an explanation for this finding, suggesting that, in highly regulated environments, the incentive and monitoring effects normally associated with private ownership are muted. More specifically, they suggest that excessive regulation inhibits the monitoring capacity of shareholders by allowing managers to hide poor management performance behind claims of distortionary government regulation (Andrew Smith and Michael J. Trebilcock 2001 at 225). Thus, while private enterprise is generally to be preferred on internal efficiency grounds, the empirical evidence suggests that ownership and competition both have a significant impact on firm efficiency and underscores that few gains can be expected from the privatization of state-owned enterprises into uncompetitive markets.

Another strand of empirical work compares the pre- and post-privatization performance of SOEs and finds that in most cases privatization has a positive effect on economic performance: divested firms mostly become more efficient,

more profitable, financially more robust, and increase their capital investment spending. While these results hold for transition and non-transition economies, the results are far more robust for upper- or middle-income countries than for low-income developing or transition economies (Megginson and Netter 2001 at 381). Indeed, much of the research focused on developing or transition economies finds that privatization generally increases social welfare when implemented in an appropriate institutional and regulatory environment, suggesting that unless privatization is coordinated with significant private sector reform, the lack of competition in low-income economies may result in minimal gains to society (Adam et al. 1992 at 71). David Parker and Colin Kirkpatrick argue that for privatization to have a significant impact on performance over the longer term, institutional capacity needs to be assessed prior to privatization to ensure that the scale, coverage, and sequencing of privatization are consistent with the available capital and management resources and that administrative capacities are in place to enable a privatization process that is fair, transparent, and efficient (Parker and Kirkpatrick (2005) at 535). Similarly, in their review of 34 empirical studies evaluating the effect of privatization on both macro- and micro-economic indicators in former communist countries, Saul Estrin et al. find that while privatization to foreign owners generally results in the rapid improvement of the performance of firms, the effects of privatizing to domestic owners are far less impressive and vary fairly significantly across countries (Estrin et al. 2009 at 29). Estrin et al. speculate that the performance differential between foreign and domestic owners is attributable to the fact that foreign firms often bring in strong expatriate managers, invest heavily in domestic training, and introduce more advanced systems of corporate governance, thus enabling them to compensate to a considerable extent for the underdeveloped legal and institutional system in many transition economies.

Finally, the importance of effective competition policies and appropriate regulation is of critical importance in the context of the welfare effects of privatizations in the infrastructure

sector. In theory, the gains in firm profitability and operational performance from selling an inefficient public sector monopoly to an unregulated private owner can easily be outweighed by the welfare losses imposed on consumers and the economy as a whole from suboptimal access to important products or services. While the existing empirical evidence has thus far concluded that divestiture substantially improves economic welfare on the whole by increasing investment and improving productivity, most privatizations have produced gains for some and losses for others. The distribution of these welfare effects therefore depends crucially on the fairness and the capacity of the regulatory system and effective competition and competition policies (Kikeri and Nellis 2004 at 97–100). Thus, in summary, the empirical evidence is consistent with the notion that that privatization can improve economic performance, but performance improvement relies also on other structural reforms.

Privatization

Assuming that privatization is desirable and that various economic and political prerequisites are to a greater or lesser extent in place, there are numerous methods of privatization that are available as models for implementation. In 1996, Josef C. Brada categorized various methods (Brada 1996 at 68), and, while not an exhaustive list, it provides a useful framework within which to understand the different ways in which privatization can be undertaken. Drawing on Brada's taxonomy, this section will survey four different types of privatization, assessing both the costs and benefits of each.

Vouchers

A form of privatization that has been used primarily in Central and Eastern Europe is mass or voucher privatization. This involves the distribution of vouchers for free, or at a nominal cost, that enables eligible citizens to bid for shares of SOEs that are being privatized (or financial intermediaries such as mutual funds). Although these schemes have resulted in fundamental changes to the ownership of business assets, often they have

not changed the effective control over these assets because of the wide diffusion of share ownership, thus leaving incumbent managers or small concentrated groups of shareholders in control of the assets with little accountability to the new owners (Megginson and Jeffrey Netter, 2001 at 324). Sweeping privatization (often through vouchers) of SOEs in the former Soviet Union in the early 1990s (“shock therapy”) without supporting legal and regulatory institutions was widely perceived to have led to a form of crony capitalism where political insiders (“oligarchs”) captured state assets at undervalue and sometimes looted them, contributing to a precipitous decline in GDP per capita.

Management Buyouts (MBOs)

Employee buyouts involve the transfer of the rights of ownership of industrial or commercial facilities to those who are employed in them. However, in the case of MBOs, the existing management is the buyer (Sondhof and Stahl 1992 at 210). MBOs have been advocated in situations involving small- and medium-sized enterprises (such as those formed by the dissolution of former combines or large enterprises) and as a way to draw upon a substantial pool of “privatization demand” from the people who work in the enterprise (Sondhof and Stahl, *ibid*; Bennett 1997).

However, given the importance of new human capital to the success of a divested SOE documented in many studies, MBOs are often a poor method of privatization (since enterprise management typically does not change as a result of an MBO). There are a number of other reasons why MBOs are often not an optimal method of divestiture:

- MBOs are often highly leveraged, leaving the new enterprise vulnerable to macroeconomic shocks.
- Managers of SOEs, given their information advantages and political influence, may be able to purchase their SOEs for less than the fair market value.
- Management may not be able to secure adequate financing, leaving the developing country government (acting as a *de facto* guarantor or lender) bearing the risk of enterprise failure.

Thus, notwithstanding that MBOs often have the support of the workforce, the MBO is not generally an efficient method of divestiture.

Privatization Through the Sale of State Property

The sale of state property is the method by which a government “trades its ownership claim for an explicit cash payment” (Megginson and Netter, 2001). As discussed by Megginson and Netter, this category can be subdivided into two different approaches. The first is *direct sale* whereby an SOE, or one of its component parts, is sold to an individual, a corporation, or a group of investors (typically the highest bidder). The second option is *share issue privatization* (also known as an SIP), where some or all of the government’s holding in an SOE is sold through a public share offering. The sale of state property serves several concrete objectives, including producing revenue for the state, restructuring organizations, and attracting foreign investors (Brada 1996).

This method of privatization can be controversial, however, as foreign investment represents a paradox to developing governments: while capital-starved developing countries are in need of foreign capital, foreign investment has often been perceived as a neocolonialist threat to newly emergent states attempting to fashion their own societies rather than be absorbed into the political economy of the developed world. Thus, domestic investment may be more attractive to many developing economies, implying that a country’s savings rate (savings as a percentage of GDP) will be a significant factor in the determination of the mode of divestiture adopted. This suggests that a developing country with a higher savings rate is likely not only to rely to a greater degree on domestic investment but also to employ the public offering more frequently as a means of placing some portion of large SOEs in domestic rather than foreign hands. Moreover, developing countries with higher savings rates are more likely to have the financial institutions necessary to facilitate a public offering. To date, direct sales to the domestic private sector in developing countries have been concentrated in smaller-scale activities associated with the commercial,

manufacturing, service, and agricultural sectors. The divestiture of large enterprises, such as utilities, has relied more on foreign direct investment and public share issues. There has also been a tendency for developing countries to privatize in a macroeconomic crisis when there is an urgent need for additional government revenues to finance budgeting deficits. However, the precipitous sale of SOEs may result in fire-sale prices or alternatively a proclivity to turn state-owned monopolies into private monopolies in order to enhance the sale price.

Partial Privatization

Partial privatization of an SOE may be a desirable interim option for governments for whom full privatization is not politically desirable or feasible. Partial privatization may take the form of a partial public offering that, if the stock is widely traded, will introduce stock prices (and changes thereto) as a benchmark of an SOE's performance, as well as creating a private constituency with a direct stake in SOE efficiency. Alternatively, government could sell a minority interest to a single commercial partner or joint venture partner with needed capital and expertise and a role in the SOE's governance, again with a direct stake in enhanced SOE performance. Under both options, full privatization may ultimately be rendered more politically feasible. Government credibility may, however, prove a serious problem with both of these options, in that the stakes of the minority investors are at risk of expropriation or devaluation by changes in policies by the controlling shareholder or government, being motivated by considerations other than efficiency. Hence, minority shareholders may be difficult to attract at all or only at severely discounted acquisition prices, reflecting political risk.

Alternative Reform Strategies

While there are many options in terms of how to privatize SOEs, there are also reform alternatives to privatization that should be identified. The following reforms, to a greater or lesser extent, seek to benefit from the ownership and/or competition effect.

Management Contracts

A management contract is an agreement between a government and a private party (often enforceable through international arbitration or governed by the law of a recognized center of commerce) by which the private party agrees to operate the assets of an existing SOE for a fee, which will often be tied to performance. In other words, management contracts privatize SOE management without divesting the SOE's assets.

While there is evidence providing some support for the use of management contracts as a means to improve SOE efficiency Shirley et al. (1995) at 136–139. Shirley studied the experience of 20 contracts in several countries and industries, measuring their impact on SOE performance in terms of increases in post-contract profitability and productivity, and found that 13 out of the 20 SOEs, or 65 %, were successes. Moreover, profitability or productivity improved in three other organizations. However, they have their drawbacks. First, in order for management contracts to yield significant benefits from the incentive of private management, remuneration under the contract must be tied to organizational performance. In practice, contracts often include fixed fees, and some do not even attempt to relate fees to performance. Secondly, the limited term of a management contract may create perverse management incentives. Because of uncertainties over whether contracts will be renewed – a problem compounded by the lack of credibility of many developing country governments – management may be tempted to maximize short-term gains (e.g., by degrading assets), often at the cost of long-term performance (see Daniels and Trebilcock (1996) at 404). And while a well-drafted management contract may mitigate such perverse incentives by including longer contract terms, technological uncertainties and uncertainties relating to market demand and surrounding government policies may negatively affect the risks associated with a project over the long term, making longer contract terms relatively unattractive to private organizations (see Daniels and Trebilcock, *ibid.*, at 405; although the management contracts of the most successful organizations in the Shirley study generally have longer terms.).

An important question concerns the relative efficiency of the management contract and divestiture. As stated earlier, privatization – assuming appropriate postdivestiture competition and regulatory policies – takes full advantage of both the ownership effect and the competition effect of private ownership. Management contracts, on the other hand, do not take advantage of the ownership (or often the competition) effect and hence may be less effective than divestiture. This is because of a number of factors:

- Private ownership is more likely to result in increased capital investment than management contracts.
- Provided that the privatized SOE does not have monopoly power, divestiture may not result in the same level of transaction costs as management contracts (e.g., the costs of ongoing negotiations and monitoring) (Shirley et al. 1995 at 149).

Thus, where the economic and political prerequisites for privatization are present, privatization is likely to be a better reform alternative than management contracts. However, the key benefit of management contracts is that they can be successfully implemented in circumstances where privatization is not feasible.

Performance Contracts

A key distinction between private and public enterprises is that private enterprises are generally motivated by one clear objective – profit maximization – whereas SOEs typically face multiple, sometimes conflicting, objectives. Poorly defined and shifting corporate objectives severely inhibit the evaluation of management performance SOEs, while a lack of autonomy from government decision-makers and a failure to tie management compensation to organizational performance deter good management performance. These are problems that *may* be resolved by a performance contract.

The three major components of a performance contract system are:

- A performance evaluation system, in which national goals are translated into explicit SOE objectives and quantified in performance criteria
- A performance information system which monitors actual achievements
- An incentive system, in which the welfare of the managers and workers of the SOE is linked to its objectives by a pecuniary or non-pecuniary reward system based on the achievement of particular target values (Hartman and Nawab 1985 at 28–30)

By the mid-1990s, performance contract systems had been implemented in 28 developing countries – mostly in Asia and Africa – but, unfortunately, despite some successes, there have also been many failures. Obstacles to successful implementation of performance contracts include ensuring that SOE managers, who enjoy a significant information advantage over their government monitors, do not use this advantage to negotiate easily attainable performance targets and ensuring that management remuneration is effectively tied to SOE performance. In one study (Dyer Cissé 1994, Background Paper, Policy Research Department, World Bank), only five of the twelve sample companies offered pecuniary bonuses as a performance award. This study also found a tendency for the governments of developing countries to breach their performance contract commitments with impunity, thereby seriously derogating from the significance attached to them by SOE managers.

Thus, the implementation of performance contracts does not appear to be a broad solution to the problem of SOE inefficiencies. Relative to privatization, performance contracts do not increase competition and they face serious obstacles in their attempt to induce good management. However, improvements can be made to the system, such as reducing bargaining inequality by channeling performance contracts through a centralized specialized agency or using published and frequently updated benchmarks from comparable entities in other sectors or countries to lower transaction costs. If these changes and assurances can be made to the performance contract systems

in LDCs, SOEs that cannot (or should not) be privatized for economic or political reasons may still benefit from improved enterprise management.

Increased Competition

While management contracts and performance contracts attempt to improve SOE performance largely through improvements in management incentives and monitoring, these types of SOE reform have little impact on the level of competition faced by an SOE. There are four means by which governments of LDCs may introduce greater levels of competition in their SOE sectors: (i) the unbundling of monopolies and exposing some elements of them to competition, (ii) the reduction of import barriers, (iii) performance-based price regulation using competitive benchmarks from other jurisdictions, and (iv) competition laws.

As in the case of the other alternative reform options, they are still a less effective option than privatization, which takes advantage of the ownership effect as well as the competition effect. However, such a strategy may be a preferable interim measure in situations where, for political or economic reasons, privatization is not a viable option. Like privatization, however, these strategies depend on the existence of strong institutions, including effective competition and regulatory authorities. The challenges in building these institutions in developing countries, however, are not negligible.

Public-Private Partnerships (PPPs)

“PPPs” is a term that exhibits a high degree of ambiguity in common usage, but it is often used to refer to projects in which the private sector designs and builds new or refurbishes and expands existing infrastructure, provides project financing, manages the asset, and operates the service. In other words, the arrangement typically involves “bundling” – or vertical integration – in the form of a consortium that brings together project designers, managers, construction companies, engineers, and financiers (Daniels and Trebilcock 1996 at 390). PPPs recover their costs through user fees, government payments,

or both. Whether the asset reverts to the public sector at the end of the contract depends upon the terms of the contract. Most PPPs implemented to date provide that ownership in the asset is retained by, or will revert to, the public sector. Over the past two decades, PPPs have been extensively employed in both developed and developing countries, mainly with respect to the construction or expansion of infrastructure facilities. In the case of many developing countries, infrastructure deficits are often massive, especially in electricity, transportation, and water supply, and governments lack the fiscal capacity to make the needed investments. PPPs provide an intermediate strategy between SOEs or their privatization (see Trebilcock and Rosenstock 2015).

PPPs are promoted on several grounds. The bundling aspect of PPPs incentivizes the builder-operator to incorporate long-term operating cost considerations in the design and construction phases of the project and arguably also reduces the coordination costs that would otherwise be faced by various departments or agencies of government in undertaking all of the functions performed by a vertically integrated consortium (Delmon 2009 at 10). Similarly, as the PPP transfers to the builder-operator the financial risks associated with design, construction, and operation, there is a heightened incentive to mitigate cost and time overruns. PPPs have also been viewed by some governments as a means of shifting spending off balance sheet or circumventing legislative budgetary limits. Clearly, this is problematic, as masking government liabilities does not reduce them nor is it transparent (Organisation for Economic Co-operation and Development 2008 at 21).

PPP contracts pose special problems of contractual design, monitoring, and enforcement. Typically, they involve large investments in durable, transaction-specific assets of the kind that Nobel Laureate Oliver Williamson describes as raising a bilaterally dependent contractual relationship between the parties (Williamson 1988 at 71). That is, while governments may have some choice among competing consortia *ex ante*, once they “lock into” a long-term relationship with a single consortium, there are significant risks of

mutual opportunism or “holdups” over the course of the subsequent contractual relationship.

Over the past two decades, the majority of PPP projects in developing countries have taken place in countries with relatively high incomes. Just under 60 % of total projects were undertaken in upper middle-income countries and 37 % in lower middle-income countries. Only 4 % of projects took place in low-income developing countries, although this latter number may underestimate the true extent that PPPs are used in low-income countries, especially in sub-Saharan Africa, given the increasing number of “resources for infrastructure” exchanges between Chinese SOEs and sub-Saharan African countries (although because the Chinese firms are often SOEs, these are not strictly public-private partnerships). A recent statistical analysis by the IMF found that after adjusting for population, PPP concentration was more likely in larger markets with a stronger rule of law, political stability, administrative capacity, greater consumer demand, and macroeconomic stability (Hammami et al. 2006).

A number of studies have attempted to evaluate the relative operating efficiencies of PPPs. A number of these studies have found that service quality and reliability have often improved and that labor productivity has greatly increased through substantial reductions in public sector employment (Andrés et al. 2008; Marin 2009 at 29). The effect of PPPs on user fees yields more mixed results (Marin, *ibid.*, at 110; Andrés et al., *ibid.*; Kirkpatrick et al. 2006). As predicted by Williamson, renegotiation rates of PPPs even early in their contractual terms are quite large. Another striking finding is the limited impact of PPPs on improving access to infrastructure relative to a continuation of the *status quo*. Often very poor consumers see relatively few benefits from a PPP and continue to rely on the relatively expensive informal sector for, e.g., access to potable water (Marin 2009 at 64–65; Estache 2005 at 1). In terms of expanding access to essential infrastructure, such as water distribution or electricity, developing country governments need to build stronger incentive mechanisms in PPP contracts to provide such access or alternatively to subsidize low-income consumers directly.

References

- Adam C, Cavendish W, Mistry P (1992) Adjusting privatization: case studies from developing countries. James Currey, London
- Aharoni Y (1986) The evolution and management of state-owned enterprises. Ballinger Publishing Company, Cambridge, MA
- Andrés LA, Luis Gausch J, Haven T, Foster V (2008) The impact of private sector participation in infrastructure: lights, shadows and the road ahead. The World Bank, Washington, DC
- Bennett A (1997) How does privatization work? Routledge, London/New York
- Brada JC (1996) Privatization is transition – or is it? *J Econ Perspect* 10(2):67
- Christiansen H (2011) The size and composition of the SOE sector in OECD countries, vol 5, OECD corporate governance working papers. OECD Publishing, Paris
- Daniels RJ, Trebilcock MJ (1996) The private provision of public infrastructure: an organizational analysis of the next frontier. *Univ Tor L J* 46:393
- Delmon J (2009) Private sector investment in infrastructure. Kluwer Law International, Alphen aan den Rijn
- Dyer Cissé N (1994) The impact of performance contracts on public enterprise performance. The World Bank, Washington, DC
- Estache A (2005) PPI partnerships versus PPI divorces in LDCs, vol 3470, Policy research working paper. The World Bank, Washington, DC
- Estrin S, Hanousek J, Kocenda E, Svejnar J (2009) Effects of privatization and ownership in transition economies, vol 4811, Policy research working paper. The World Bank, Washington, DC
- Gillis M (1980) The role of state enterprises in economic development. *Soc Res* 47(2):248
- Hammami M, Ruhashyankiko J-F, Yehoue EB (2006) Determinants of public-private partnerships in infrastructure, vol 06/99, International monetary fund working paper. International Monetary Fund, Washington, DC
- Haririan M (1989) State-owned enterprises in a mixed economy: micro versus macro economic objectives. Westview Press, London
- Hartman A, Nawab SA (1985) Evaluating public manufacturing enterprises in Pakistan. *Finance Dev* 22(3):27
- Hodgetts J (1971) The Canadian public service 1867–1970. University of Toronto Press, Toronto. Chapter 7
- Iacobucci E, Trebilcock M (2012) The Role of Crown Corporations in the Canadian Economy, School of Public Policy research paper, vol 5, Issue 9. University of Calgary, Calgary
- Kikeri S, Nellis J (2004) An assessment of privatization, vol 19, no 1. World Bank Research Observer, Washington, DC
- Kirkpatrick C, Parker D, Yin-Fang Z (2006) An economic analysis of state and private-sector provision of water services in Africa. *World Bank Econ Rev* 20:143
- Marin P (2009) Public-private partnerships for urban water utilities. The World Bank, Washington, DC

- Meggison WL, Netter JM (2001) From state to market: a survey of empirical studies on privatization. *J Econ Lit* 39(2):321
- Muir R, Saba JP (1995) Improving state enterprise performance: the role of internal and external incentives. The World Bank, Washington, DC
- Organization for Economic Co-operation and Development (2008) Public-private partnerships: in pursuit of risk sharing and value for money. OECD, Paris
- Parker D, Kirkpatrick C (2005) Privatization in developing countries: a review of the evidence and the policy lessons. *J Dev Stud* 41(4):513
- Shapiro D, Globerman S (2012) The international activities and impacts of state-owned enterprises. In: Sauvart KP, Sachs L, Schmit Jongbloed WPF (eds) *Sovereign investment: concerns and policy reactions*. Oxford University Press, Oxford, UK
- Shirley M, Walsh P (2000) Public versus private ownership: the current state of the debates, vol 2420, Policy research working paper. The World Bank, Washington, DC
- Shirley MM et al (1995) Bureaucrats in business: the economics and politics of government ownership. World Bank, Washington, DC
- Smith A, Trebilcock M (2001) State-owned enterprises in less developed countries: privatization and alternative reform strategies. *European J Law Econ* 12:217
- Sondhof H, Stahl M (1992) Management buy-outs as an instrument of privatization in Eastern Europe. *Intereconomics* 27(5):210
- Toninelli PA (ed) (2000) *The rise and fall of state-owned enterprise in the Western world*. Cambridge University Press, Cambridge
- Trebilcock M, Rosenstock M (2015) Infrastructure public-private partnerships in the developing world: lessons from recent experience. *J Dev Stud* 51(4):335–354
- Vernon R, Aharoni Y (eds) (1981) *State-owned enterprise in the Western economies*. Croom Helm, London
- Vickers J, Yarrow G (1991) *Privatization: an economic analysis*. MIT Press, Cambridge, MA
- Vickers J, Yarrow G (1998) *Privatization: an economic analysis*. MIT, Cambridge, MA
- Williamson OE (1988) The logic of economic organization. *J Law Econ Org* 4:65

Static Market Power

Rafael Emmanuel Macatangay
Centre for Energy, Petroleum and Mineral Law
and Policy School of Social Sciences, University
of Dundee, Dundee, Scotland, UK

Abstract

Static market power refers to the ability of economic agents profitably to move prices away

from competitive levels during one time period. Market power, a form of market failure, prevents the achievement of an efficient allocation of resources. The simplest example of market power is monopoly, a market in which there is only one firm. Another example of market power is oligopoly, a market in which there is a small number of firms. The determination of whether or not prices profitably deviate from competitive levels is at the heart of antitrust law and economics. Methods for analysing the unilateral change in incentives arising from merging two firms have become influential as a screening device for identifying potentially anti-competitive mergers in the EU, US, and UK.

Definition

Static market power refers to the ability of economic agents profitably to move prices away from competitive levels during one time period (i.e., a one-shot simultaneous moves game).

In a competitive economy, consumers or producers assumed to be price takers behave as if the demand or supply functions they face are infinitely elastic at prevailing market prices (Mas-Colell et al. 1995). However, if there are only few agents on either side of a market, they are likely to have market power, the ability profitably to move prices away from the marginal cost (i.e., the competitive or socially optimal level). As a form of market failure, market power prevents the achievement of an efficient allocation of resources. Static market power involves agents acting in only one period (i.e., one-shot simultaneous moves). Agents possessing market power are typically sellers. The ability of a firm to exercise market power depends on whether or not consumers can find substitutes and, therefore, varies inversely with the demand elasticity (Church and Ware 2000). The potential for supply substitution depends on the extent to which consumers can switch to alternative suppliers of the same product (i.e., homogeneous goods). If consumers cannot find substitute suppliers capable of providing the same product, then the firm will have market power. The potential for demand

substitution depends on the extent to which consumers can switch to alternative suppliers of products that are acceptable substitutes (i.e., differentiated goods).

If products are sufficiently differentiated, then they are not close substitutes and consumers will not reduce their consumption when price rises above the marginal cost. The possibilities for supply or demand substitution increase the elasticity of demand.

The simplest example of market power is monopoly, a market in which there is only one firm (Mas-Colell et al. 1995). If the monopolist faces a market demand that is continuously decreasing in price, then the monopolist will find it profitable to raise price above the marginal cost. The monopolist recognizes that a reduction in the quantity it sells allows it to increase the price on all its sales and consequently to increase its profits. The monopolist's profit-maximizing output is thus below the socially optimal level. Another example of market power is oligopoly, a market in which there is a small number of firms. If the efficient scale of operation for a firm is large relative to the size of market demand, the equilibrium number of firms could be small. Competition in an oligopoly involves strategic interaction requiring the analytical tool of game theory. In a static model of oligopoly, there is only one period of strategic interaction and firms act simultaneously. A widely used oligopoly model is the Cournot model (named after French mathematician Augustin Cournot writing in 1838) in which quantity is the strategic variable (another oligopoly model, the Bertrand model, in which the strategic variable is price, yields the social optimum even with just two firms, but is not discussed here). In a Cournot duopoly, each firm forecasts the other's output choice and chooses a profit-maximizing output for itself. Price then adjusts to clear the market. In equilibrium each firm is maximizing its profits and confirms its beliefs about the other. Using the terminology of game theory, their strategies constitute a Nash equilibrium (named after American mathematician John Nash writing in 1951) if one's choice is optimal given the other's choice and the other's choice is optimal given one's choice (Varian 2010). Neither

firm will find it profitable to change its output choice once it discovers the output choice actually made by the other.

The Cournot price and profit are lower, respectively, than the monopoly price and profit, and the Cournot output is higher than the monopoly output (Mas-Colell et al. 1995). In a Cournot equilibrium, as each firm estimates the profitability of selling an additional unit of output, it does not consider the reduction in the other's profit due to the decline in the price. If for some reason the firms in aggregate produced the competitive quantity and consequently earned zero profits, either one could benefit from a slight reduction in its output. Moreover, the slope of the marginal cost function is a key determinant of the market outcome (Vives 1999). Flat marginal costs are conducive to socially optimal behavior, but steep marginal costs, indicative of inflexible technologies, are conducive to Cournot behavior. A Cournot model can be configured to nest the two extreme outcomes of social optimality on one end and monopoly on the other (Mas-Colell et al. 1995). If the number of firms is very large, price, aggregate output, and profits approach the socially optimal levels, but if the number of firms is very small (or set to one), they approach the monopoly levels. Each firm's influence on price is negligible if there are many firms.

One criticism of the Cournot model is that, in some situations, firms seem to choose prices rather than quantities (Mas-Colell et al. 1995). One way of addressing this is to interpret quantity choices in the Cournot model as long-run choices of capacity, and thus to let price determination through the inverse demand function serve as a proxy for short-run price competition, given the capacity choices. Another criticism of the Cournot model is that, in some situations, firms seem to choose both price and quantity (Vives 1999). One way of addressing this is to deploy an analytical approach in which firms compete in supply functions. Under a supply function equilibrium ("SFE"), the strategy for a firm is a price-quantity pair that is optimal given the price-quantity pairs of other firms. The market outcome is a Nash equilibrium in supply functions. One criticism of SFE is that there are infinitely many supply

functions that satisfy the optimal market price and quantity. One way of addressing this is to introduce demand uncertainty that limits the multiplicity of supply function equilibria.

The choice of oligopoly model is ultimately a function of the features of the market (Vives 1999). For example, in the world market for crude oil, the Organization of Petroleum Exporting Countries (“OPEC”) is a major player and its aims are effectively those of a cartel (Carvajal et al. 2013). In principle firms that are not colluding ought to be playing the static Nash equilibrium at each time period. A test of Cournot rationalizability determines if the market outcome could be explained as a static Nash equilibrium. If it could not, then antitrust authorities could consider alternative explanations, such as complex or possibly collusive (i.e., dynamic) forms of strategic behavior. An analysis of data on OPEC and non-OPEC countries suggests that the behavior of oil-producing countries is not Cournot rationalizable (i.e., the Cournot hypothesis is rejected).

Whether the price or quantity competition model is better for oligopolies remains an open question (Vives 2011). SFE, which appears to be a viable alternative, has been used to model bidding for government procurement contracts, management consulting, or airline pricing reservation systems. The analysis of competition in wholesale electricity markets in different countries has been a fertile area for testing alternative oligopoly models. One of the major lessons from the international experience of electricity industry liberalization over the last two decades is that the wholesale market for electricity is prone to the exercise of unilateral market power (Wolak 2014). The elasticity of electricity demand tends to be inelastic, close to zero in the short run, and well below one within a year (Stoft 2002). Market power in wholesale electricity markets is exercised strategically by reducing output that could be produced profitably at the market price (i.e., quantity withholding) or raising the price of output above the marginal cost (i.e., financial withholding). Either would steepen the supply curve and the outcomes deviate from the socially optimal levels for price, quantity, and profit. For the analysis of competition in wholesale

electricity markets, Cournot models have been used in California, the Pennsylvania-New Jersey-Maryland area, and New England and SFE models, in England and Wales and Texas (Willems et al. 2009). An analysis of competition in the German electricity market using both Cournot and SFE models shows that each model explains similar proportions of observed price variations.

The determination of whether or not prices profitably deviate from competitive levels is at the heart of antitrust law and economics. The upward pricing pressure (“UPP”) methodology and the gross upward pricing pressure index (“GUPPI”) have become influential as a screening device for identifying potentially anticompetitive mergers in the EU, USA, and UK (Davis and Fletcher 2013). The underlying rationale is to focus on the key likely anticompetitive effect of a merger, the unilateral change in incentives arising from merging two firms. One interesting implication is that high profits are not necessarily indicative of market power. High profits can occur in a situation of effective competition (e.g., through superior management), or a low profit margin could occur in monopoly (Bork and Sidak 2013). Using a firm’s profit data to infer market power might lead a court or competition authority to the wrong conclusion.

The current Merger Regulation of the European Commission (the “Commission”) ensures that competition and consumers are not harmed by excessive concentrations of market power (European Commission 2014). One of the key assessments is a test for a significant impediment of effective competition (“SIEC”). Although SIECs most prominently arise from the creation or strengthening of a dominant position, the test also addresses the issue of mergers in which the merged entity is not dominant but potentially has anticompetitive “non-coordinated effects.” In an oligopoly in which no firm is individually dominant and collusion is unlikely, merging firms may have an incentive to increase their prices unilaterally even if they do not become dominant, and the remaining firms, benefitting from the reduction in competitive pressure due to the merger, might also increase their prices. The test is designed to

address mergers which allow firms unilaterally to raise prices but do not create or reinforce a single or collective dominant position. If competitive constraints are sufficiently strong, a merger resulting in a high combined market share of the merging firms may not have a negative impact on competition or consumers.

The merger control regime in the UK generally but not exclusively involves the Competition and Markets Authority (“CMA”) for UK companies covered by the provisions of the Enterprise Act 2002, as amended; the commission for mergers under the Merger Regulation; and the relevant sectoral regulators for natural gas and electricity markets, water services, rail services, civil aviation, health, and (from 1 April 2015) financial services (Competition and Markets Authority 2014). The CMA has a duty to refer, for an in-depth “phase 2” investigation, any relevant merger situation that risks resulting in a substantial lessening of competition (“SLC”) in a UK market. UK competition authorities have been using techniques very closely related to UPP or GUPPI to diagnose unilateral price effects (Shapiro 2010). The restoration of the dynamic process of firm rivalry through remedies that reestablish the market structure expected in the absence of the merger (i.e., structural remedies, such as divestitures) is projected to address the root cause of adverse effects (Competition Commission 2008). Structural remedies are generally preferable to behavioral remedies (e.g., price caps, supply commitments, or restrictions on long-term contracts) which not only are deemed unlikely to deal with an SLC and its adverse effects as comprehensively as structural remedies but also may result in distortions to the socially optimal outcome.

The 2010 Horizontal Merger Guidelines of the US Department of Justice and the Federal Trade Commission reflect the ascendancy of unilateral effects as the theory most often pursued by US competition authorities (Shapiro 2010). A key element of the analysis is the disincentive created by the merger to sell additional units of product 1 if they cannibalize unit sales of product 2 (the disincentive is an opportunity cost borne by the merged firm for selling product 1). A merger

generates net upward pricing pressure for product 1 if the opportunity cost exceeds the efficiencies for it. The treatment of unilateral price effects rests on economic principles related to how firms account for opportunity costs (i.e., cannibalization), when pricing and promoting product lines containing substitute products, and how higher costs tend to lead to higher prices. Moreover, there is additional flexibility in the implementation of the hypothetical monopolist test (“HMT”). Under the HMT, a market consists of a product (or group of products) and a geographic area in which it is sold, such that a hypothetical, profit-maximizing firm (that is both not subject to price regulation and the only present and future seller of the product in the area) would impose a small but significant and non-transitory increase in price (“SSNIP”) above prevailing or likely future levels. The purpose of defining the market and measuring market shares is to assist in the analysis of competitive effects. The use of market shares and concentration is in the context of the smallest relevant market satisfying the HMT.

References

- Bork RH, Sidak JG (2013) The misuse of profit margins to infer market power. *J Comp Law Econ* 9(3):511–530
- Carvajal A, Deb R, Fenske J, Quah JK-H (2013) Revealed preference tests of the Cournot model. *Econometrica* 81(6):2351–2379
- Church J, Ware R (2000) *Industrial organization: a strategic approach*. Irwin McGraw-Hill, Boston
- Competition and Markets Authority (2014) *Mergers: guidance on the CMA’s jurisdiction and procedure*. Competition and Markets Authority, London
- Competition Commission (2008) *Merger remedies: Competition Commission guidelines*. Competition Commission (now the Competition and Markets Authority), London
- Davis P, Fletcher A (2013) Contributions to competition economics. *Econ J* 123(572):F493–F504
- European Commission (2014) Commission staff working document accompanying the document “Towards more effective EU merger control.” European Commission, Brussels
- Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic theory*. Oxford University Press, New York/Oxford
- Shapiro C (2010) The 2010 horizontal merger guidelines: from hedgehog to fox in forty years. *Antitrust Law J* 77:701–759

- Stoft S (2002) Power system economics: designing markets for electricity. Wiley-IEEE Press, New York
- Varian H (2010) Intermediate microeconomics a modern approach. W. W. Norton, New York/London
- Vives X (1999) Oligopoly pricing old ideas and new tools. MIT Press, Cambridge/Massachusetts/London
- Vives X (2011) Strategic supply function competition with private information. *Econometrica* 79(6):1919–1966
- Willems B, Rumiantsev I, Weigt H (2009) Cournot versus supply functions: what does the data tell us? *Energy Econ* 31(1):38–47
- Wolak FA (2014) Regulating competition in wholesale electricity supply. Chapter 4. In: Rose NL (ed) *Economic regulation and its reform: what have we learned?* University of Chicago Press, Chicago

Strict Liability Versus Negligence

Vaia Karapanou
Athens, Greece

Abstract

Strict liability and negligence are the basic rules that courts apply to affirm tort liability and award damages. Law and Economics is concerned with the efficiency of the rules and therefore compares strict liability and negligence based on the incentives they provide accident parties to minimize total accident costs. To analyze the implications of the two liability rules for economic efficiency, a distinction is made between unilateral accidents, in which only the injurer can affect the probability of accident occurrence and the magnitude of the accident losses, and bilateral accidents, in which both accident parties can affect the probability of accident occurrence and the magnitude of the losses. Provided that damages are perfectly compensatory and that the due care level is set to be equal to the optimal level of care, both rules induce optimal precaution in unilateral accident settings; however, only the rule of strict liability induces the injurer to engage in the optimal activity level. In bilateral accident settings, all rules besides the rule of simple strict liability induce both accident parties to take optimal care; however,

none of the rules can simultaneously create both accident parties incentives to engage in the optimal activity level. The litigation costs per trial are higher under a negligence rule and its variations because courts have to determine if the injurer and/or the victim complied with the set level of due care. In accident contexts where a preexisting market relationship exists between the injurer and the victim, such as in the case of a defective product, a critical factor is whether consumers (victims) are aware of the risk level of the product. If they are, all liability rules create the injurer optimal precaution incentives; if they are not, only the rule of strict liability with a defense of contributory negligence creates optimal care incentives. Today, the rule of negligence is the most commonly used rule of tort liability.

Definition

The rules of strict liability and negligence are the basic rules that courts apply to affirm tort liability and award damages. Under a rule of strict liability, a person is liable for all the accident losses she causes. Under a rule of negligence, a person is liable for the accident losses she causes only if she was negligent. Being negligent implies that the person took less care than the minimum acceptable level prescribed by the law and/or by the court. Law and Economics is concerned with the efficiency of the rules and therefore compares strict liability and negligence based on the incentives they provide to reduce accident losses.

Introduction

Tort law stipulates under which circumstances a person suffering civil harm such as loss or impairment of property, reputation, health and life is entitled to damage compensation. In his seminal book *The Costs of Accidents*, Calabresi (1970) proposes that the principal goal of tort law besides justice is to improve social welfare by minimizing the total costs of accidents. These include the costs of taking precautions to avoid accident

occurrence, the accident losses that are still expected to occur, the costs involved with spreading the losses through the insurance mechanism, and the administrative costs of the legal system relating to the adjudication of tort damages. The prospect of being held liable and having to bear accident losses provides potential wrongdoers incentives to take care and to better control their activity. Economic analysis of tort law compares liability rules to determine under which circumstances strict liability or negligence is more suitable to improve social welfare by creating incentives to take precautions that reduce accident losses.

Brown (1973) and later on Shavell (1980) developed models formalizing the ideas of Calabresi (1970) with regard to the implications of various liability rules for economic efficiency. These basic models are not concerned with the costs of spreading the losses and the administrative costs of the legal system but rather focus on reducing the costs of precaution and the expected accident losses. The models assume that accidents may occur between two parties, an injurer and a victim, causing losses L . Both parties make decisions regarding their activity, depending on the liability rule that is in place, to avoid having to bear accident losses. The decisions involve choices regarding how much care x to take and how often to engage in the activity. The probability of accident occurrence, denoted by p , is affected by these decisions as well as the magnitude of the expected accident losses denoted by pL . Figure 1 provides a graphical representation of how the costs of accidents are influenced by the choices of an accident party regarding her level of care. The cost $B(x)$ of taking a certain amount of care is depicted as well as the expected accident losses given the level of care $p(x)L$ (Landes and Posner 1987; Cooter and Ulen 2011). The upward slope of line $B(x)$ indicates that the marginal cost of precaution increases as more care is taken. The downward slope of line $p(x)L$ indicates that the marginal expected accident losses decrease as more care is taken. The goal is to choose a level of care that minimizes the sum of the costs of care and the expected accident losses, which together constitute the expected total costs of accidents

(TC). The following sections make a distinction between unilateral and bilateral accidents (Shavell 1980, 1987) to discuss the implications of strict liability and negligence for economic efficiency. The analysis is mainly based on Cooter and Ulen (2011), Landes and Posner (1987), Polinsky (2011), Posner (2010), Schäfer and Ott (2005), and Shavell (1980, 1987, 2004).

Unilateral Accidents

In a unilateral accident, it is assumed that only the injurer can influence the probability of accident occurrence and the magnitude of damages, while on the other hand, only the victim incurs accident losses. In this setting a liability rule is efficient if it provides the injurer incentives to exercise optimal precaution that minimizes total accident costs.

Under a rule of *strict liability*, the injurer bears the cost of precaution and the accident losses that are still expected to occur, namely, she has to pay for the total accident costs. The injurer will take this prospect under consideration and compare the costs of taking precautions and the benefits arising thereof in the form of a reduction in expected accident losses. Given that her private costs from the accident coincide with total accident costs, the injurer takes a level of care that is also the socially optimal level of care x^* , as it minimizes total accident costs. Besides choosing the level of precaution, the injurer will also decide on how often to engage in the activity. Given that an increase in the level of activity will result in an increase of the accident probability and/or of the magnitude of accident losses, the injurer will compare her benefits from engaging in the activity more often and the costs arising thereof in terms of an increase in expected accident losses. She will therefore also be induced to choose a socially optimal level of activity.

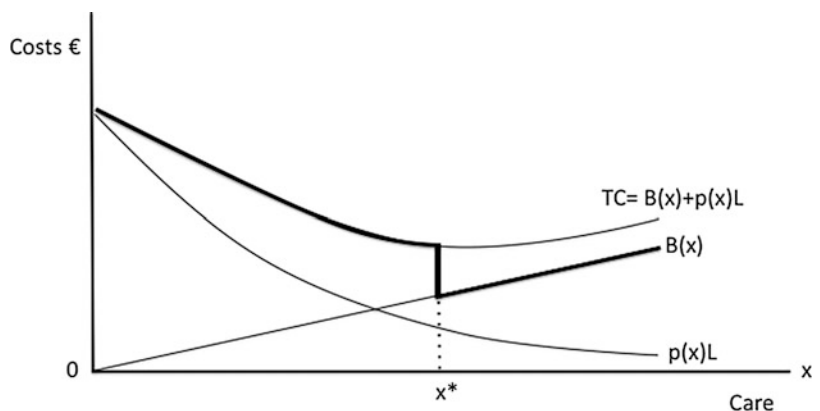
Under a *negligence* rule, the injurer will have to pay for the total accident costs only if she did not take the level of due care set by the law and/or by the courts. Due care, frequently also referred to as reasonable care, is usually defined as the ordinary precaution that a reasonable person would have exercised under the circumstances,

compared to the precaution that the injurer actually exercised. To decide whether the injurer is negligent, the court will have to consider if the injurer took the due level of care. The most famous and commonly used test of negligence was developed by the American judge Learned Hand (1947), who suggested that an injurer is negligent if the burden B of her precautions is less than the gravity of the loss L multiplied by the probability p that an accident occurs, that is, if $B < pL$. Assuming that judge Hand had marginal values in mind, his rule implies that an injurer can escape liability if she takes additional precautions to the extent that the marginal costs of her precautions equal the marginal benefit arising from the resulting decrease in expected accident losses. The level of care that satisfies this equality is the socially optimal level of care. Expressed in algebraic terms, the rule obtained is $B'(x^*) = -p'(x^*)L$. If the due care level is set to be equal to the socially optimal level of care x^* , then total accident costs will be minimized as the injurer will prefer to comply with the level of due care to avoid having to pay for the total accident losses. The negligence rule hence implies that the injurer will be responsible for the accident losses circumscribed by the bold line in Fig. 1. Given that the injurer will only have to take due care to escape liability, she will only consider her private benefits from engaging in the activity and will not take into account the effect of the activity on expected accident losses. She may therefore choose to engage in the activity too often.

Liability Rules Compared

A comparison of the rules of strict liability and negligence shows that both rules can provide the injurer incentives to exercise socially optimal precaution that minimizes total accident costs. However, only the rule of strict liability can create incentives for choosing a socially optimal level of activity (Shavell 1980). Furthermore, the strict liability rule is efficient only if damages awarded by courts are perfectly compensatory in order to induce the potential injurer to take optimal care and engage in optimal activity levels. If damages are systematically underestimated or overestimated, then the injurer will be respectively induced to take a suboptimal or an excessive level of precaution and activity. Under a negligence rule on the other hand, the damages need not reflect the total loss incurred but rather be high enough to make taking optimal care more attractive than taking lower care. As long as the costs of taking optimal care are lower than the total accident costs with less than optimal care, the injurer will choose to take optimal care. However, inefficient precaution may also occur under a negligence rule if the level of due care is not set to be equal to the socially optimal level of care. As the injurer will always comply with the due care level to avoid being held liable, a due care level that is set below the socially optimal level of care will motivate her to take suboptimal precaution, while a due care level that is set above the socially optimal level of care will induce her to take excessive precaution. In the latter case, if the costs of

Strict Liability Versus Negligence, Fig. 1 The influence of the injurer's care choices on the costs of accidents in a unilateral accident setting



precaution exceed total accident costs, the injurer may choose to take a lower level of care that is closer to the socially optimal level. The effect of random errors will be greater under a negligence rule than under a strict liability rule. Random errors in the assessment of damages under a strict liability rule will not change the incentives for precaution as long as damages are correct on average. On the other hand, random errors in setting the due care level under a negligence rule will lead to an increase in precaution as they will induce the injurer to take more care than the due care level to make sure that she will avoid liability (Calfee and Craswell 1984, Craswell and Calfee 1986).

Another difference between the rules of strict liability and negligence regards the administrative costs of trial. Under a rule of strict liability, the court will only need to establish causation between the accident losses and the activity of the injurer, whereas under a rule of negligence, the court will also have to determine if the injurer is negligent by examining whether her level of care fell short of the due care level. It is therefore safe to conclude that the cost per trial will be higher under the rule of negligence. Given that it is easier to prove harm under strict liability, the number of damage claims is likely to be larger than under negligence; however, as it will be easier to predict the outcome of the trial, many of the claims under strict liability will be settled out of court.

Bilateral Accidents

In bilateral accidents it is assumed that both the injurer and the victim can influence the probability of accident occurrence and the magnitude of damages. In this setting a liability rule is efficient if it provides both the injurer and the victim incentives to exercise a level of precaution that minimizes the expected total costs of the accident.

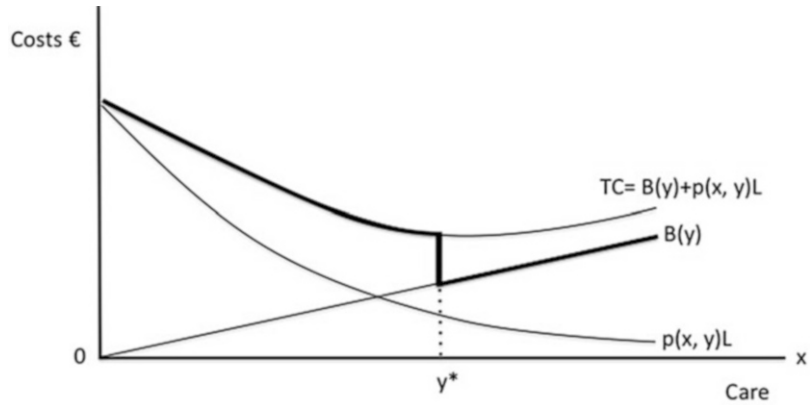
Evidently, a simple *strict liability* rule cannot motivate both accident parties to exercise optimal precaution. As explained in the previous section, under the rule of strict liability, the injurer is liable for the total accident costs and therefore will take

an optimal level of care and of activity in order to minimize them. However, given that the victim's losses will be fully compensated, the victim will not be induced to take any precaution or control her activity level. Therefore, the rule of simple strict liability will not provide efficient incentives for precaution in the case of bilateral accidents.

A variation of the rule of strict liability that deals with the problem of inadequate incentives for the victim is the rule of *strict liability with a defense of contributory negligence*. Under this rule the injurer is liable for the total accident costs unless the victim was negligent, in which case the victim has to pay for her losses. Being negligent implies that the victim took less care than the due care level set by the law and/or by courts. Under this rule the victim will choose to take due care to avoid having to pay for her accident losses. If courts set the due care level for the victim to be equal to the socially optimal level of care y^* , then total accident costs will be minimized as victims will be induced to exercise optimal precaution to escape liability. The bold line in Fig. 2 shows the costs that the victim will have to bear under a rule of strict liability with a defense of contributory negligence depending on her care decisions. Note that the figure depicts expected accident costs and total accident costs as a function of both the victim's and the injurer's care decisions as both actors can influence the probability of accident occurrence and the magnitude of accident losses in a bilateral accident setting. Given that the victim will choose to comply with the due care level, the injurer will be liable for the total accident losses. Therefore she will choose to take optimal precautions and engage in a correct level of activity that minimizes the expected total costs of accidents. The rule hence induces both accident parties to take a socially optimal level of precaution. However, only the injurer will choose an optimal level of activity. The victim will just need to take due care to avoid liability; therefore, she may engage in the activity too often.

Strict liability with a defense of comparative negligence is another variation of the simple strict liability rule that can induce both accident actors to take optimal precautions (Haddock and Curran

Strict Liability Versus Negligence, Fig. 2 The influence of the victim's care choices on the costs of accidents in a bilateral accident setting



1985). This rule is similar to the previous one with the difference that if the victim is negligent, she will have to bear only the fraction of the accident losses that is in proportion to the amount of care that the victim failed to exercise.

Under a rule of *negligence*, the injurer will take due care to avoid being held liable. If courts set the level of due care equal to the socially optimal level of care, the injurer will exercise socially optimal precaution. Given that the injurer takes due care, the victim will be the residual cost bearer, namely, she will have to pay for the total accident losses as designated in Fig. 2 by the curve TC . Therefore, she will take into account the benefit arising from taking care and engaging in a correct activity level in the form of a reduction in expected accident losses and will choose to exercise a socially optimal level of care y^* and of activity that minimizes total accident costs. Hence, under this rule both actors will exercise optimal precaution. However, only the victim will consider the effect of her activity level on expected accident losses and engage in the optimal level of activity. The injurer will just need to take due care to escape liability and therefore she will only consider her private benefits arising from the activity; she is consequently likely to engage in the activity too often.

The rule of *negligence with a defense of contributory negligence* is a variation of the simple negligence rule. Under this rule a negligent injurer will be able to escape liability if the victim is also negligent. In other words the injurer will have to bear the total accident losses if she fails to take the due care level set by the law and/or by courts

while the victim does. On the other hand, the victim will have to pay for the total accident losses if she does not comply with the due care level y^* set by the law and/or by courts irrespective of whether the injurer exercised due care. It should be noted that contrary to the rule of simple negligence where the court does not have to determine a due care level for the victim, under this rule the court has to set a level of due care for both the injurer and the victim. If the due care level is set by courts to be equal to the socially optimal level of care, then both accident parties will be induced to exercise socially optimal precaution. The injurer will comply to the due care level to escape liability, while the victim will also be motivated to exercise due care to avoid having to bear total accident losses. Under this rule, if both parties are negligent, only the victim will bear the total accident losses compared to the rule of simple negligence under which if both parties are negligent, only the injurer has to pay for the total accident losses.

A variation of the negligence rule that is used to divide the accident losses between the injurer and the victim when both of them have acted negligently is the rule of *comparative negligence*. Under this rule both accident parties escape liability if they comply with the due care level set by the law and/or by the courts. However, if both of them fall short of exercising the due level of care, then each of them will be liable in proportion to her negligence. In this case the court will assign liability by comparing the actual level of precaution with the due care level that each party should have

exercised to avoid being held liable. Under this rule both parties are motivated to comply with the due care level to avoid having to pay for the total accident losses. Hence, if the due care level is set to be equal to the socially optimal level of care, both parties will be induced to take optimal care.

Liability Rules Compared

A comparison of the rules of strict liability and negligence and their variations shows that apart from the rule of simple strict liability, all other liability rules are efficient as they can provide both injurers and victims incentives to exercise optimal precaution that minimizes total accident costs. However, it is also clear that, irrespective of the rule chosen, in bilateral accidents it is not possible to provide both actors with the proper activity incentives. Only the residual risk bearer will compare her private benefit from the activity with the total accident costs; the other actor will compare her private benefit from the activity with the cost of taking due care. The rules of strict liability with a defense can therefore control the activity levels of injurers but fail to control the activity levels of victims (Shavell 1980; Landes and Posner 1987), while negligence rules can motivate victims to engage in the optimal level of activity but do not provide injurers proper activity incentives. It follows that in bilateral accident settings, the choice between the rules of strict liability with a defense and the rules of negligence depends on whose activity level one wants to control, the injurer's or the victim's.

Under the rules of strict liability with a defense, fully compensatory damages will not be necessary to induce the injurer to take optimal precaution. Given that victims will take due care to escape liability, the injurers will have to bear the expected accident losses and therefore will be induced to take optimal care to minimize total accident costs. Hence, damages need not fully reflect all losses but rather be high enough to make taking optimal care a better choice than taking lower or no care. The same will be true under the rules of negligence. Moreover, the problem of inefficient precaution that arises under negligence rules if the due care level is not set to be equal to the optimal level of care will also be relevant for the rules of

strict liability with a defense. Hence, if courts set the due care level for the victim below the socially optimal level of care, the victim will be induced to take inadequate precaution, whereas if the due care level is set above the optimal care level, the victim will be motivated to take excessive precaution. In order to determine what should be the due care level for the victim under a rule of strict liability with a defense, courts will have to consider the injurer's due care level as well, given that in bilateral accident settings the care decisions of both accident parties affect the magnitude of expected accident losses. This implies that the administrative costs of trial will be of a similar magnitude for both the rules of strict liability with a defense and the rules of negligence. Under both types of rules, courts will have to determine if the injurer and/or the victim is negligent by examining whether her level of care fell short of the due care level. Under a simple negligence rule, the court will have to determine if the injurer complied with the due care level, while under a negligence rule with a defense, the court will additionally have to determine if the victim complied with the due care level. The same will be true for the rules of strict liability with a defense, under which the court will have to set the due care levels for both the victim and the injurer but will only need to determine what was the victim's actual level of care. Therefore, in a bilateral accident setting, a simple negligence rule will impose the lowest cost per trial, whereas under a rule of strict liability with a defense, the cost per trial will be lower than under a rule of negligence with a defense.

Accidents Occurring Under a Preexisting Market Relationship

The same analysis of liability rules, albeit with some variations, holds also for accidents where parties have a preexisting market relationship such as accidents occurring as a result of a defective product. A critical factor is whether consumers (victims) are aware of the risk level of the product. If they have perfect knowledge of the risk, all rules of liability will lead to optimal care levels because the knowledge of the risk induces consumers to take optimal care, and the

force of the market drives firms (injurers) to take optimal care so as to be able to offer their product at the lowest full price, consisting of the market price and the expected accident losses. If consumers perceive the product risk correctly, then under a rule of *negligence*, consumers will regard as full price the price offered by the producer plus the expected accident losses they expect to bear themselves. Therefore, consumers will choose to exercise optimal precaution. Under a rule of *strict liability*, the price offered by the producer will already be the full price, as it will incorporate the accident losses the producer is expected to bear in the occurrence of an accident. Under both liability rules, the producer's costs of care will be included in the offered price. If however consumers are not aware of the product risk, then only *strict liability with a defense of contributory negligence* creates optimal care incentives, as in this case firms will take optimal care to minimize the expected total accident costs and consumers will take due care irrespective of their perception of the risk to avoid being held liable and having to bear the accident losses. It is important to note that the analysis of liability rules in product-related accidents changes somewhat depending on the type of the product. If the product is a service or a nondurable, the analysis above remains valid. However, if the product is a durable, then besides inducing the consumer to take due care, it will also be important to induce the consumer to engage in the optimal activity level (Shavell 1987, 2004). This may occur only under a rule of negligence if the consumer has a perfect knowledge of the risk because then the consumer will be the residual risk bearer, which will induce her to engage in optimal activity. If the consumer is not aware of the product risk, no liability rule will induce her to engage in the optimal activity level.

Strict Liability Versus Negligence and Related Topics

The rules of strict liability and negligence can be further differentiated based on their ability to create optimal incentives under special accident conditions. The rule of negligence, for instance, is,

under certain conditions, considered to be more appropriate than strict liability for cases where the injurer is judgment proof, meaning that her wealth does not suffice to cover accident losses (Shavell 1986; Dari-Mattiacci and De Geest 2005). This is so because an injurer who is unable to pay for total accident costs will have no incentive to take optimal care under a rule of strict liability but may be induced to take optimal care under a negligence rule provided she has enough wealth to afford optimal precaution because in that case she will escape liability. Both rules of strict liability and negligence will create injurers optimal incentives in cases where multiple injurers cause an accident jointly (Landes and Posner 1980; Kornhauser and Revesz 1989). However, only the rule of negligence will induce injurers to take optimal care in cases where multiple injurers cause an accident independently. In that case, under a rule of negligence, each injurer will assume that other injurers will take optimal care and therefore will be induced to take optimal care to avoid liability. On the other hand, taking optimal care under a rule of strict liability will only reduce each injurer's expected accident losses by a fraction, and hence injurers will not be motivated to take optimal care. An additional advantage of the negligence rule is that it generates information regarding safety technology (Schäfer and Müller-Langer 2009). Under a rule of negligence, firms will disclose information on their private cost levels to courts in order to enable them to determine the due level of care. The dissemination of the resulting court decisions will provide other firms valuable information on safety technology free of charge.

The preceding analysis clarifies that from a Law and Economics point of view, the choice between the rule of strict liability, the rule of negligence, and their variations in a given accident setting depends primarily on their ability to create accident parties incentives leading to the minimization of total accident costs. Litigation costs involved with the application of each rule also play a role as to which rule is most desirable. Provided that courts are able to assess damages correctly and that law and/or courts can set the due care level equal to the optimal level of care, it is

concluded that strict liability is a better liability rule in a unilateral accident setting because it induces the injurer to take optimal levels of care and activity. In bilateral accident settings, all liability rules apart from simple strict liability create optimal care incentives to injurers and victims. However, no rule can induce both actors to engage in the optimal activity level.

Today, the rule of negligence is the most commonly used rule of tort liability. The rule of strict liability (with a defense of contributory negligence) is also employed, most notably in cases where it is important to control the care and activity level of the injurer as is the case with product liability and environmental liability.

Cross-References

- ▶ [Economic Analysis of Law](#)
- ▶ [Efficiency](#)
- ▶ [Tort Damages](#)

References

- Brown JP (1973) Toward an economic theory of liability. *J Legal Stud* 2(2):323–349
- Calabresi G (1970) *The costs of accidents. A legal and economic analysis*, Yale University Press, New Haven
- Calfee JE, Craswell R (1984) Some effects of uncertainty on compliance with legal standards. *Va Law Rev* 70(5):965–1003
- Cooter R, Ulen T (2011) *Law and economics*, 6th edn. Pearson Addison Wesley, USA
- Craswell R, Calfee JE (1986) Deterrence and uncertain legal standards. *J Law Econ Org* 2(2):279–303
- Dari-Mattiacci G, De Geest G (2005) Judgment proofness under four different precaution technologies. *JITE* 161(1):38–56
- Haddock D, Curran C (1985) An economic theory of comparative negligence. *J Legal Stud* 14(1):49–72
- Hand L (1947) *United States v. Carroll Towing Co.*, 159 F. 2d 169 (2nd Cir. 1947)
- Kornhauser LA, Revesz RL (1989) Sharing damages among multiple tortfeasors. *Yale Law J* 98(5):831–884
- Landes WM, Posner RA (1980) Joint and multiple tortfeasors: an economic analysis. *J Legal Stud* 9(3): 517–556
- Landes WM, Posner RA (1987) *The economic structure of tort law*. Harvard University Press, Cambridge/London
- Polinsky MA (2011) *An introduction to law and economics*, 4th edn. Wolters Kluwer, New York
- Posner RA (2010) *Economic analysis of law*, 8th edn. Aspen Publishers, New York
- Schäfer HB, Müller-Langer F (2009) Strict liability versus negligence. In: Faure M (ed) *Tort law and economics*. Edward Elgar, Cheltenham
- Schäfer HB, Ott C (2005) *Lehrbuch der ökonomischen Analyse des Zivilrechts (Textbook economic analysis of civil law)*, 4th edn. Springer, Berlin
- Shavell S (1980) Strict liability versus negligence. *J Legal Stud* 9(1):1–25
- Shavell S (1986) The judgment proof problem. *Int Rev Law Econ* 6(1):45–58
- Shavell S (1987) *Economic analysis of accident law*. Harvard University Press, Cambridge
- Shavell S (2004) *Foundations of economic analysis of law*. Harvard University Press, Cambridge

Further Reading

- Arlen JH (1990) Re-examining liability rules when injurers as well as victims suffer losses. *Int Rev Law Econ* 10(3):233–239
- Arlen JH (1992) Liability for physical injury when injurers as well as victims suffer losses. *J Law Econ Org* 8(2):411–426
- Dari-Mattiacci G, Mangan B (2008) Disappearing defendants versus judgment-proof injurers. *Economica* 75(300):749–765
- Grady M (1983) A new positive economic theory of negligence. *Yale Law J* 92(5):799–829
- Kahan M (1989) Causation and incentives to take care under the negligence rule. *J Legal Stud* 18(2):427–447
- Posner RA (1972) A theory of negligence. *J Legal Stud* 1(1):29–96
- Wittman D (1981) Optimal pricing of sequential inputs: last clear chance, mitigation of damages, and related doctrines in the law. *J Legal Stud* 10(1):65–91

Subjective Punishment

Josef Montag^{1,2} and Tomáš Sobek³

¹International School of Economics, Kazakh-British Technical University, Almaty, Kazakhstan

²Faculty of Law, Charles University, Prague, Czech Republic

³Faculty of Law, Masaryk University, Brno, Czech Republic

Abstract

A punishment implies some discomfort for its recipient – else it would not punish. Indeed, criminal justice can be viewed as an intricate system that calibrates the severity of

punishment – and therefore the amount of the associated discomfort – to individual offenses and offenders. This is done primarily by adjusting the nominal size of punishment, given by the length of a prison sentence, the number of hours of community service, or the amount of a fine. However, the discomfort from a punishment is codetermined by a host of other factors, such as differences across prison facilities, judicial delays, the punishee's psychological setup, her wealth, luck, family relations, and so on. It is the *interaction* of the nominal punishment with these subjective factors that determine the total amount of discomfort that a punishment creates in a punishee. "Subjective punishment" (or individualized sentencing) is an umbrella term for a variety of theories that suggest these factors should be accounted for when courts decide on a punishment. Their common denominator is that – in order to maintain the equality of the (total) punitive effect for the same crime – they often imply that different offenders should receive different nominal punishment. This essay aims to provide a broad overview of these theories and their potential implications for criminal justice.

Introduction

The primary purpose of criminal sanctions is to give individuals incentives to avoid actions that are socially undesirable, or to reduce the extent of such activities, in other words to deter. In this sense, sanctions are the opposite of rewards, which give individuals incentives to do something socially desirable. For sanctions to be effective in achieving deterrence, their implementation must be associated with a discomfort, or utility reduction, experienced by the recipient. A fine, for instance, effectively reduces an individual's disposable income and thus the amount of goods and services she can afford to buy, lowering her utility. A prison term prevents an individual from undertaking actions, legitimate or otherwise, she would choose to undertake if at large, again resulting in lower utility.

Importantly, this utility reduction from sanctions is fundamentally a subjective phenomenon. As a result, the discomfort from a certain punishment is likely to vary across individuals. In this essay, we try to take a bird's eye view and look at the major factors that generate variation in punishment, other than the sentence pronounced by a court. In particular, two main sources of this variation are relevant to our discussion: (i) differences in the punishees' personae, that is their ability to cope with, or to "absorb," a particular punishment; and (ii) idiosyncratic circumstances that are external to the punishment itself but interact with it. An example of such an idiosyncratic factor is a time lag between the punishment and a crime, which results in a variation in the de facto punishment because delayed punishment is less severe due to temporal discounting. An example of individual, or psychological, factor can be differences in time preferences across individuals, which result in differences in the discounted value of a delayed punishment.

A question naturally arises, whether, to what extent, and how should be such variation in the disutility associated with a punishment taken into account when determining punishment in criminal cases. These issues were recently discussed in a series of high-profile law review articles (see, mainly, Bronsteen et al. 2009; Gray 2010; Kolber 2009; Markel and Flanders 2010). Our goal in this essay is to summarize the main points, questions, and findings in this debate and put them within the context of the law and economics of crime literature.

In order to keep the ideas focused, we make two qualifications. (i) We abstract from individual variation in the benefits from crime that may produce variation in the social cost of crime and may have implications for the punishment. To give an example, jaywalking by a surgeon on his way to an urgent operation may be socially less harmful (in fact it may be socially beneficial) than jaywalking by a student who is late for her lecture. Whenever we talk about two different individuals committing the same offense, we assume that the offenses are equivalent in terms of the total social costs as well as the offenders' benefits. (ii) When thinking about punishment, we are

mainly concerned with the consequentialist reasons for punishing individuals, particularly with deterrence, and are abstracting from retributivist approaches (see, e.g., Kolber 2009; Gray 2010; Markel and Flanders 2010, for discussion of retributivism and subjective punishment).

Punishment and Subjective Discomfort

The magnitude of the discomfort that any given nominal punishment creates in a punishee is determined by a number of factors other than the formal punishment itself. Wealth is a prominent, and controversial, example of such a factor. A fine of €50 for jaywalking is likely to produce a different amount of displeasure in a person living off unemployment benefits than in a successful rock star. This is because the goods that the rock star would buy for those €50, such as a pair of luxury socks, would be probably less valuable to her than the value to the unemployed of the goods she would purchase for her €50, perhaps a winter jacket. As a result, the utility reduction produced by this fine may differ across the two offenders. By the same token, however, a prison term may generate different amount of discomfort in a wealthy individual from say a homeless person. This is because the change in lifestyle, the opportunities forgone while incarcerated, and the postrelease repercussions of incarceration are likely to be different for the two individuals (Kolber 2009).

Wealth is not the only determinant of the discomfort an individual punishee experiences, however. Individuals' ability to adjust to prison might vary, for example, a claustrophobe is likely to suffer in prison more than an inmate without such a condition. Some prisons may be better than others in terms of food, wards treatment of prisoners, safety, or hygiene.

These idiosyncrasies in severity of the actual punishment as experienced by the punishee are inherently problematic since they imply that individuals who committed the same offense and receive the same sentence may eventually be punished differently. Put differently, individuals will, in reality, face different *de facto* sanctions for the same offense. That, in turn, means that

individuals' incentives not to offend differ, which may not be desirable.

Mechanisms Causing Variation in De Facto Sanctions

Wealth

As already noted, offenders' wealth affects the severity of monetary as well as nonmonetary sanctions. This is because a sanction can cause different amount of utility reduction, depending on individual's wealth. However, wealth also affects the ability to pay a fine, which has important implications for deterrence. Take an extreme example of a potential offender who possesses zero wealth. As a consequence, the maximum fine such an offender can pay is also zero – no fine generates any deterrence for such an offender. Because deterrence produced by monetary sanctions is clearly limited by offenders' wealth, low wealth offenders may be underdeterred. This has important implications for the design of penal system. Specifically, optimal fines and prison sentences may vary with wealth, at least in theory.

To see this, note a well-known theoretical result by Becker (1968) stating that fines should be as high as possible, that is equal to individuals' wealth, in order to allow for minimal enforcement intensity. At the same time, incarceration should be used as sparsely as possible. This is because law enforcement and nonmonetary sanctions are socially costly, whereas monetary sanctions are just transfers of money from offenders to the state. Whenever a fine can be increased and the probability of apprehension decreased while keeping the expected sanction (and thus deterrence constant), doing so increases social welfare, for fewer resources will be necessary for law enforcement in order to achieve the same crime rate.

This result, however, breaks down when individuals differ by wealth (Polinsky and Shavell 1991). Suppose there is only one type of crime that people commit and all potential offenders are the same, except for their wealth. Initially, a fine is in place that is equal to the total wealth of low-

wealth individuals. Then increasing the fine and lowering the probability of apprehension may result in the same expected sanction for high-wealth individuals, but will result in a *lower* expected sanction for low-wealth individuals, since for them the fine remains the same (equal to their entire wealth), while the probability of conviction decreases. When nonmonetary sanction is available, it is optimal to set fine equal to the total wealth of high-wealth individuals and possibly supplement it with nonmonetary sanction (Polinsky and Shavell 1984). Wealthy offenders will then pay high fine and bear low prison sentence while low-wealth offenders would pay low fine and face more severe prison sentence (Polinsky 2006).

Punishment Lag

Another important source of differences in the severity of punishment is judicial delays. A 12-month sentence that commences, say, 7 years after the crime is very different from a 12-month sentence that commences within a year. This is especially true from the *ex ante* perspective, that is when a decision as to whether commit a crime or not is being contemplated. The effects of delays on the *ex ante* punishment severity are magnified by the fact that criminals tend to discount future more heavily (Åkerlund et al. 2016). More specifically, Mastrobuoni and Rivers (2016) estimate the average yearly discount factor among criminals to be about 0.75.

We illustrate the implications of temporal discounting on sentence severity with the help of a simple model and a numerical example. Consider an individual contemplating whether to commit a crime that would carry a 12-month jail sentence if she is caught. Suppose that S is the value to her of the discomfort from a 12-month prison sentence in case it would commence immediately. The probability of apprehension is fixed and equal to P and let D denote the delay, measured in years, between the day of the crime and the commencement of the prison term. Finally, define the “certainty equivalent of an instantaneous prison sentence” (CEIPS) as the amount of discomfort associated with a hypothetical prison sentence that would commence

immediately after the crime and with 100% probability. CEIPS thus represents the cost of a potential future prison sentence to an offender in terms of a prison sentence that she would begin to serve immediately and is equal to

$$\text{CEIPS} = 0.75^D \cdot P \cdot S,$$

using Mastrobuoni and Rivers’ (2016) estimated discount factor and assuming, for simplicity’s sake, that the offender is risk neutral (and time-consistent). For an offender who is to start serving a 12-month sentence today, CEIPS is simply equal to S . Note that CEIPS is the relevant quantity that enters into the criminal’s determination whether to commit a crime or not, i.e., it is the *ex ante* cost of criminal punishment to the criminal that can be avoided by not committing the crime.

Suppose that about two-thirds of offenders are captured and convicted, i.e., the probability of conviction is $P = 2/3$. As a result, in case of instantaneous conviction, that is when $D = 0$, CEIPS is equal to $(2/3) \cdot S$, which is equivalent to receiving 8 months in prison starting immediately. Now consider a case in which it takes a year to convict a criminal and send her to prison, i.e., when $D = 1$. The implied CEIPS will be $0.75 \cdot (2/3) \cdot S = S/2$, i.e., 6 months in jail. However, court delays are in some cases are much longer. If, for example, the time to conviction is equal to 7 years, that is when $D = 7$, then the associated CEIPS will be $0.75^7 \cdot (2/3) \cdot S = 0.09 \cdot S$, just a little more than 1 month in prison. In other words, a prison sentence delayed by 7 years is worth one-sixth of the same prison sentence that would begin 1 year after the crime.

Judicial delays therefore significantly limit the deterrence that the system of criminal law provides. At the same time, judicial delays may vary from case to case. As a result, two criminals that receive formally equivalent sentence may receive very different de facto punishment, if one is sent to prison quickly and the other with a lag.

In order to compensate for these disparities, Listokin (2007) proposes maintaining constant discounted sentencing terms by adjusting individual sanctions to account for the lag between crime and punishment. By the same token, pretrial

detainees should receive “interest” in addition to credit for time served since their sentences begin earlier and have greater discounted values. Equivalent reasoning, of course, applies to monetary punishment. A fine of €100 due 10 years from now is less severe than the same fine that is due right now. Delayed fines should be higher to compensate for their lower discounted value.

Hedonic Adaptation

Another important mechanism affecting the de facto sentence severity is adaptation. A substantial body of psychological research suggests that people tend to adapt to shocks to their life, positive or negative (see Bronsteen et al. 2009, 2014). A few findings of the existing research stand out from the point of view of criminal sanctions: (i) sanctions, monetary or imprisonment, tend to generate a decline in punishees’ well-being; (ii) this is however followed by a recovery so that the negative shock from a sanction tends to be short-lived; (iii) people tend to forget about past experience of adaptation and therefore tend to overestimate the effects of future shocks to their well-being; (iv) the postrelease effects of prison experience, in the form of poor health, economic effects from the stigma, or worsened family relations generate long-term losses to ex-prisoners’ well-being that are difficult to adapt to.

The patterns of adaptation while in prison and postrelease effects of prison have profound consequences for how punishment affects individuals. As Bronsteen et al. (2009) point out the fact that individuals fail to anticipate adaptation suggest that the social cost of imprisonment is lower than previously thought. This is because adaptation mitigates the losses in punishee’s well-being during incarceration while the individuals’ tendency to ignore this effect mean that punishment retains its deterrent power.

At the same time, these findings may have important implications for the trade-off between fines and imprisonment. If individuals adapt to the negative shock from fines and if they fail to anticipate this pattern, fines too are socially less costly than they would be in the absence of adaptation, while keeping their deterrent effect. To the extent there are other costs related to incarceration itself and the postprison effects on individuals’ well-

being are substantial (for they tend to occur over a longer timespan than the prison sentence) the social costs of incarceration are increased. In addition, these long-term effects may not generate much deterrence due to discounting of future events, as discussed above. On the balance therefore, the economic case for monetary punishment may be stronger than previously thought.

Health

Individuals have different physical and psychological setup. This in turn implies differences in their ability to cope with a punishment. A claustrophobe may not be able to adapt to prison confinement as easily as a person without such condition and so may suffer in prison more. A physically weak punishee may be more likely to become a victim of prison violence, rape, or exploitation. Even in the absence of actual violence, its increased likelihood may put individuals under increased strain.

Additionally, being a prisoner is not a healthy lifestyle (Fazel and Baillargeon 2011). The prevalence of diseases such as HIV, Hepatitis, and Tuberculosis imply a higher risk of infection faced by prisoners. Prisoners also exhibit higher rates of mental disorders and higher suicide rates, particularly among women and juveniles. Importantly, the impact of prison on health does not cease upon the release. Exprisoners suffer from increased prevalence of infectious and stress-related diseases (Massoglia 2008a, b). All of these factors generate differences between individual offenders in the amount of discomfort from a punishment.

Subjective Effects of Other Types Sanctions

The subjective punishment considerations are not limited to incarceration and pecuniary penalties. A variety of other types of punishments are prescribed for various types of crimes, examples include death penalty, community service, shaming punishments, or revocation of driver’s license. The severity of all these types of punishments is codetermined by subjective factors on the side of convicted persons.

Consider perhaps the most severe type of punishment, the termination of the offender's life. A significant element of the punishment from death penalty is the loss of the rest of life. However different people may have a different expected length of life. The life expectancy of young people is larger than that of old people. Moreover the quality of life is reduced by chronic illnesses typical of old age. Therefore the capital punishment is likely to be more severe for a young person than for an old one (see Alberini and Ščasný 2011, for evidence of individual-level heterogeneity in the valuation of life).

Similarly, in the case of community service that can be more burdensome, symbolically or physically, for a white collar senior compared to a blue collar young man. The symbolical burden may be however diminished, possibly to a larger degree for the blue collar worker, by the fact that the service for the public good may not be viewed as shameful. The so-called shaming punishment is a type of alternative sanction whose purpose is to satisfy a popular expectation that punishment expresses an authoritative moral condemnation (see Kahan 1996, for a debate of alternative sanctions and shame). A thief is ordered, for example, to publish his photo with his full name together with a moralistic or self-degrading commentary. However the severity such a sanction will again substantially depend on individual punishee's circumstances. The same punishment may carry much more shame in a village community than in a big anonymous city. It may or may not result in a reputation damage, depending on the actual reputation a punishee possesses.

Potential Benefits of Tailoring Sanctions

Recognizing how nominally equivalent punishment may generate different amounts of discomfort in different punishees is relatively straightforward. Whether and how these factors should enter into the determination of punishment is a more complicated question. Indeed, the United States Sentencing Guidelines Manual discourages judges from calibrating sentences based on convicts' subjective experience (Kolber 2009).

While clearly controversial (see, e.g., Gray 2010; Markel and Flanders 2010), the idea of tailoring sanctions to individual offenders is theoretically appealing. Suppose courts are perfectly informed, that is they have exact knowledge of offenders' benefits from an offense as well as the costs to her from a sanction. Then, if a court could *ex post* set a sanction in each case so that the sanction-related costs to the criminal exceed the benefits from the crime she may have received, each offender would be perfectly deterred. As a result, no sanctions would have to be implemented and crime level would be zero (Shavell 1987).

That may sound like a theoretical possibility, however one can expect that with the progressing scientific knowledge and advances in technology, our ability to measure the discomfort associated with punishment as well as to estimate the benefits from a crime should also improve. We are already beginning to be able to see and measure the neural processes associated with pain and pleasure, although the resulting data are still highly imperfect and their usability at court is limited as of now (see, e.g., Brodersen et al. 2012; Miller 2009; Schulz et al. 2012; for philosophical discussions about advances in neuroscience and the law see Greene and Cohen 2004; Morse 2006, 2011). In the future, however, these developments will inevitably lead to an important dilemma faced by the criminal justice, namely, to what extent should such information be taken into account when determining a criminal punishment?

Experimental Evidence

While legal scholars debate to what extent should the subjective factors be accounted for by the criminal law and sentencing practice, a natural question arises as to whether such legal innovation would conform with people's preferences toward punishment and their view of justice. This question is relevant because for criminal justice to function properly and generate the desired consequences, criminal law and practice should be clear, well understood by its subjects, and should not contradict the common sense of what is just. Two experimental studies designed to

shed light on this question have been recently conducted.

In Montag and Sobek (2014), we run a vignette experiment in which about 200 subjects, mostly business or economics undergraduates, were asked to decide on a hypothetical punishment for a convicted defendant. The subjects could choose any combination of fine and a prison sentence. Each subject was presented with a sequence of two cases, one at a time, and provided stylized facts about each case. The two cases differed from each other only, but significantly, in the wealth of the defendant. In addition, the sequence was randomized across subjects so that one half first received a case with a poor defendant, after which they were presented with the same case featuring a rich defendant. For the second half of the participants, the order of cases was reversed.

This experiment thus generated two types of variation: (i) within-subject variation, when comparing the punishment for the poor with the punishment for the rich assigned by each subject; and (ii) between-subject variation when the decisions of the group that first received the case featuring poor defendant are compared with the decisions of the group that first received the case with rich defendant. Both sources of variation allow us to look at the relationship between wealth and punishment, but with a slightly different interpretation. The within-subject variation is essentially telling us what is the explicit *rule* that the subject would adhere to – under the assumption that their first decision is a precedent, i.e., it is binding when subjects are deciding their second case. The between-subject variation, i.e., comparing the first decisions of the two groups, allows us to look at subjects' "gut feelings" about the appropriate punishment, since this decision is done in the absence of any precedent.

The within-subject results suggest that a slight majority of subjects, about 55%, would give higher fine to a wealthier defendant. At the same time, a vast majority, 87%, assigned the same prison sentence to both defendants, strongly rejecting the notion that rich defendants could receive shorter sentences. In addition, the rules elicited via within-subject variation were highly

consistent with the between-group results: the subjects that first received the rich defendant assigned higher fines on average than the group who first had to punish the poor defendant. However, both groups assigned the same average prison sentence in their first decisions. These results suggest that the subjects had a rather consistent notion of whether and how wealth should enter into the determination of criminal punishment.

However, vignette experiments are subject to criticism due to the absence of incentives. Furthermore, the subjects' decisions in this experiment had no consequences, the punishments assigned were purely hypothetical. At the same time, it is also possible that the subjects did not perceive wealth as a cause of potentially differential effects of prison on the punishee.

Montag and Tremewan (2016) therefore run a laboratory experiment that aims to address these criticisms. The "crime" in this experiment consists of subjects decision whether to take back a donation to a charity (the Red Cross) that would be otherwise subtracted from their earnings at the end of the experiment. They run six experimental sessions with two treatments, monetary and non-monetary: In the monetary treatment, the punishment is a fine between €0 and €5 subtracted from subjects' earnings in the experiment. In the non-monetary treatment, the punishment consists of the punishee having to place between 0 and 100 sliders displayed on the computer screen exactly in the middle (using mouse only) at the end of the experiment.

Importantly, the authors elicit subject's willingness to pay, i.e., the maximal amount each subject would be willing to pay, to avoid placing the sliders. This is a measure of subjective discomfort associated with placing sliders. During this elicitation, the subjects earned between 0 and €5, depending on how quickly they place 100 sliders. In addition, subjects receive a random payment of 0 to €5. So there are two sources of income: earned and a windfall.

Then the punishers were matched with punishees and given three pieces of information: the punishee's willingness to pay to avoid placing the sliders, her earned wealth, and her random wealth. This allows looking at the role of these

factors as punishment determinants. The results of the experiment suggest that that people tend to apply higher fines to wealthier individuals and assign fewer punishment sliders to those with a higher willingness to pay to avoid this tedious task. These results generally seem to support the notion that subjective discomfort should enter into punishment determination.

Summary

Every criminal justice system faces a trade-off between the principle of systemic consistency and the premises of individualized justice. The former demands equal treatment before the law, rule-guided judgment, predictability, and legal certainty. The latter holds that judges need some measure of discretion and flexibility because they should impose sentences that are appropriate according to relevant circumstances of particular case. However, the two principles may not be in conflict once it is recognized that nominally equal treatment may lead to treatment that is unequal de facto.

Legal scholars have recently proposed that improvements in scientific knowledge and advancements in technologies (such as functional magnetic resonance imaging), which allow us to measure subjective perceptions and feelings, need to be and should be incorporated in our penal systems. This would facilitate calibrating the punishment not only for the crime but also to the offender's persona, so that different people experience equally tough punishment for the same crime. These developments raise a number of interesting questions as to the nature, the purpose, and possibly the future of punishment.

The experimental evidence available so far suggests that in the case of monetary punishment, calibration to the individual punishee's circumstances may well conform to people's perception of justice and punishment. In the case of non-monetary punishment, the results are rather inconclusive. More research is required in order to better understand the costs and benefits of accounting for subjective factors in punishment determination.

Cross-References

- ▶ [Becker, Gary S.](#)
- ▶ [Behavioral Law and Economics](#)
- ▶ [Cognitive Law and Economics](#)
- ▶ [Consequentialism](#)
- ▶ [Crime and Punishment \(Becker 1968\)](#)
- ▶ [Crime, Attitude Towards](#)
- ▶ [Crime: Economics of, the Standard Approach](#)
- ▶ [Criminal Sanctions and Deterrence](#)
- ▶ [Death Penalty](#)
- ▶ [Experimental Law and Economics](#)
- ▶ [Judicial Decision-Making](#)
- ▶ [Judicial Delay](#)
- ▶ [Prisons](#)
- ▶ [Retributivism](#)

References

- Åkerlund D, Golsteyn BHH, Grönqvist H, Lindahl L (2016) Time discounting and criminal behavior. *Proc Natl Acad Sci* 113:6160–6165
- Alberini A, Ščasný M (2011) Context and the VSL: evidence from a stated preference study in Italy and the Czech Republic. *Environ Resour Econ* 49: 511–538
- Becker GS (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Brodersen KH, Wiech K, Lomakina EI, Lin C-s, Buhmann JM, Bingel U, Ploner M, Stephan KE, Tracey I (2012) Decoding the perception of pain from fMRI using multivariate pattern analysis. *NeuroImage* 63:1162–1170
- Bronsteen J, Buccafusco C, Masur J (2009) Happiness and punishment. *Univ Chic Law Rev* 76:1037–1082
- Bronsteen J, Buccafusco C, Masur JS (2014) Happiness and the law. University of Chicago Press, Chicago
- Fazel S, Baillargeon J (2011) The health of prisoners. *Lancet* 377:956–965
- Gray D (2010) Punishment as suffering. *Vanderbilt Law Rev* 63:1619–1693
- Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. *Philos Trans Roy Soc Lond B Biol Sci* 359:1775–1785
- Kahan DM (1996) What do alternative sanctions mean? *Univ Chic Law Rev* 63:591–653
- Kolber AJ (2009) The subjective experience of punishment. *Columbia Law Rev* 109:182–236
- Listokin Y (2007) Crime and (with a lag) punishment: the implications of discounting for equitable sentencing. *Am Crim Law Rev* 44:115–140
- Markel D, Flanders C (2010) Bentham on stilts: the bare relevance of subjectivity to retributive justice. *Calif Law Rev* 98:907–988

- Massoglia M (2008a) Incarceration as exposure: the prison, infectious disease, and other stress-related illnesses. *J Health Soc Behav* 49:56–71
- Massoglia M (2008b) Incarceration, health, and racial disparities in health. *Law Soc Rev* 42:275–306
- Mastrobuoni G, Rivers DA (2016) Criminal discount factors and deterrence. IZA discussion paper no. 9769. Institute for the Study of Labor, Bonn
- Miller G (2009) Brain scans of pain raise questions for the law. *Science* 323:195
- Montag J, Sobek T (2014) Should Paris Hilton receive a lighter prison sentence because she's rich: an experimental study. *Kentucky Law J* 103:95–125
- Montag J, Tremewan J (2016) Let the punishment fit the criminal: an experimental study. ISE working paper no. 3. International School of Economics, Kazakh-British Technical University, Almaty
- Morse SJ (2006) Brain overclaim syndrome and criminal responsibility: a diagnostic note. *Ohio State J Crim Law* 3:397–412
- Morse SJ (2011) Neuroscience and the future of personhood and responsibility. In: Rosen J, Wittes B (eds) *Constitution 3.0: freedom and technological change*. Brookings Institution Press, Washington, DC, pp 113–129
- Polinsky AM (2006) The optimal use of fines and imprisonment when wealth is unobservable. *J Public Econ* 90:823–835
- Polinsky AM, Shavell S (1984) The optimal use of fines and imprisonment. *J Public Econ* 24:89–99.
- Polinsky AM, Shavell S (1991) A note on optimal fines when wealth varies among individuals. *Am Econ Rev* 81:618–621
- Schulz E, Zherdin A, Tiemann L, Plant C, Ploner M (2012) Decoding an individual's sensitivity to pain from the multivariate analysis of EEG data. *Cereb Cortex* 22:1118–1123
- Shavell S (1987) The optimal use of nonmonetary sanctions as a deterrent. *Am Econ Rev* 77:584–592

to strengthen the accountability, transparency and integrity of government and public sector entities as well as to make a difference to the lives of citizens. Moreover, SAIs should notify policy makers of the emerging trends and future challenges in order to help them to take action to avoid crises. Despite their differences and various models from country to country, SAIs are often mentioned as “*watchdogs*” or “*cornerstones of transparency and accountability*” of public financial management and they constitute essential components in a democratic system.

Definition

Each country has a supreme audit institution (SAI), which is responsible for exercising oversight over government revenues and expenditures and publishing an annual report to inform the parliament and the public.

These institutions are organized according to different models (i.e., Anglo-Saxon model (National Audit Office), Court of Accounts (Cour des Comptes), etc.), which reflect the different policies and systems of governance and different regimes and cultures of each country. Despite their differences, their main objective is to oversee the management of public funds as well as the quality and reliability of financial data included in the reports addressed to governments.

Supreme Audit Institutions

Georgia Kontogeorga
International Board of Auditors for NATO
(IBAN), Brussels, Belgium
University of Paris 1- Pantheon Sorbonne, Paris,
France

Abstract

Supreme Audit Institutions (SAIs) are independent external services charged with the audit of public funds. Their main objective is

Introduction

Public expenditure in a democratic society should rely on an effective financial management and control system. “SAIs help the executive to make the best possible use of public funds, in other words to ensure that the political objectives of the expenditure are achieved at a minimum cost and that the accounts drawn up are transparent. Furthermore, the publication of audit findings enables citizens to become familiar with and to legitimise the actions of their government and representatives” (Vallés 2005, Foreword, p. xi).

According to the International Standards for Supreme Audit Institutions (ISSAIs), SAIs are external audit services and do not constitute part of the organizational structure of the audited entities (INTOSAI, ISSAI 1, Sect. 3, § 1). Their legal mandates, reporting relationships and effectiveness, vary, reflecting different governance systems and government policies (Stapenhurst and Titsworth 2001).

However, each country has the choice of its own SAI model, and the way in which it operates varies greatly from one country to another.

Theoretical Framework

SAIs vary according to the different legal, financial, and political traditions of each country. Although the literature mentions several categories of SAIs, according to ISSAIs (1000, § 47), the three most common systems are:

The Jurisdictional Model (Napoleonic System or Court of Accounts)

In this system, the SAI, also known as the Court of Accounts (Court of Audit), is a judicial and administrative authority independent of legislative and executive powers. The SAI is an integral part of the justice system, making judgments on government compliance with laws and regulations and ensuring that expenditures are justified.

According to this model, the SAI is an integral part of the judicial system functioning independently of the executive and legislative powers. One of the main differences of the judicial model is the fact that public officials are normally held personally responsible for any unauthorized or illegal payments in installments. The Court of Accounts is competent for imposing penalties where illegal transactions are found.

Under judicial models, the SAI often has significant independence from the parliament. Susceptibility to political influence should be relatively low through this arm's length relationship but depends on the autonomy of the judiciary and the specific terms of tenure for judges. A potential incentive problem can arise in

circumstances where financial penalties imposed flow to the court itself rather than back to central government budgets (Clench 2010).

Judicial systems traditionally concentrate on compliance with detailed rules and regulations to ensure that money has been properly spent. However, there is often less focus on wider financial management issues relating to the economy, efficiency, and effectiveness of expenditure (Blume and Voigt 2011).

The drawbacks include that the system is inherently complex and can be difficult to operate effectively if there is limited human capacity (in either numerical or knowledge terms) or a lack of other resources. In such circumstances, controls may be applied mechanistically or not applied at all. This model is widespread in countries that have a French legal origin and is followed in France, Italy, Spain, Portugal, Brazil, Belgium, El Salvador, Turkey, Greece, and in most Francophone and Latin American countries (DFID 2004).

The Anglo-Saxon Model (Westminster Model or Monocratic Model or Auditor General-National Audit Office Model)

In this model, the SAI is headed by a single auditor general (comptroller and auditor general – C&AG) and often acts as an auxiliary body in the parliament. The Office of the Auditor General is an independent body reporting to the parliament, namely, the parliamentary committee (PAC). Traditionally, the head of the PAC is a member of the opposition (Blume and Voigt 2011).

With one individual at the top of the SAI, power is considerably centralized. As a result, much of the integrity of the institution rests on the credibility and personal motivations of the auditor general. However, susceptibility to political influence should generally be low, given that the auditor general answers to the legislature as a whole, and typically cannot easily be removed from office (Clench 2010).

The Office of the Auditor General does not have legal competence, but when necessary, the results can be transmitted to the judicial authorities for further investigation (Stapenhurst and

Titsworth 2001). Its priority is to provide support to the parliament and secondarily to provide assistance to judicial institutions.

The Westminster model has a strong focus on financial audit and on the value for money with which audited bodies have used their resources, with less emphasis on compliance with detailed legislation and regulations (DFID 2004). This model focuses much more on financial aspects than the Napoleonic model.

It is widespread in Anglo-Saxon countries, the United States, Australia, India, the United Kingdom, Canada and Latin America, Chile, Colombia, Mexico, Peru, in several Caribbean and Pacific countries, and in English-speaking African countries.

The Collegiate Model (Board Model)

In this model, the emphasis is on the collective decisions of the organization, but without legal responsibility. Its main task is to analyze public revenue and expenditure and to present the report on its findings to the parliament.

The basic structure of an SAI's relationship with the parliament, the government, and the administration is very similar to the Westminster model except for the internal structure of the audit institution. This system is independent of the executive and assists the parliament in the exercise of oversight but is not managed by a single person (as in the Westminster model) but by a board made up of several people who make decisions on operations and management by consensus (Clench 2010; Blume and Voigt 2011). This can be both an advantage and a disadvantage: the result of the work does not depend so much on one person, but on the other hand, its collective structure can lead to a complex and slow decision-making process. Clench (2010) mentions that depending on the degree of autonomy of each board member, there is a potential for multiple and conflicting interests. This can undermine the consistency of auditing across different government bodies with difficulties in imposing minimum standards. These risks can be minimized by the use of international standards that may be a point of reference for achieving consensus across the board.

This model is very widespread in Asia, Indonesia, Japan, Argentina, and the Republic of Korea.

Discussion

Generally, the Napoleonic model is more common in countries where the system has a French legal origin and the Westminster model in countries where the system has a common law origin, but it is also widespread in countries such as Chile and Peru, which, although included in the French system, have SAIs following the Westminster model (Blume and Voigt 2011).

Santiso (2006) points out the most important differences between these models, which include:

- Time of the audit: (ex ante or ex post audit)
- The nature of the audit: if the SAI focuses on compliance or performance (audit performance)
- The effects of the audit: (follow-up of audit recommendations)
- Their status: (the legal status of audit decisions)

The OECD mentions that the two major types of SAIs are generally described as "courts" and "offices." However, in practice, SAIs are unique hybrids that do not easily fit into a traditional separation of powers and can combine various elements of these models.

Given above, it is obvious that in the field of external public financial auditing, almost every country has a unique system, different even from the country that seems closest to it in terms of norms.

References

- Blume L, Voigt S (2011) Does organizational design of supreme audit institutions matter? A cross-country assessment. *Eur J Polit Econ* 27(2):215–229
- Clench B (2010) Auditing the auditors – international standards to hold supreme audit institutions to account, March 2010 Number: 238, U4 Expert Answer. Available at: <https://fr.scribd.com/document/351779985/Auditing-the-Auditors> Accessed 30 July 2017
- DFID (Department for International Development) (2004) Briefing characteristics of different external

- audit systems. Part of the policy division info series. Ref no: PD Info 021. © Crown copyright 2004
- INTOSAI – ISSAI 1. The lima declaration of guidelines on auditing precepts. Available at: http://www.issai.org/en_us/site-issai/issai-framework/. Accessed 6 Jan 2017
- Santiso C (2006) Improving fiscal government and curbing corruption: how relevant are autonomous audit agencies? *Int Publ Manag Rev* 7:97–108
- Stapenhurst R, Titsworth J (2001) Features and functions of supreme audit institutions. The World Bank Prem notes, Public sector, Number 59, October
- Vallés JMF (2005) Foreword. In: Crespo MG (ed) *Public expenditure control in Europe: coordinating audit functions in the European union*. Edward Elgar, p. xi. Cheltenham, UK, Northampton, MA, USA

Sustainable Development, and Code of Best Practices

- [Codes of Conduct](#)

Systemic Risk

- [Risk Management, Optimal](#)

Takings

Thomas J. Miceli and Kathleen Segerson
Department of Economics, University of
Connecticut, Storrs, CT, USA

Abstract

This entry discusses the economics of eminent domain, which is the government's power to take or regulate privately owned property for the common good. It discusses the origins of the power as well as its limits, particularly as embodied in the public use and just compensation requirements. It also reviews the economics literature on how eminent domain affects incentives for efficient land use.

JEL codes: K11, K32

Definition

Takings: The acquisition of privately owned property by the government using its power of eminent domain.

The right of the community to exercise eminent domain and thereby expropriate individually held property for the common good seems to have been universally acknowledged and practiced by societies throughout history. As Reynolds (2010, p. 11) asserts, "...the principle that land might

be taken from individuals when the community needed it has been so generally accepted that it did not need to be stated or argued about until recent times." For example, the very phrasing of the Takings Clause in the US Constitution, which states "nor shall private property be taken for public use, without just compensation," implicitly acknowledges that governments have the right to take property; its main purpose is to limit that power by requiring that the taking be for a public purpose and that just compensation be paid.

The idea that compensation should be paid when land is taken also has ancient roots and is therefore taken for granted in the law of most countries. Even in communist China, where land is collectively owned, the 1982 Constitution provides for farmers to receive compensation for "use rights" (as proxied by crop yield) when their land is taken in the "public interest." In cases of physical takings, the debate about compensation has always been about the *amount* of compensation that should be paid, not about whether it should be paid. However, the question about the specific meaning of "common good" or "public purpose" seems to be primarily an issue in American law and likely reflects the explicit inclusion of "public use" as a limitation on eminent domain in the US Constitution.

Takings questions also arise in contexts where land is not physically taken, but its use is restricted in some way, for example, through land use regulations. Such regulations are often termed

“regulatory takings.” In addition to the provisions for compensation for acquisitions under eminent domain, laws governing compensation for land use regulations exist in most countries, though they tend to vary greatly in the rights to compensation that they afford. In a study of 13 countries, including the USA, Canada, several European countries, Israel, and Australia, Alterman (2010) found a range of compensation rights, from very little to nearly full, with the USA lying in between. Notably, there seems to be no way to correlate a country’s laws with its geographic, legal, demographic, or other attributes.

From a theoretical perspective, the issue of eminent domain has generated an enormous amount of legal scholarship and case law. Much of the discussion focuses on the meaning of the “public use” and “just compensation” requirements in the US Constitution, which together restrict the scope of government acquisition of private property. However, the economic questions regarding the appropriate limits on the power to take property or restrict its use are relevant in all countries where governments are granted these powers. The economic literature asks whether there is an economic rationale for bestowing the right to acquire property or restrict its use without the owner’s consent on the government alone (and not on private parties) and seeks to understand the limits of that right, as embodied in the public use and just compensation requirements. The primary focus is on how granting and restricting that right affects incentives for efficient land use.

Public Use

On its face, a public use requirement would seem to limit the use of eminent domain to provision of public goods like highways, airports, or parks. This interpretation is appealing both in terms of the plain meaning of the phrase “public use” and in view of the well-accepted role of the government in providing these types of goods. Specifically, because public goods have the characteristic of non-excludability, meaning consumers can enjoy the benefits of the goods without first

having to pay for them (i.e., to free ride), ordinary markets will underprovide them. Economists have long argued, therefore, that the government should provide public goods so as to ensure that the efficient quantity will be supplied and then use its tax powers to coerce consumers to contribute to the cost. In this sense, government provision and financing of public goods amount to a kind of “forced purchase” by consumers. Merrill (1986) refers to the proposition that the government alone should have the power of eminent domain and should only use it to provide public goods, as the “ends approach” to defining public use because it concerns the use to which the taken land will be put (also see Miceli (2011), Chap. 2).

The logic of the ends approach, however, does not explain why the land (and possibly other inputs) needed to provide a public good must also be forcibly acquired from the owner(s). The economic justification for this form of coercion, which involves a “forced sale” from input owners to the government, is a different kind of market failure, referred to as the holdout problem. The holdout problem arises in the context of any development project requiring the assembly of multiple contiguous parcels of land. The difficulty developers face in this context is that once the scope of the project becomes public knowledge, individual landowners acquire a kind of monopoly power, given that each parcel is essential for completion of the overall project. Imagine, for example, a road or railroad builder who has decided on the optimal route and has assembled several parcels. Refusal of any additional owners to sell would greatly increase the cost of the project, and as a result, all owners can hold out for prices above their true valuations. The likely result is underprovision of projects involving assembly (Miceli and Segerson 2012). One solution to the holdout problem is to take away an owner’s right to refuse a sale at the offered price. The power of eminent domain represents such a forced sale at a price set by the court. Merrill (1986) refers to this justification for eminent domain – i.e., as a response to the holdout problem – as the “means approach” to public use because it concerns the manner by which land for a government project is acquired.

From an economic perspective, the ends approach is the correct justification for eminent domain, but that does not mean that it is the correct interpretation of the “public use” requirement. The reason for this paradox is twofold. The first derives from the fact that the holdout problem and the attendant inefficiencies are not unique to public projects. Private developers also face holdouts for large-scale commercial developments, as do local development authorities undertaking urban renewal. In some cases, private developers can mitigate the problem by means of secret purchases (Kelly 2006), but at some point the scope of the project becomes apparent, and the holdout problem arises. The logic of the above argument therefore implies that, from an efficiency perspective, the power of eminent domain should be available to any developer, private or public, engaged in land assembly.

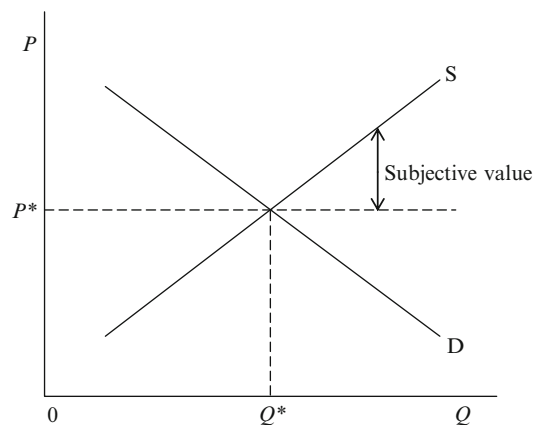
Second, in explicating the constitutional basis for a public use requirement, courts routinely invoke the ends approach by appealing to the “public purpose” behind the project in question. This is an easy argument for truly public goods because the benefits of the resulting project are available to all, but the logic is less persuasive when the project is largely private, as has been true in most US Supreme Court cases involving the public use issue. For example, in the case of *Kelo v. New London*, despite the fact that the primary beneficiaries were private businesses, the US Supreme Court upheld the city’s use of eminent domain to assemble land for purposes of a redevelopment project by emphasizing the jobs and enhanced tax revenues that would materialize from the project (125 S. Ct. 2655, 545 U.S. 469 (2005); also see Merrill (1986) and Kelly (2006) for a discussion of other notable cases in this vein). As Merrill (1986, p. 67) notes, however, this strategy is not unusual, as courts strain to make their reasoning consistent with the plain meaning of the US Constitution. The difficulty in defining the public use requirement therefore comes down to a divergence between the economic and legal justifications for eminent domain. Whereas economists focus on the use of eminent domain as an efficient response to the holdout problem, regardless of the context,

courts often emphasize the need to identify some public benefits as flowing from the use of government coercion.

Just Compensation

A second requirement for the use of eminent domain is generally that the landowner must be paid “just compensation.” Although no further definition of what amount of compensation is “just” is given, courts have typically interpreted it to mean “fair market value.” Economists, however, argue that this measure systematically undercompensates owners because it ignores their “subjective value” (Fischel 1995a). The idea can be easily illustrated in a simple supply-and-demand diagram for some particular category of land (Miceli and Segerson 2007, p. 20) (Fig. 1).

In the graph, the demand curve represents the willingness to pay for individual parcels of land by potential buyers, while the supply curve represents the reservation price, or opportunity cost, of current owners. The equilibrium price, P^* , can be interpreted as the market value of a unit of the land in question, reflecting the price of recently sold parcels. Note that the price partitions the market into the parcels that are sold over some time period (those between 0 and Q^*) and those that are not sold (those to the right of Q^*). In other words, owners between 0 and Q^* voluntarily sold because the price exceeded their opportunity cost,



Takings, Fig. 1 Subjective value

but those to the right of Q^* did not sell because the price was below their opportunity cost.

Now suppose that one of the parcels to the right of Q^* is taken by the government for use in a public project, necessitating the payment of just compensation. In order for the transaction to be consensual on the part of the seller, compensation would have to be set at or above the relevant point on the supply curve. The problem with this “value-to-the-owner” measure of compensation is that it is a private information of the seller and therefore can be misrepresented, thus creating the risk of a holdout problem (Knetsch and Borcharding 1979). Market value, in contrast, is observable but clearly undercompensates owners, with the difference between market value and value to the owner representing the owner’s subjective value, as shown in the graph.

Epstein (1985) has argued that the loss that market-value compensation imposes on landowners is only justifiable in two circumstances. The first is when the surplus created by the taking – the difference between the value of the parcel in its new use minus the owner’s opportunity cost – is widely distributed rather than being concentrated in a few hands. In essence, Epstein is arguing for the ends approach to public use – that is, eminent domain should only be used to provide public goods – for in that case the distributional requirement is clearly satisfied (also see Ulen (1992)). Epstein’s argument is motivated primarily by concerns about distributional equity, but it also reflects the idea that when gains from coerced transactions are dispersed, private interests will not find it worthwhile to engage in rent-seeking efforts to acquire eminent domain power.

The other circumstance in which undercompensation in a monetary sense is justifiable according to Epstein is when the government action provides “in-kind” compensation. An example is a zoning ordinance that deprives owners of some uses of their land, particularly those that would be harmful to their neighbors. Although an owner could claim that this type of regulation prevents him from engaging in certain profitable activities, like opening a gas station in a residential area, that “loss” is calculated based on his unilateral departure from an efficient land use pattern in which all other owners refrain from

engaging in the externality-producing activity. In fact, when all landowners comply with the regulation, and assuming the regulation is efficient, their property values are enhanced relative to a situation in which no regulations are imposed. In this sense, all owners receive in-kind compensation from the regulation and therefore are not entitled to further monetary compensation. This is in contrast to regulations that “single out” individual owners to surrender their land for the benefit of others, in which case monetary compensation is required.

Compensation and Land Use Incentives

The preceding discussion of compensation has focused on its role in limiting excessive takings. However, most recent economic scholarship on eminent domain, especially since the publication of the seminal article by Blume et al. (1984) (hereafter BRS), has been to evaluate the incentive properties of the compensation rule regarding the land use decisions of property owners whose land is at risk of a taking. The incentive effects of compensation can be shown using a simplified version of the BRS model. Specifically, let $V(x)$ be the value of a piece of land after x dollars have been invested in improvements, where $V' > 0$ and $V'' < 0$. Also let p be the probability that the land will be taken by eminent domain, let B be the value of the land in public use, and let $C(x)$ be the amount of compensation paid to the landowner, which may depend on the amount of improvements. Initially, both p and B will be treated as exogenous.

The timing of events is as follows. First, the landowner chooses the level of investment, which is irreversible, taking the probability of a taking and the compensation rule as given. Once x is in place, the taking decision is made. If the land is taken, the landowner is paid compensation and the land is converted to public use. The requirement that x must be chosen before the taking decision occurs is crucial because otherwise, the landowner could simply wait until the taking decision is made and only invest if the land is not taken. (This assumption is not restrictive in the sense that private land is always subject to a taking risk.)

Similarly, the irreversibility of x precludes salvaging the cost of the investment if the land is taken.

In this setting, the socially optimal level of investment maximizes the expected value of the land:

$$(1 - p)V(x) + pB - x. \quad (1)$$

The first-order condition defining the optimal investment, denoted x^* , is

$$(1 - p)V'(x) - 1 = 0. \quad (2)$$

Note that the resulting investment is decreasing as the probability of a taking increases and is generally less than what the owner would invest in the absence of a taking risk, reflecting the fact that any improvements are lost if the land is converted to public use.

The actual level of investment by the owner maximizes the expected private value of the land, which includes the expected amount of compensation:

$$(1 - p)V(x) + pC(x) - x. \quad (3)$$

The resulting first-order condition is

$$(1 - p)V'(x) + pC'(x) - 1 = 0. \quad (4)$$

Comparing this to Eq. 2 shows that $C' \equiv 0$ is necessary and sufficient for efficiency – that is, compensation must be lump sum. Intuitively, compensation potentially creates a moral hazard problem. If, for example, landowners expected to be fully compensated, or $C(x) = V(x)$, they would overinvest because they would ignore the possibility that the land might be taken, which would render any improvements worthless. At the other extreme, zero compensation, or $C(x) \equiv 0$, results in the efficient level of investment. This represents the famous “zero compensation result” from BRS. (Of course, any positive lump-sum amount would also be efficient.)

The conclusion that zero compensation is efficient, however, is at odds both with general notions of fairness and constitutional requirements for just compensation. This raises the question of whether this result would hold under a more realistic depiction of the taking decision. For example, the above

argument assumes that the probability that the land will be taken is independent of its private value. In reality, a government that behaves in a benevolent or “Pigovian” way (see Fischel and Shapiro 1989) will only take land when it is efficient to do so, thereby making the probability of a taking dependent on the landowner’s decision. Suppose, for example, that the value of the land in public use is a random variable whose value is only observed after landowners have made their investment decisions. Upon observing the realized value of B , the government takes the land if $B \geq V(x)$, that is, if the land is more valuable in public use, given its current value to the owner. The resulting probability of a taking from the landowner’s perspective, prior to the choice of x , is now $1 - F(V(x))$, where $F(V(x))$ is the probability that $B \leq V(x)$, with $F' > 0$.

The expected social value of the land in this case is

$$\begin{aligned} & F(V(x))V(x) \\ & + [1 - F(V(x))]E[B|B \geq F(V(x))] \\ & = F(V(x))V(x) + \int_{V(x)}^{\infty} B dF(B) - x, \end{aligned} \quad (5)$$

and the first-order condition defining x^* is

$$F(V(x))V'(x) - 1 = 0, \quad (6)$$

which is the analog to Eq. 2 with $F(V(x)) = 1 - p$. The expected private value is

$$F(V(x))V(x) + [1 - F(V(x))]C(x) - x, \quad (7)$$

and the first-order condition defining the privately optimal level of investment is

$$\begin{aligned} & F(V(x))V'(x) + [1 - F(V(x))]C'(x) \\ & + F'(V(x))V'(x)[V(x) - C(x)] - 1 = 0. \end{aligned} \quad (8)$$

Note that $C' \equiv 0$ is no longer sufficient for efficiency, but it is still necessary. Now, in addition to being lump sum, compensation must also equal the value of the land at its efficient level of investment; that is, $C = V(x^*)$. The reason for this additional requirement is that landowners view the probability of a taking as depending on their

level of investment. In particular, by investing more (less), they can reduce (increase) that probability of a taking because it becomes less (more) likely that $B \geq V(x)$. Consequently, if compensation is less than full ($C < V(x)$), owners will overinvest to decrease the probability of a taking, and if compensation is more than full ($C > V(x)$), they will underinvest to increase that probability. Combining this with the lump-sum requirement yields $C = V(x^*)$ (Miceli 1991).

Hermalin (1995) shows that two other compensation rules will also achieve the efficient outcome in this context. The first requires the government to pay owners the full value of the land in public use, or $C = B$. In this case, landowners invest efficiently because they internalize the full social value of the land. Alternatively, compensation could be set at zero, but the landowner could be given the option to “buy back” the land in the event of a taking for a price equal to the realized value of the land in public use, B . The difference between these two rules depends on the assignment of property rights in the land when it potentially has public value. Under the first rule, the landowner holds the right, whereas under the second, society holds the right.

The preceding discussion assumed that the government behaved in a socially benevolent way in making its taking decision, but many have argued that an important reason for requiring compensation for takings is to prevent the government from converting too much land to public use (Johnson 1977). A government that considers the dollar costs of a taking as embodied in the compensation rule, rather than the true opportunity costs, is said to have “fiscal illusion” (BRS). To reflect this view of government behavior, suppose that once B is realized, a taking occurs if $B \geq C(x)$. It follows that “full” compensation will be necessary to prevent excessive takings, but simply setting $C(x) = V(x)$ will revive the moral hazard problem that initially gave rise to the BRS no-compensation result. One solution is the lump-sum rule, $C = V(x^*)$, which will simultaneously eliminate moral hazard by landowners (because compensation is lump sum) and induce only efficient takings by the government (because compensation is full).

Consider also the two rules proposed by Hermalin, both of which were shown above to induce efficient investment. Under the rule that sets $C = B$, the government will be indifferent between taking the land and not taking it for any realization of B . The landowner, however, will only want efficient takings to occur – that is, those for which $B \geq V(x)$. Thus, efficiency of the takings decision will depend on whether the government conforms to the wishes of the landowner. As for the buyback rule, since compensation is nominally set at zero, the government will initiate a taking for any $B > 0$, but the landowner will buy back the land whenever $V(x) \geq B$ or whenever a taking is not efficient. Thus, the outcome will be efficient.

A final compensation rule that balances the incentives of landowners and the government is the threshold rule proposed by Miceli and Segerson (1994). According to this rule, the government pays full compensation if it acts inefficiently to take (or regulate) the land, but it pays zero compensation if it acts efficiently. Formally, the rule is written as follows:

$$C = \begin{cases} V(x), & \text{if } B < V(x^*) \\ 0, & \text{if } B \geq V(x^*) \end{cases} \quad (9)$$

The efficiency of this rule is established as follows. Assume the landowner invested efficiently. Then, once B is realized, if $B \geq V(x^*)$, the taking is efficient and compensation is zero. Thus, a government with fiscal illusion will incur a net benefit of $B > 0$ and so will go ahead with the taking. In contrast, if $B < V(x^*)$, a taking is not efficient and compensation is full. Thus the government will incur a net loss of $B - V(x^*) < 0$ and so will refrain from the taking. In both cases, the government acts efficiently. Moving back to the land use decision, the landowner, who anticipates the government’s behavior, will choose x to maximize

$$F(V(x^*))V(x) - x, \quad (10)$$

which has x^* as its solution. The Nash equilibrium therefore involves efficient behavior by both the landowner and the government.

The noteworthy feature of the compensation rule in Eq. 9 is that, in an efficient equilibrium, no compensation is paid for takings. This is clearly contrary to legal practice in takings cases involving physical acquisitions where compensation is typically required, but, as will be discussed below, it is consistent with not paying compensation in regulatory takings cases.

A final class of takings models, referred to as “constitutional choice” models, views landowners as designing the compensation rule from behind a veil of ignorance regarding which parcels will be taken for public use. In this setting, all landowners are equally at risk of having their land taken, given that it is efficient to devote some land to public use, but landowners also know that any money paid to takings victims must be raised by taxes assessed on all landowners. Thus, rational individuals will presumably account for both sides of the public budget in designing the compensation rule and will therefore not be overly stingy (in case their land is taken) or overly generous (in case it is not).

The prototypical model of this sort is by Fischel and Shapiro (1989) (also see Nosal (2001)), which is identical to the BRS model except that the probability of a taking is written as $p = s/n$, where n is the total number of parcels in the jurisdiction and s is the number that will be randomly taken. The public benefit is written as $B(s)$, which will be enjoyed by all landowners (including those whose land is taken), where $B' > 0$, $B'' < 0$. Compensation is written as a fraction of property value, or $C = \alpha V(x)$, where α is a nonnegative constant, and per-person tax liability is also assessed as a fraction of property value, or $T = tV(x)$, where t is the tax rate. The resulting expected wealth of a landowner is

$$(1 - p)V(x) + pC + B(s) - T - x \\ = (1 - p + p\alpha - t)V(x) + B(s) - x. \quad (11)$$

The landowner chooses x to maximize this expression, taking p , s , α , and t as given. The resulting first-order condition is

$$(1 - p + p\alpha - t)V'(x) - 1 = 0. \quad (12)$$

As for the government, if it has fiscal illusion, it will choose s , the number of parcels to take, to maximize

$$nB(s) - sC. \quad (13)$$

The resulting first-order condition is

$$nB'(s) - C = 0. \quad (14)$$

Equations 12 and 14 jointly determine the optimal behavior of landowners and the government, given the compensation rule, as reflected by α , and the tax rate t . As noted, these parameters are determined by citizens from behind a veil of ignorance to maximize overall welfare, subject to a balanced budget condition. The latter is given by $sC = nT$ or

$$p\alpha V(x) = tV(x) \quad (15)$$

for any x . It follows that $p\alpha = t$, which immediately implies that $x = x^*$ in Eq. 12. Thus, landowners invest efficiently regardless of the value of α . Although both the compensation rule and the proportional property tax are potentially distortionary with respect to the choice of x , in the current model the two distortions exactly offset through the budget constraint, resulting in an efficient level of investment for any amount of compensation (Miceli 2008). Finally, substituting for C in Eq. 14 implies $nB'(s) - \alpha V(x^*) = 0$. If $\alpha = 1$, this becomes

$$nB'(s) = V(x^*), \quad (16)$$

which says that parcels should be taken until the aggregate benefit from the last parcel taken just equals the opportunity cost. This, of course, is the Samuelson condition for efficient provision of a pure public good. Thus, if compensation is full, the efficient level of the public good is provided. This, therefore, is the choice citizens would make in designing the optimal compensation rule.

Regulatory Takings

Much more pervasive than physical acquisitions are government regulations that reduce the value of land without seizing title to it. Examples include zoning, environmental and safety regulations, and historic landmark designations. As noted, many courts have granted the government broad police powers to enact such regulations in the public interest without requiring them to compensate landowners, but in some cases the regulation goes so far in reducing the value of the regulated land that courts have declared it to be a regulatory taking for which compensation is due. The question is where the dividing line should be between compensable and non-compensable regulations (Michelmann 1967; Fischel 1995b).

From an economic perspective, there is no substantive difference between a government action that seizes land for purposes of providing a public good and one that merely regulates that same property for purposes of preventing an external harm. In both cases, the government imposes costs on some landowners in order to confer benefits on others. The only difference is the extent of the taking. This point is clear in the context of the above models, which apply equally to full or partial takings. From a legal perspective, however, the two types of cases are often treated very differently by courts – as noted, full takings typically require payment of compensation, whereas partial takings do not. This dichotomy poses a significant challenge for developing a positive economic theory of the compensation question.

One answer is provided by the threshold compensation rule in Eq. 9, which, recall, only requires compensation to be paid for government actions that are inefficiently imposed. In this sense, the rule resembles the famous test for compensation announced by the US Supreme Court in the landmark case of *Pennsylvania Coal v. Mahon* (260 U.S. 393, 1922). According to this rule, referred to as the “diminution of value” test, compensation is only due if a regulation “goes too far” in reducing the landowner’s value, where, in light of Eq. 9, “too far” can be interpreted as

“inefficient.” The efficiency of the Nash equilibrium under this rule, as demonstrated above, implies that compensation will be rarely paid, which is consistent with legal practice in many regulatory takings cases.

Another perspective on the distinction between full and partial takings as regards compensation is Epstein’s idea, discussed above, that compensation need not always be monetary – it can also be in kind. Regulations that are enacted for the purpose of preventing externalities impose costs on individual landowners by limiting those things they can do with their property, but they also confer reciprocal benefits that serve as in-kind benefits. If the regulation is efficient, the benefits exceed the costs on average, which justifies non-payment of monetary compensation. In contrast, physical takings are more likely to “single out” a small number of landowners to bear costs for the benefit of many, as when land is taken from a few to provide a public good. Since in-kind compensation is not generally available to these takings victims, monetary compensation is necessary to satisfy a just compensation requirement.

Acknowledgments We acknowledge the input of James Wen.

References

- Alterman R (2010) Takings international: a comparative perspective on land use regulations and compensation rights. American Bar Association, Chicago
- Blume L, Rubinfeld D, Shapiro P (1984) The taking of land: when should compensation be paid? *Q J Econ* 99:71–92
- Epstein R (1985) Takings: private property and the power of eminent domain. Harvard University Press, Cambridge, MA
- Fischel W (1995a) The offer/ask disparity and just compensation for takings: a constitutional choice approach. *Int Rev Law Econ* 15:187–203
- Fischel W (1995b) Regulatory takings: law, economics, and politics. Harvard University Press, Cambridge, MA
- Fischel W, Shapiro P (1989) A constitutional choice model of compensation for takings. *Int Rev Law Econ* 9:115–128
- Hermalin B (1995) An economic analysis of takings. *J Law Econ Organ* 11:64–86

- Johnson M (1977) Planning without prices: a discussion of land use regulation without compensation. In: Siegan B (ed) *Planning without prices*. Lexington Books, Lexington
- Kelly D (2006) The ‘public use’ requirement in eminent domain law: a rationale based on secret purchases and private influence. *Cornell Law Rev* 92:1–65
- Knetsch J, Borchering T (1979) Expropriation of private property and the basis for compensation. *Univ Toronto Law J* 29:237–252
- Merrill T (1986) The economics of public use. *Cornell Law Rev* 72:61–116
- Miceli T (1991) Compensation for the taking of land under eminent domain. *J Inst Theoretical Econ* 147:354–363
- Miceli T (2008) Public goods, taxes, and takings. *Int Rev Law Econ* 28:287–293
- Miceli T (2011) *The economic theory of eminent domain: private property, public use*. Cambridge University Press, New York
- Miceli T, Segerson K (1994) Regulatory takings: when should compensation be paid? *J Leg Stud* 23: 749–776
- Miceli T, Segerson K (2012) Land assembly and the hold-out problem under sequential bargaining. *American Law and Econ Rev* 14:372–390
- Miceli T, Segerson K (2007) *The economics of eminent domain: private property, public use, and just compensation*, vol 3, *Foundations and trends in microeconomics*. Now Publishers, Boston
- Michelman F (1967) Property, utility, and fairness: comments on the ethical foundations of ‘just compensation’ law. *Harv Law Rev* 80:1165–1258
- Nosal E (2001) The taking of land: market value compensation should be paid. *J Public Econ* 82: 431–443
- Reynolds S (2010) *Before eminent domain*. University of North Carolina Press, Chapel Hill
- Ulen T (1992) The public use of private property: a dual constraint theory of efficient government takings. In: Mercurio N (ed) *Taking property and just compensation: law and economics perspectives of the takings issue*. Kluwer, Boston

Tariff Benefits

- [Preferential Tariffs](#)

Tariff Concessions

- [Preferential Tariffs](#)

Tax Amnesty

Carla Marchese

Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, POLIS Institute, Università degli Studi del Piemonte Orientale “Amedeo Avogadro”, Alessandria, Italy

Definition

Tax amnesty is the opportunity given to taxpayers to write off an existing tax liability (including interests and fines) by paying a defined amount. Such offers are usually presented as being exceptional and available for only a limited period of time. Amnesties can either be general or restricted to certain groups of taxpayers or taxes, and they routinely include the waiving of criminal and civil penalties.

Prevalence and Types of Tax Amnesties

Both local and central authorities grant tax amnesties. Over the past 50 years, the central governments of some developing countries (e.g., Argentina, Colombia, Brazil, India, the Philippines, Turkey) have repeatedly offered amnesties (Le Borge and Baer 2008), as have the central governments of developed countries plagued by specific economic problems such as recession, financial crisis, and large public debt (e.g., Ireland, Italy, Spain, Greece, Portugal). Many developing and developed countries have also occasionally resorted to some form of amnesty to foster flight capital repatriation or to ease economic liberalization and openness to international trade. Local governments, too, often resort to tax amnesties. Many US states have made repeated use of waves of amnesties (Alm and Beck 1993; Mikesell et al. 2012) in response to a variety of motivations, including decreased central support for local tax enforcement in the 1980s (Dubin et al. 1982) or the dwindling of local tax revenue coupled with a mandatory balanced budget in the 2000s.

Lawmakers and local administrators are constantly devising new types of tax amnesties. Innovation and differentiation in this field is likely sparked by the need to capture the attention of the public and, where tax amnesties are frequent, to appeal to groups that have not yet been reached by previous offers.

In terms of the immediate financial benefits to participants, some tax amnesties not only reduce or waive sanctions and interest but also reduce the principal on the tax. These are the so-called extensive tax amnesties (Franzoni 1996; Macho-Stadler et al. 1999), which also often provide immunity from audits for past, and sometimes future, tax liabilities.

The timing of amnesties is a key feature in their functioning: how long the program is available, whether or not extensions will be granted, and the frequency with which amnesties are offered clearly affect their results (Mikesell 1986). In many countries, unaudited taxpayers who spontaneously report tax evasion can be granted a standing permanent tax amnesty (Andreoni 1991), although these amnesties are never of the extensive type and are sometimes available only for a limited time after the violation. Standard tax amnesties instead can be designed to cover recent or past liabilities that still fall within the expiration term of the tax obligation. The benefits of amnesties for the participants sometimes also extend to the future, as various provisions can be introduced to reduce expected future liabilities. For example, in some amnesties, participants who increased their subsequent reported income by a given percentage for several years were exempted from future audits on those years, barring major violations. Portugal granted an amnesty of this type in 1986 (Baer and Le Borgne 2008, p. 10).

Another important aspect of tax amnesties is the information disclosed by participants, which serves to condition their future expected payments. Some amnesties even provide for the free writing-off of past liabilities, so long as the taxpayer's latest tax return was honest (Pommerehne and Zweifel 1991). When a new tax return is filed, the tax administration ordinarily maintains its full powers of auditing. Taxpayers participating in an amnesty may also be subject to special

surveillance in subsequent years. Of course, these policies reduce the amnesty's appeal and its potential as an immediate revenue source. If the government is primarily interested in raising revenue and encouraging participation to boost the immediate amnesty's proceeds, the auditing powers can be limited or excluded, and anonymity can be offered to the amnesty participants. One way this can be achieved is by allowing participants to disclose their liabilities and to make the amnesty payments to a third party (such as a bank), which releases a certificate to be used as a shield in case of future tax audits: the 2001 Italian tax amnesty for capital repatriation with these characteristics was called a "scudo fiscale" (tax shield). However, it is also true that the government's commitment not to access such information may be more or less credible. In Italy, a 2011 law introduced a new tax on capital that benefited from that 2001 amnesty. The new tax was justified on the basis of the benefit principle, with the benefit being continued secrecy to those who entered the amnesty despite newly introduced legislation granting the tax administration easier access to taxpayers' bank accounts.

In terms of the extent of coverage, amnesties are often granted only to taxpayers not yet under investigation. These could be taxpayers who missed a filing deadline (e.g., for VAT), or who failed to file one or more tax returns, or simply those who reported regularly but cheated. However, amnesties can also include those whose liability has already been assessed or liquidated (i.e., the so-called accounts receivable). By granting an amnesty, the tax administration surrenders the right to collect payments from taxpayers through standard means, such as audits, injunctions, or litigation in courts. The ensuing opportunity cost is likely to be larger if the ordinary collection process has already been initiated, as in the case of accounts receivable. Another aspect of coverage is the type of tax or tax base to which the amnesty refers. From this point of view, in principle, all types of payments can be considered, including social contributions, charges and fees, and so on. Amnesties tend to be general to assuage taxpayer fears that further audits will ensue if one specific hidden tax base is disclosed. With

reference to the tax base, it is also important to distinguish standard amnesties from those specifically targeted to flight capital, since the latter involve problems of international relations and tax competition. Amnesties can also involve leverage, so in some cases participants have been required to invest the hidden tax base in special public debt bonds. These securities delivered no yield or a low yield and could not be traded before a given date. As in the case of intermediation by third parties, these special bonds could be used as a shield in case of tax audits. An amnesty of this type was granted in Spain in 1991 (Macho-Stadler et al. 1999). In other cases (e.g., the 1987 amnesty in Argentina) the evaded tax base could also be invested in private enterprises, provided that the investment was twice the evaded tax base amount.

Amnesties are also characterized by the interventions supporting them, which range from global reforms of the tax system to specific provisions aimed at strengthening tax enforcement. The underlying reasoning is some stick should be given along with the carrot represented by the amnesty, in order to avoid negative effects on compliance. These further interventions often include harsher penalties for tax evasion, reorganization of the activities and of the legal capabilities of tax auditors, modifications of laws regulating tax shelters, and use of the funds collected through the amnesty for financing enforcement activities. Moreover, specific information can be sent to the perspective participants to inform them of their likely tax liability, together with the threat of naming and shaming evaders who do not participate or of increasing penalties specifically for them.

A traditional justification often provided for granting a tax amnesty is that special circumstances may motivate unwanted breaches of the law or mistakes by citizens. This is mainly true for amnesties that accompany huge reforms in taxation or other related fields or that are granted after major upheavals such as political regime change, changes in currency, and so on. In these cases, the amnesty is well grounded in terms of equity and should not be harmful in terms of efficiency, since it should be unexpected (and will not induce an ex ante evasion

increase) and unlikely to be repeated (and will not encourage subsequent evasion). The resort to tax amnesties, however, is more frequent and widespread than one would expect on the basis of exceptional circumstances alone. The Philippine government, for example, called its 1980 amnesty the “final amnesty,” although many others followed.

Amnesties are actually a well-established, ancient, and widely used institution. There is evidence of a tax amnesty in the Rosetta stone (from around 200 B.C.): the priests of a Memphis temple thank the Monarch for not demanding a large sum of tax arrears due by the people. Tax amnesties can also be considered a form of pardon that shares some features with other past or recent institutions (Cassone and Marchese 1999) in the field of religion (jubilees, indulgences), criminal justice (plea bargaining), and penal or civil violations (amnesties for illegal immigration, breach of regulatory rules, and so on). The provision of statutes of limitation for some legal obligations and even for crimes can also be considered as a limiting case of a permanent standing amnesty, which might be rationalized on the basis of the fact that as time passes without any actual law enforcement, the net advantages of delayed enforcement tend to vanish or even turn to disadvantages. These considerations suggest that amnesties might exert some positive social function. However, since tax amnesties are also harshly criticized, it is important to understand both the pros and the cons of this institution.

Pros

Among the pros is the idea that amnesties encourage repentance of violators and/or foster their willingness to behave cooperatively in the future (Malik and Schwab 1991). People may be unable to clearly assess ex ante the costs and benefits of violating rules, such as engaging in tax evasion, and while ex post repentance may ensue, the fear of heavy sanctions for past conduct sometimes discourages disclosure of the violation. Amnesties have the positive effect of rendering repentance less costly and the return to honest behavior (such

as the regular payment of future taxes) therefore more likely. The benefits of amnesties in this case are the partial restoration received by society (through payments or other forms of contribution by the violators) and by expected improvement in future compliance.

While repentance implies some form of limited rationality, the full rationality of taxpayers can also be assumed when justifying the participation of citizens and the granting of tax amnesties by governments. Participation can be justified if it is in the best interest of the taxpayer to revise her economic calculus because the costs of evasion have increased, as a result of actual or anticipated changes in law enforcement, for example, which increase the expected sanctions for past misconduct. These revisions, however, are more likely for small evaders that can easily adapt to even marginal changes in the law, while the effects on repeat evaders who have hidden large sums are probably limited.

Good economic news might also provide the motivation for granting an amnesty. If the economy is growing quickly thanks to policies of liberalization and of opening to international trade, it may be the case that firms can benefit more in the new environment if they are legal and have a clean tax record, as this paves the way for accessing the credit market or for listing on the stock exchange and so on. For a firm that has been operating for some time in the hidden economy, however, shifting to legality could imply a huge cost in terms of sanctions for past evasion, and an amnesty can grease the wheels of change. The results of the amnesty in this case should be evaluated not only with reference to the effects on tax revenue but also in terms of the effects on GDP that should ensue thanks to productivity increases in firms that were able to shift to legality (Bose and Jetter 2012).

Amnesties can improve efficiency when they are designed as an optimal discriminatory policy in the field of taxation (Marchese and Cassone 2000). Amnesties are in general discriminatory since they imply a more favorable treatment for those who evaded tax – and for those who can regularize their position at a discount – than for those who complied from the outset on the one

hand or for those who were discovered and punished before the amnesty on the other. Amnesties can perform a role similar to that of price discrimination which can increase a firm's profits. For example, selling a good in different markets at different prices can boost a firm's revenue, and this might occur even if some arbitrage arises; in other words, for example, even if a few of those who were expected to buy at a higher price manage to buy at a lower one. Amnesties can be seen as a way of opening another market beyond that of compliance. If such an offer appeals to a large enough number of new "customers", government revenue can increase even if the number of regular tax compliers decreases somewhat. Price discrimination, however, can be optimally designed only if there are characteristics that distinguish markets or, in our case, that distinguish perspective evaders and compliers.

For a case in which an efficient discrimination can be applied, consider economic shocks that affect particularly some firms or individuals (e.g., sectorial crises, adverse life events, and so forth). Tax evasion can work as an extreme method of increasing disposable income in such circumstances. Those more likely to be harshly hit by a negative shock are also those more likely to resort to such extreme measures. In this case, an amnesty helps the unluckiest to improve their circumstances as well as easing their return to legal behavior. Since there might always be a share of the population experiencing these problems, a standing permanent tax amnesty can be justified both in terms of efficiency (supplying insurance) and of equity (helping those more in need) (Andreoni 1991).

Another possible efficiency-based rationale for resorting to amnesties as a discriminatory policy is related to the exploitation of differences in the visibility of tax evasion (Marchese and Cassone 2000). Citizens who are more confident about their ability to remain undetected are less likely to file their tax returns regularly, but they may still be willing to pay to eliminate the risk of evading the law and thus be interested in an amnesty. While it might be difficult to distinguish *ex ante* those who are more difficult to tax, it should become clear *ex post*, since the more visible should comply immediately, while the less visible

dare to wait. If a favorable amnesty is offered, the latter can enter it. The tax administration bears an opportunity cost since without the amnesty it might have enforced sanctions on the evaders, but if the payment asked for entering the amnesty equals the expected sanctions plus a risk premium, the result could be positive in terms of government revenue. Note that this approach can justify expected, repeated, and periodic tax amnesties, which should work like a form of sales (Cassone and Marchese 1995). Sales, too, aim at discriminating among customers according to their impatience. It is well known that efficient periodic sales can be compatible with a market in equilibrium, in which the share of customers buying in periods in which the full price is applied is stable and the total revenue is larger than if sales are not held. Similarly efficient periodic tax amnesties are compatible with the stability of the number of regular compliers. A characteristic of both efficient amnesties and sales is the greater exploitation of those with the lowest demand (those who buy during sales or those who participate in a tax amnesty), while a better deal should be offered to regular buyers or compliers. This might seem somewhat counterintuitive, but the idea is that those who buy during sales and those who participate in amnesties, while paying less in absolute terms, should be fully expropriated of their willingness to pay, while those characterized by high demand (regular buyers and compliers) should be left with some net gain.

When amnesties are offered to tax evaders already under investigation, they can be rationalized as a form of plea bargaining. In other words, the tax administration uses the amnesty to profitably renounce part of the expected proceeds as long as this also involves a partial cashing of its credits and a substantial reduction in the implementation costs it would have borne. As in the case of plea bargaining, it is expected that those who are more willing to participate in the amnesty are also those who more patently violated the law, since they would lose if they tried to contest their liability. It is widely held that plea bargaining can be considered as an efficient selection tool, as on the one hand the guilty, who risk more, should reveal themselves by pleading guilty and thus be

properly sanctioned, while by not pleading guilty, the innocent should go on to trial, where they are likely to be acquitted (Grossman and Katz 1983). A similar type selection is possible through amnesties. However, some problems may ensue with tax evasion, as long as only monetary sanctions are foreseen. In this case, *ceteris paribus*, the amnesty is more appealing to those who have enough wealth to bear the liability and are thus easier to reach through standard means of enforcement. An amnesty, instead, is not relevant for those who are sanction-proof, as they have nothing to lose if an audit occurs.

While efficient discrimination is a potential, overall justification for tax amnesties in both developed and developing countries, in the latter countries, a further motivation is that they can help compensate for organizational problems related to performing audits and dealing with taxpayers' tax-related appeals. Here, the main idea is that, whenever the productive capacity of the tax administration is modest and fixed in the short run, performance can be improved by using the amnesty to deal with the oldest unsolved cases, in order to concentrate resources on the most recent and probably more visible and thus easier to resolve ones. Moreover, amnesties can also be used by governments to periodically curtail the rents extracted from evaders by corrupted auditors seeking bribes. As long as the amnesty provides a large enough discount, taxpayers will prefer to settle their liabilities directly with the state.

Cons

On the efficiency grounds, the main argument against tax amnesties is that credibility problems may arise concerning the ability and/or commitment of the state to enforcing the tax law (Stella 1991). The idea is that governments in general, when granting an amnesty, try to look tough for the future, in order to induce compliance. Citizens, however, base their beliefs on past experience. Even if amnesties are accompanied by declarations about new stronger enforcement efforts, taxpayers, who take past auditing policy into account, are likely to only slightly correct

their perception of the riskiness of tax evasion. Hence, the collection of resources from past evaders is likely to be low, and the expectation of future amnesties can induce previous compliers to evade. If an amnesty leads to larger evasion, a further subsequent amnesty might appear even more justified, but this might lead to a slippery slope, in which more and more generous and frequent amnesties are granted. Amnesties, in sum, could reveal the weakness behind the feigned tough stance of a government, thus permanently endangering its taxing capacity, i.e., reducing the overall taxpayers' willingness to pay, either through regular compliance or in amnesties.

As long as a government lacks the commitment to enforce the tax law, the aforementioned discriminatory function of amnesties would actually be impossible to implement due to the lack of credibility of the threat of punishment. Discriminatory policies have also been criticized because the tax administration might lack enough information to design them, if taxpayers' attitudes differ along many dimensions (such as income, age, sex, location, and so on). It might even be technically impossible to devise a sufficiently fine-grained discriminatory mechanism.

A further problem raised by amnesties is that they are costly, and their costs are difficult to anticipate and to correctly evaluate (Baer and Le Borgne 2008). As already seen, a large share of these costs is represented by the opportunity costs of renouncing pursuit of the standard enforcement activity over the participants, thus renouncing the revenue that they would otherwise have produced. When the amnesty foresees payment by installment, some installments may be missed, and the initial assessment of the amnesty's proceeds proves incorrect. Other costs can be due to a temporary suspension of enforcement activities, which is often granted during the period in which the amnesty is pending. Moreover, amnesties need to be well designed and advertised: "Get to us before we get to you" is the famous slogan used to advertise the Michigan 2002 tax amnesty. Interventions can include mass mailings of information to prospective participants, the provision of an information hotline, and so forth. Last but not

least, if amnesties endanger the future compliance of honest taxpayers, this prospective revenue loss also needs to be accounted for. Difficulty in evaluating the total costs involved in an amnesty may actually contribute to their popularity, as they would thus be based on a kind of fiscal illusion.

In terms of equity, most critics are of the view that amnesties are unacceptable because they introduce discriminatory treatment of citizens according to law enforcement. More specifically, when an amnesty is granted *ceteris paribus*, the participants can fulfill their tax obligations by paying different amounts than those who complied from the outset or from those who were caught in the meantime. On the other hand, if one focuses not on the equity of rules but on the social outcome instead, one sees that even without amnesties, honest taxpayers and tax evaders routinely end up with differences in their actual contributions to the financing of the public budget. Moreover, as long as the amnesty collects revenue from tax evaders, these differences are reduced, thus securing more horizontal equity. Even vertical equity could increase, if the rich evade more and take advantage of tax amnesties more often. Here too, as is generally acknowledged, focusing on justice in procedures or on justice in outcomes leads to different conclusions.

Even by considering amnesties as discriminatory in principle, the economic approach would suggest considering the potential equity efficiency trade-off resulting from possible compensation for those who are negatively hit thanks to the proceeds that an efficient amnesty should produce. However, the compensation principle is problematic and a source of widespread debate: it would either have to involve the actual payment of compensations (and this would be difficult to plan for and is therefore practically never done) or a comparison on paper of utility gains and losses, which is not acceptable to those who claim there is no objective unit of measurement for making these calculations. At any rate, some types of amnesties, such as those aimed at capital repatriation or at easing the opening of the economy to international trade, have clear-cut and widely recognized general economic benefits extending beyond the effects on tax revenue. They are

therefore often considered more acceptable from an equity point of view.

Another criticism of amnesties is that external incentives might endanger internal incentives to legal behavior. In other words, as long as tax compliance in modern societies largely rests on the internal incentive represented by a moral imperative, the introduction of external incentives based on the gains that an amnesty can produce for both those who participate and those who do not could push the public reasoning away from the moral perspective. This might have negative effects on compliance, since in many cases, pure economic calculus shows that tax evasion pays, so widespread evasion should be expected. Amnesties, however, are generally considered as acceptable on the grounds of equity if they discriminate in favor of more deserving people: amnesties that work as a form of insurance for the most unfortunate or exceptional amnesties granted to citizens faced with the difficulties posed by large-scale overhauls of the tax code or other major changes.

Not everyone who files for an amnesty is a tax evader. Even honest taxpayers may fear having made mistakes on their tax returns and want to avoid being audited or going to court or they may also wish to avoid the inconvenience and cost of being audited. Particularly in developing countries, where enforcement is often intrusive and burdensome for citizens, amnesties can appeal to honest taxpayers. The role of amnesties in this case is ambiguous. On the one hand, they produce a benefit for participants who are honest and deserving from a social point of view; but, on the other, they can imply perverse incentives for the tax administration (Franzoni 2000) to capitalize on its own malfunctioning.

The Observed Effects of Amnesties

As for participation, amnesties tend to deliver somewhat extreme results. Either the entire potential population participates or only a very few members do. This is due to the fact that amnesties ordinarily reduce the perspective workload of auditors, who no longer need to focus on those who entered the amnesty and can thus target

nonparticipants more. Hence, the risks for non-participants (and thus their motivation for taking part) increase with the number of those who have already applied. In a certain sense, there is a network effect, and even in successful amnesties, participation tends to accelerate toward the end of the period of validity, as people wait and see how large participation has been. It is thus suggested that the period in which an amnesty is pending should not be too long (around 90 days), in order to avoid wasting time and to reduce costs. Moreover, this period should not coincide with that in which tax reports are filled-in, since it has been shown that this might exacerbate the trade-off that might arise between reporting for the amnesty and reporting for paying taxes of the current year (Alm and Beck 1990). If both reports have to be made at the same time, when confronted with two ways for reducing their risk in the field of taxation, past evaders can more easily assess the relative advantages. If the amnesty is very cheap, they might even reduce the overall amount they pay, with a negative result for the public budget.

The economic effects of amnesties can be estimated using econometric techniques. The reliability of the results is limited by the fact that one either studies a specific amnesty – but then the data collection should be very detailed, and the results might at any rate not have general significance – or one can pool many amnesties, but then one needs to control for the specificities of each case considered. Moreover, it is not easy to disentangle the specific effects of amnesties from those of the supporting programs often jointly enacted.

Many studies have focused on the large number of amnesties granted in the USA, which present the advantage of providing a large data set of cases sharing a common legal and economic background. Mikesell and Ross 2012 arrive at a total of 117 amnesties granted by 41 States in the period 1980–2011. The results of these econometric studies show that some features of amnesties are decisive for boosting the proceeds, such as the effort made in advertising, the inclusion of accounts receivable, and the possibility for installment payments. Other relevant variables are those

linked to tax evasion opportunities, either in terms of self-employment or with reference to less auditing in the state by the central government. While the amnesties granted in the 1980s also included cases where eligibility was very restricted, the most recent amnesties are more often characterized by large admission criteria. Moreover, recent amnesties tend to waive interest to a larger extent (whereas the principal is never reduced) and are less often accompanied by supportive measures aimed at reinforcing future compliance. This might also be due to the repetition of amnesties over time, a fact that tends to reduce the number of new enforcement interventions not yet enacted that can be introduced and to undermine the credibility of further threats, so that they are mainly dispensed with. From the point of view of gross revenue collected, US states tax amnesties were sometimes successful (with an impact at any rate always below 3% of the yearly tax revenue), while their long-term effect is unclear and probably nil or negative (Alm and Beck 1993; Luitel and Sobel 2007). It has been noted that the participants totally unknown to the tax administration mainly continued to report after they entered an amnesty, but their contribution to the tax revenue has often been scant (Christian et al. 2002), thus raising some questions about the role of amnesties in significantly enlarging the future tax base. The average sums paid in amnesties were often small and related to recent years. Since there is no reason to expect this on an objective basis, the explanation may either be a lack-of-recall of past evasion or a rational calculus about the fact that evasion made far in the past and not yet discovered is even less likely to be found out in the future. In general, there seems to have been some change over time in the goals that the states pursued by granting an amnesty, which mainly ranged from boosting future compliance to providing an immediate source of revenue.

For countries that often granted tax amnesties (including Italy, Ireland, India, the Philippines, and Turkey), the best results seem to have been reached whenever an amnesty was offered together with a policy aimed at strengthening enforcement capacity and at improving the overall efficiency of the tax collection system (Baer and

Le Borgne 2008). In some cases, instead, such as in Argentina, where measures of this type were not introduced, amnesties gave rise to a spiral in which forgiveness for not paying what was due according to a previous tax amnesty was also offered and participation in amnesties faded out over time. When a negative spiral is avoided, amnesties mainly modify somewhat the intertemporal profile of tax revenue; that is, they imply a temporary increase, also due to the fact that they anticipate the cashing in of some future revenue that would have been obtained through ordinary enforcement activity, followed by a subsequent decrease, with no change in the revenue trend.

Experiments have also been used to assess the effects of tax amnesties (Alm et al. 1990). It turns out that amnesties actually do foster subsequent increases in evasion. However, combining amnesties for the past with harsher penalties for the future prompts greater compliance. This combination is reminiscent of a feature that has also been deemed necessary for the efficient and equitable functioning of plea bargaining, where it is found that the judge should use discretionary power to threaten harsher penalties, so that those who plead guilty, while obtaining a discount compared to the new level of penalty, are actually given the sanction due (Grossman and Katz 1983).

Amnesties as a Public Choice Issue

Amnesties imply a kind of temporary modification of the social contract between citizens and the state on taxation. As with contract renegotiations, such modifications can be justified by the arrival of new information not available when the pact was signed or by changes in the preferences and objectives of the parties involved. It might, however, also arise out of intertemporal inconsistency, whenever the parties find it in their interest to renege on their past promises. In the specific case of taxation, the former case would occur with efficient amnesties that ease the adjustment to major changes or perform discriminatory tasks, while the latter would correspond to cases in which governments risk their credibility as law

enforcers in order to secure an immediate increase in revenue. The short-run perspective that often plagues the functioning of democratic governments can easily lead to intertemporal inconsistency. Notwithstanding the long-term benefits of tough tax enforcement, politicians who stay in power for a limited period are likely to be tempted by amnesties that grant immediately available proceeds, while they might not be so worried about the damage that will materialize only in the long run, after they step down. To avoid these drawbacks, in some countries, amnesties can only be introduced with the approval of large parliamentary majorities (as in Italy for amnesties waiving criminal offenses) or they are subjected to approval by referendum (as in Switzerland).

Amnesties can be tempting from the political perspective, too, because they represent means of gaining a quick increase in revenue without burdening the entire tax-paying population with changes in the tax law and rate increases. This might be particularly welcome in preelection periods, as politicians are keen on spending to boost economic growth and consent (the so-called political business cycle). Amnesties, however, might also be seen as a way of squeezing taxpayers, besides offending regular compliers. In fact, from the point of view of political consent, amnesties have not performed well. The US state governors granting one during the electoral year proved more likely not to be reelected (Le Borgne 2006).

Amnesties are relevant for politicians also when they participate in one. If information leaks and reaches the press, it can damage the social reputation and the possible political career of participants. Companies also risk damaging their reputation if it is discovered that they took advantage of an amnesty.

Economic psychology has characterized the implicit psychological contract between the state and the citizens concerning taxation. While citizens are obliged to pay taxes, the state must treat them respectfully yet punish those who fail to comply. If it does not punish them, the citizens who did comply may feel betrayed. In fact, in experiments where there are collective gains from cooperation, the participants often demand punishment for violators. In many cases, it turns

out that those who cooperated are ready to sacrifice a share of their gains in exchange for implementing the punishment. If amnesties are perceived as a breach of the psychological contract of taxation, they will encourage further tax evasion. However, it is also true that honest taxpayers might consider participants in an amnesty as willing to change their behavior. Frustration that some evaders may go unpunished in amnesties can also be dealt with if the hidden evaders are threatened with harsher penalties. In fact, punishment can serve two main purposes: retribution for illicit conduct and restoration of legal order. While the retribution recouped via amnesties is lower than that provided for by standard rules, amnesties can convey some advantages in terms of restoration as long as they foster greater future compliance. From this point of view, amnesties should be favored by those who are more generally in favor of alternative penalties aimed at facilitating the social rehabilitation of those who breached the law (Rechberger et al. 2010). The ambiguous role that amnesties can play implies that public debate over an amnesty program may have important consequences. This is confirmed by experiments in Switzerland and in Costa Rica conducted by Torgler and Schaltegger in 2005. They found that only amnesties approved by referendum lead to increased compliance. This effect can be traced back to the formation of public opinion through the public discussions among participants in the experiment that accompanied the referendum. Participants perceived the amnesty not as an imposition from above but as an agreed-upon intervention with useful functions, and this in turn increased the social pressure for cooperation.

As for public opinion on amnesties, the Bank of Italy (Cannari and D'Alessio 2007) conducted interviews in 1992 and 2004 in which questions were asked about the government motivations for granting an amnesty, the results expected, and the respondent's evaluation of such a policy. Regarding the first question, the majority of respondents think that the Italian state resorts to amnesties either because it is powerless to punish evaders or because groups of evaders had lobbied for preferential treatment. Yet in reference to the evaluation of the consequences and the moral

judgment of this policy, only about 30% of respondents have clear-cut negative feelings (that evasion will increase and the policy offends honest citizens), while the remaining respondents express more nuanced opinions. The negative feelings, however, were more frequent in 2004 than in 1992, possibly due to some deterioration of the government's credibility in light of repeated tax amnesties.

Besides being relevant for internal political affairs, amnesties can also be linked to the state of cooperation or competition between governments or levels of government, since they represent a means of dealing with externalities in taxation and in enforcement policies. It is also the case that forms of imitation or competition often arise among neighboring countries, so amnesties sometimes spread from one country to another. The importance of externalities in this field is confirmed by an empirical analysis of the motivations leading states in the USA to grant an amnesty, which revealed that the likelihood of amnesties increased as the effort of the federal government in auditing taxpayers within the state decreased (Dubin et al. 1982; Le Borgne 2006). In the field of international relations, in 2010, the OECD suggested offshore voluntary disclosure programs as a solution to help governments benefit quickly in terms of revenue from the effects of improvements in international cooperation for information exchange and transparency that have occurred since the onset of the financial crisis. Voluntary disclosure implies a "limited-time offer by the government to a specified group of taxpayers to settle undisclosed or unpaid tax liabilities for a previous period in return for defined concessions over civil or criminal penalties. In some cases, there are also concessions over the amount of tax and/or interest payable" (OECD 2010, p. 11): the definition is very close to that of an amnesty.

International organizations such as the IMF have studied more in general the policy of tax amnesties (Baer and Le Borgne 2008). They arrive at a substantially negative evaluation of this institution. Their suggestion in this field is to avoid the resort to amnesties, while pursuing alternative policies instead, such as: (i) trying to reduce tax evasion by addressing its basic determinants (unsustainable

tax system, insufficient and improper enforcement, the malfunctioning of courts, etc.); (ii) resorting to permanent programs for encouraging disclosure of tax evasion and for granting extended payment agreements to taxpayers under economic stress for personal or conjunctural reasons; and (iii) improving the functioning of the tax administration, for example, by granting it with the power of disposing of cases unlikely to lead to net contributions to the revenue. The basic idea is to consolidate taxpayers' expectations about the commitment of the state to fighting tax evasion, while also dealing with the problems that often motivate the granting of an amnesty in other ways. These policy suggestions have sound economic foundations. They can be likened to the commercial practice that uses price discrimination systematically rather than intermittently, so that, for example, special sales can be replaced by permanent offers at outlets specializing in major discounts. Following these suggestions, however, is not easy, particularly when one considers developing countries and countries where corruption is frequent. Whenever interventions such as a standing amnesty are introduced, it is possible that corrupt auditors will accept bribes in order to say that a taxpayer voluntarily disclosed her evasion. Problems of this type have arisen in the past even in the USA (Andreoni 1991). Likewise, whenever a personalized deal (such as an individual installment plan for tax payments) must be designed, the risk of corruption of officials tends to be greater than when general public interventions such as tax amnesties are implemented, given that they are often regulated by the law. These considerations, coupled with the existence of genuine unanticipated phenomena that cannot be dealt with efficiently through other means or with discrimination opportunities not available elsewhere, suggest that amnesties are and are likely to remain an accepted tool in tax administration.

References

- Alm J, Beck W (1991) Wiping the slate clean: individual response to state tax amnesties. *South Econ J* 57: 1043–1053
- Alm J, Beck W (1993) Tax amnesties and compliance in the long run: a time series analysis. *Natl Tax J* 46:53–60

- Andreoni J (1991) The desirability of a permanent tax amnesty. *J Public Econ* 45:143–159
- Baer K, Le Borgne EL (2008) Tax amnesties. IMF, Washington, DC
- Bose P, Jetter M (2012) Liberalization and tax amnesty in a developing economy. *Econ Model* 29:761–765
- Cannari L, D'Alessio G (2007) Le opinioni degli Italiani sull'evasione fiscale, vol 618, Temi di discussione. Bank of Italy, Roma
- Cassone A, Marchese C (1995) Tax amnesties as special sales offers: the Italian experience. *Public Finance/Finance Publiques* 50:51–66
- Cassone A, Marchese C (1999) The economics of religious indulgences. *J Inst Theor Econ* 155:429–442
- Christian CW, Gupta S, Young JC (2002) Evidence on subsequent filing from the state of Michigan's income tax amnesty. *Natl Tax J* 55:703–721
- Dubin JA, Graetz MJ, Wilde LL (1992) State income tax amnesties: causes. *Q J Econ* 107(3):1057–1070
- Franzoni LA (1996) Punishment and grace: on the economics of tax amnesties. *Public Finance* 51:353–368
- Franzoni LA (2000) Amnesties, settlements and optimal tax enforcement. *Economica* 67:153–176
- Grossman GM, Katz ML (1983) Plea bargaining and social welfare. *Am Econ Rev* 73:749–757
- Le Borgne E (2006) Economic and political determinants of tax amnesties in the U.S. States. IMF WP 706/222
- Luitel HS, Sobel RS (2007) The revenue impact of repeated tax amnesties. *Public Budg Finance* 27:19–38
- Macho-Stadler I, Olivella P, Pérez-Castrillo D (1999) Tax amnesties in a dynamic model of tax evasion. *J Public Econ Theory* 1:439–463
- Malik A, Schwab RM (1991) The economics of tax amnesties. *J Public Econ* 46:29–49
- Marchese C, Cassone A (2000) Tax amnesty as price-discriminating behavior by a monopolistic government. *Eur J Law Econ* 9:21–32
- Mikesell JL (1986) Amnesties for state tax evaders: the nature of and response to recent programs. *Natl Tax J* 39:507–525
- Mikesell JL, Ross JM (2012) Fast money? The contribution of state tax amnesties to public revenue systems. *Natl Tax J* 65:529–562
- OECD (2010) Offshore voluntary disclosure, comparative analysis, guidance and policy advice. OECD, Paris
- Pommerehne WW, Zweifel P (1991) Success of a tax amnesty: at the polls, for the fisc? *Public Choice* 72:131–165
- Rechberger S, Hartner M, Kirchler E, Hämmerle FK (2010) Tax amnesties, justice perceptions, and filing behavior: a simulation study. *Law Policy* 32:214–225
- Stella P (1991) An economic analysis of tax amnesties. *J Public Econ* 46:383–400
- Torgler B, Schaltegger CA (2005) Tax amnesties and political participation. *Public Finance Rev* 33:403–431

Tax Evasion by Firms

Laszlo Goerke

IAAEU (Institute for Labour Law and Industrial Relations in the European Union),

University Trier, Trier, Germany

IZA, Bonn, Germany

CESifo, Munich, Germany

Abstract

A standard finding in the analysis of tax evasion and avoidance by firms is that the decision about the firm's activity level can be separated from the evasion choice and vice versa, irrespective of the tax under consideration. The implications, relevant empirical evidence, and the robustness of this separability feature are surveyed. The article finishes with speculations about topics of future research with regard to tax evasion or avoidance by businesses.

JEL-Classification: H 25, H 26, K 34

Definition

Tax evasion (avoidance) by a firm represents the attempt to illegally (legally) reduce the payment of taxes which have to be remitted by a profit-maximizing entity to below the level prescribed by law.

Introduction

A substantial fraction of tax revenues in OECD countries is remitted by firms (OECD 2013). This is the case either because firms are legally obliged to pay, for example, corporate income, payroll, property, or consumption taxes, or because firms act as withholding agents, inter alia with respect to personal income taxes and social security contributions. Therefore, firms and businesses have ample scope for tax avoidance and tax evasion

activities. Slemrod (2007, p. 28), for example, reports that the average tax gap in the United States – that is, the difference between the amount of taxes due and the amount paid voluntarily and in time – was about 16% in 2001. Moreover, the tax gap for income taxes of small corporations and for business income was substantially higher. As a further piece of evidence, the VAT compliance gap is estimated to exceed 10% in a number of European Union member states (Keen 2013). Consequently, tax evasion and avoidance by firms is not only feasible but also seems to be widespread.

Despite this evidence, the vast majority of contributions on tax evasion has focused on and analyzed the decision by individuals to evade income taxes. (The relevant literature is surveyed, for example, by Andreoni et al. (1998), Alm (1999), Slemrod and Yitzhaki (2002), Marchese (2004), Franzoni (2009), and Sandmo (2012). The contributions by Cowell (2004) and Slemrod (2007) also include substantial sections on firms. See also ► “Tax Evasion by Individuals”, this encyclopedia.) The extant literature on tax evasion by firms has then focused on results which are specific to these entities. Such features which differentiate a firm from an individual are, inter alia, (1) all payoffs can be defined in terms of money; (2) firms often have market power; (3) risk neutrality may be an appropriate approximation of preferences; and (4) within a firm, conflicts of interest between owners and managers are likely to arise.

Basic Model

Subsequently, we take up some of these features and analyze a firm’s tax evasion and avoidance behavior. We consider a single firm which maximizes expected profits. In order to do so, the firm determines the level, x , of economic activity and the extent of tax evasion or avoidance. The overall rate of all taxes the firm has to remit is labeled t . If the firm does not evade or avoid taxes, its (legitimate) profits are given by $\pi(x, t)$. We assume that a profit-maximizing level of activity, x^* , exists. This activity choice, x^* , is defined by

$\pi_x(x^*, t) = 0$ and $\pi_{xx}(x^*, t) < 0$, where subscripts denote partial derivatives. Taxes remitted by firms will generally reduce profits, implying that $\pi_t < 0$ holds. This will be the case unless firms (1) can fully shift taxes forward to customers or backward to employees or suppliers of input goods or (2) act as withholding agents. In such a situation, $\pi_t = 0$ will apply. Furthermore, a pure profit tax will generally not affect a firm’s optimal activity choice ($\pi_{xt} = 0$). In general, however, higher taxes will reduce the incentives to exert an economic activity. This implies that the gain from higher activity shrinks with the tax rate ($\pi_{xt} < 0$).

If the firm avoids or evades taxes, it decides about the under-declaration of taxes, which is labeled E , such that the resulting monetary gain amounts to Et . We subsequently assume that taxes are under-declared ($E > 0$) and that E is less than the tax base, in contrast, for example to Virmani (1989) and Cremer and Gahvari (1992), who analyze models of tax evasion by firms which allow for corner solutions, i.e., outcomes in which either no tax evasion takes place or no taxes are paid. To ensure an interior solution, evasion or avoidance has to be costly. These costs are given by $C(E, t, F)$, where C is increasing in the under-declaration, E , for $E > 0$ at an increasing rate ($C_E(0) = 0 < C_E(E > 0)$, $C_{EE} > 0$) and also rising in the tax rate, t ($C_t > 0$), and a parameter F , which can capture the penalty which a tax-evading firm has to pay if evasion is detected ($C_F > 0$).

In the case of tax avoidance, profits may be considered as certain with regard to the outcome of tax payments. If tax evasion takes place, there is a probability, p , that evasion is successful and profits amount to $\pi(x, t) + Et$ and a converse probability $1 - p$ that the firm is audited and evasion is detected, so that profits are given by $\pi(x, t) + Et - C(E, t, F)$. If the firm evades taxes, expected nonlegal profits hence equal

$$\pi^N(x, E) = \pi(x, t) + Et - (1 - p)C(E, t, F). \quad (1)$$

The maximization of nonlegal profits π^N implies two first-order conditions:

$$\pi_x = 0 \quad (2a)$$

and

$$t - (1 - p)C_E = 0. \quad (2b)$$

The optimal activity choice, x^* , maximizes legal profits, π , while the optimal under-declaration, E^* , balances the marginal gain in terms of lower tax payments with the greater expected costs of evasion or avoidance.

Major Findings

From conditions (2a) and (2b), a number of fundamental insights can be derived:

1. Analytically, there is no difference between tax evasion ($0 < p < 1$) and tax avoidance ($p = 0$). This is the case because a penalty or, alternatively, the costs of avoidance affect the firm's payoff, namely, expected profits, qualitatively in the same way. Consequently, from now on, we will only refer to tax evasion, although the exposition obviously applies to avoidance activities as well.
 2. The firm's activity decision is separable from its evasion choice. That is, the model predicts, first, that the firm will always choose that level of economic activity, x^* , which it would have selected in the absence of evasion activities. Second, the optimal extent of tax evasion, E^* , is the same, irrespective of activity choices. Accordingly, large firms, which are characterized by a high activity level, under-declare the same amount as otherwise identical small firms which exhibit a lower activity level. Note that this separability feature will also arise if the audit probability, $1 - p$, depends on the under-declaration, such that $p = p(E)$, because the firm's optimal activity choice is still determined by (2a).
- The separability prediction is due to two features. First, the monetary gains from legal profits, π , and from tax evasion, $E t - (1 - p)C$, affect the firm's payoff in qualitatively the same way and do not reinforce or weaken each other. This is in contrast to usual findings
- with regard to tax evasion by individuals because an increase in activity, say in working time, reduces leisure which, in turn, alters the marginal utility from income (see ► [“Tax Evasion by Individuals”](#)). Second, the net gain from tax evasion, $E t - (1 - p)C$, is independent of the activity level, x .
- The empirical evidence on the lack of a correlation between firm size and tax evasion activities is mixed (Rice 1992; Nur-tegin 2008; Hanlon et al. 2007; Cai and Liu 2009; Tedds 2010; Hoopes et al. 2012). Since firm size in these studies is measured by the number of employees, value added, or assets, these measures are imperfect proxies for economic activity. Moreover, the costs of evasion may vary with firm size. Accordingly, the empirical findings do not necessarily provide comprehensive evidence with regard to the separability prediction.
3. A higher tax rate, t , will raise evasion activities as long as a rise in t does not increase the marginal costs of evasion, C_E , by more than $1/p$. This will, for example, always be the case if the costs of evasion depend neither on the tax rate ($C_t = 0$) nor vary with the amount of taxes evaded, $E t$. Accordingly, if adhering to the law becomes more expensive, violations of these regulations will become more severe. This is a straightforward and intuitive prediction which also generally obtains in standard models of the economic analysis of crime but does not always apply in the case of tax evasion by individuals (see Yitzhaki 1974 and the literature cited above in section [“Introduction”](#)). The theoretical prediction that higher tax rates induce firms to expand evasion activities is generally consistent with empirical findings (Rice 1992; Nur-tegin 2008; Cai and Liu 2009).
 4. If a higher fine, F , raises the marginal costs of tax evasion, so that $C_{EF} > 0$ holds, a rise in F will reduce tax evasion. A greater probability, $1 - p$, of being audited will have the same effect. The empirical evidence, although scarce, is consistent with these predictions (Nur-tegin 2008; Hoopes et al. 2012). Moreover, the separability feature implies that changes in tax enforcement have no impact on a firm's activity level.

5. The results outlined above are not affected by the type of tax considered (Yaniv 1995). Accordingly, the findings are independent of the tax base (such as corporate income, the payroll, value of property, sales, value added), the curvature of the tax schedule, and whether firms act as withholding agents or not. In addition, the theoretical analysis implies that empirical findings for one tax are applicable to others as well.

Robustness

Many analytical contributions of tax evasion by firms have investigated the robustness of the separability result and of its implications and have identified conditions under which the result will no longer hold. In terms of Eq. 1, this can be the case if nonlegal profits, π^N , are affected by a variation in activity, x , not only via official net profits $\pi(x, t)$, i.e., if $E t - (1 - p)C$ varies with activity, x . Such dependence can result under a variety of circumstances, and we discuss four of them below:

1. Assume that $p = p(x)$ holds. (Cf. Virmani 1989). This could be the case, for example, because tax authorities condition the audit probability, $1 - p(x)$, on firm size or because it is related to the frequency of transactions between buyers and sellers which, in turn, are positively correlated to activity. Instead of directly assuming a relationship between p and x , Lee (1998), Marrelli (1984), Marrelli and Martina (1988), and Wang (1990) indirectly include the activity level into the respective specification of the audit probability. Further, Bayer and Cowell (2009) derive such linkage on the basis of a relative auditing rule. According to this rule, the audit probability is positively related to tax payments by other comparable firms and their output levels in an oligopolistic market.

Given the modification of $p = p(x)$, the first-order conditions for a profit maximum are given by $\pi_x + p'C = 0$ and by (2b). If the audit probability, $1 - p(x)$, declines with

activity, because larger firms can avoid detection more easily, tax evasion will induce a firm to expand activity. Fighting tax evasion, for example, by raising the fine, F , and the marginal costs of evasion, C_E , will still reduce evasion activities for a given activity level. Therefore, the gain from an activity expansion will become less pronounced and the overall activity change in response to a higher fine will generally be ambiguous. Even more surprising, a rise in activity, x , enhances the gain, $(1 - p(x))C_E$, from evasion because of the decline in the audit probability. Consequently, the change in evasion activities resulting from a higher fine is also potentially ambiguous (Virmani 1989). Note finally that if knowing the activity level would enable tax authorities to infer the tax base, the assumption of $p = p(x)$ could be inconsistent with the notion of the firm being able to evade taxes. However, knowledge of activity, x , does not necessarily imply that the true tax base is known, unless they are perfectly correlated as, for example, in the case of a unit tax on output.

2. Assume, alternatively, that the firm is not free to choose the under-declaration, E , but can only select a fraction or multiple, e , of activity, x , so that $E = ex$. Such a situation may arise if activity is easily observable, for example, in the case of output levels. The first-order conditions for a profit maximum are given by $\pi_x - e(t - (1 - p)C_E) = 0$ and $t - (1 - p)C_E = 0$. Therefore, activity and evasion decisions are separable if firms can select e optimally. However, if institutional restrictions limit a firm's choice of e , $t - (1 - p)C_E \neq 0$ will hold, and the separability feature will no longer apply (Marrelli 1984; Wang and Conant 1988; Yaniv 1995). Since the restriction on evasion will only be relevant if it is binding, the above result suggests that firms which are easier to monitor and cannot choose e optimally will evade less tax, while evasion and activity choices will be related. The empirical evidence that state-owned companies (Nur-tegin 2008) and publicly traded firms (Rice 1992; Hanlon et al. 2007; Tedds 2010) evade less is consistent with this interpretation.

3. A relationship between economic activity and tax evasion can also arise due to market equilibrium effects. To illustrate, suppose that aggregate activity is positively related to profits because higher profits will induce more firms to enter the market. A higher fine, F , not only raises the costs of evasion ($C_{EF} > 0$) and reduces profits but also makes market entry less attractive. In consequence, the market structure will change along with any policy to reduce tax evasion, and such a policy may actually mitigate competition. Conversely, policies to enhance competition, such as the abolition of product market regulations, may have the detrimental effect of intensifying tax evasion activities by firms (cf. Cai and Liu 2009; Goerke and Runkel 2011).
4. Finally, consider the existence of principal-agent problems. They are likely to arise, for example, in firms run by managers. Managers are unlikely to maximize profits but will pursue their own objectives, at least to some extent. In the presence of such principal-agent problems, the objective function of the firm's decision-maker could be given by $\pi^N(x, E) + K$, where K represents the manager's additional objective. K may be increasing in x and could depend on the under-declaration, E . Moreover, K will vary with the fine and audit probability if the manager is legally or otherwise responsible for tax evasion activities by the firm. Hence, $K = K(x, t, E, F, p)$. Since the manager's first-order condition for a maximum with regard to output is given by $\pi_x + K_x = 0$, it is obvious that activity, x , may vary with tax enforcement and that separability may no longer hold if the decision-maker's marginal payoff is altered by evasion (i.e., if $K_{xE} \neq 0$). As a consequence, changes in the structure of corporate governance; the relative importance of the nonprofit objective, K ; restrictions on the level and composition of a manager's remuneration; and the manager's preferences can affect the choice of economic activity, x , and create a relationship, as well as influence the possible linkage between activity and tax evasion activities (see, e.g., Joulfaian 2000; Crocker and Slemrod 2005; Goerke 2007).

Future Directions

In lieu of a summary, we may speculate about future areas of research. For example, the interaction of evasion and avoidance behavior of individuals on the one hand and of firms and businesses on the other can play a greater role in the future. Moreover, the relationship between the structure of input and output markets and tax evasion by firms which are active on these markets is likely to become a more prominent topic. Furthermore, the ability of firms to shift activities across jurisdictions can affect tax evasion and avoidance behavior. In addition, firms are likely to have better outside options than individuals in negotiations with tax authorities. Such bargains between tax payers and authorities also appear to be a promising area of future work. Additionally, the consequences of the division of labor have been an issue which has hardly been looked at. The tax declaration may, for example, be decided upon by a different agent within the firm or outside its realm (think of tax preparers, etc.) than the one who determines the activity level. Such a separation of responsibilities will create additional principal-agent problems. When looking at individuals, recent years have seen intensified attempts to apply various facets of behavioral economics to the analysis of tax evasion (Hashimzade et al. 2013). This may also be an aspect which becomes relevant to the investigation of firm behavior (Alm and McClellan 2012). Finally, it is noteworthy that the economic analysis of crime focuses very much on normative issues. The question of how much tax evasion is optimal has not been an issue looked at intensively with reference to taxes remitted by firms. Hence, it can be conjectured that welfare issues will also play a more prominent role in the future.

References

- Alm J (1999) Tax compliance and tax administration. In: Hildreth WB, Richardson JA (eds) Handbook on taxation. Marcel Dekker, New York, pp 741–768
- Alm J, McClellan C (2012) Tax morale and tax compliance from the firm's perspective. *Kyklos* 65(1):1–17

- Andreoni J, Erard B, Feinstein J (1998) Tax compliance. *J Econ Lit* 36(2):818–860
- Bayer R, Cowell FA (2009) Tax compliance and firms' strategic interdependence. *J Public Econ* 93(11–12):1131–1143
- Cai H, Liu Q (2009) Competition and corporate tax avoidance: evidence from Chinese industrial firms. *Econ J* 119(537):764–795
- Cowell FA (2004) Carrots and sticks in enforcement. In: Aaron H, Slemrod J (eds) *The crisis in tax administration*. Brookings Institution Press, Washington, DC, pp 230–275
- Cremer H, Gahvari F (1992) Tax evasion and the structure of indirect taxes and audit probabilities. *Public Financ/Financ Publiques* 47(Suppl):351–365
- Crocker KJ, Slemrod J (2005) Corporate tax evasion with agency costs. *J Public Econ* 89(9–10):1593–1610
- Franzoni LA (2009) Tax evasion and avoidance. In: Garoupa N (ed) *Encyclopedia of law and economics. Criminal law and economics*, vol 3. Edward Elgar, Cheltenham/Northampton, pp 290–319
- Goerke L (2007) Corporate and personal income tax declarations. *Int Tax Public Financ* 14(3):281–292
- Goerke L, Runkel M (2011) Tax evasion and competition. *Scott J Polit Econ* 58(5):711–736
- Hanlon M, Mills L, Slemrod J (2007) An empirical examination of corporate tax noncompliance. In: Auerbach AJ, Hines JR, Slemrod J (eds) *Taxing corporate income in the 21st century*. Cambridge University Press, Cambridge, UK, pp 171–210
- Hashimzade N, Myles GD, Tran-Nam B (2013) Applications of behavioural economics to tax evasion. *J Econ Surv* 27(5):941–977
- Hoopes JL, Mescall D, Pittmann JA (2012) Do IRS audits deter corporate tax avoidance. *Account Rev* 87(5):1603–1639
- Joulfaian D (2000) Corporate income tax evasion and managerial preferences. *Rev Econ Stat* 82(4):698–701
- Keen M (2013) The anatomy of the VAT. *Natl Tax J* 66(2):423–446
- Lee K (1998) Tax evasion, monopoly, and nonneutral profit taxes. *Natl Tax J* 51(2):333–338
- Marchese C (2004) Taxation, black markets, and other unintended consequences, Chapter 10. In: Backhaus JG, Wagner RE (eds) *Handbook of public finance*, vol 1. Kluwer Academic, Boston, pp 237–275
- Marrelli M (1984) On indirect tax evasion. *J Public Econ* 25(1–2):181–196
- Marrelli M, Martina R (1988) Tax evasion and strategic behaviour of the firms. *J Public Econ* 37(1):55–69
- Nur-tegin KD (2008) Determinants of business tax compliance. *B E J Econ Anal Policy* 8(1), (Topics), Article 18
- OECD (2013) *Revenue statistics 2013*. OECD Publishing
- Rice EM (1992) The corporate tax gap: evidence on tax compliance by small corporations. In: Slemrod J (ed) *Why people pay taxes – tax compliance and enforcement*. The University of Michigan Press, Ann Arbor, pp 125–161
- Sandmo A (2012) An evasive topic: theorizing about the hidden economy. *Int Tax Public Financ* 19(1):5–24
- Slemrod J (2007) Cheating ourselves: the economics of tax evasion. *J Econ Perspect* 21(1):25–48
- Slemrod J, Yitzhaki S (2002) Tax avoidance, evasion, and administration, Chapter 22. In: Auerbach AJ, Feldstein M (eds) *Handbook of public economics*, vol 3. Elsevier, Amsterdam, pp 1423–1470
- Tedds LM (2010) Keeping it off the books: an empirical investigation into the characteristics of firms that engage in tax non-compliance. *Appl Econ* 42(19):2459–2473
- Virmani A (1989) Indirect tax evasion and production efficiency. *J Public Econ* 39(2):223–237
- Wang LFS (1990) Tax evasion and monopoly output decisions with endogenous probability of detection. *Public Financ Q* 18(4):480–487
- Wang LFS, Conant JL (1988) Corporate tax evasion and output decisions of the uncertain monopolist. *Natl Tax J* 41(4):579–581
- Yaniv G (1995) A note on the tax evading firm. *Natl Tax J* 48(1):113–120
- Yitzhaki S (1974) A note on income tax evasion: a theoretical analysis. *J Public Econ* 3(2):201–202

Tax Evasion by Individuals

Laszlo Goerke

IAAEU (Institute of Labour Law and Industrial Relations in the European Union),
University Trier, Trier, Germany
IZA, Bonn, Germany
CESifo, Munich, Germany

Abstract

The basic deterrence model of tax evasion is described, its main predictions are derived, and limitations and flexibility are outlined. Further, the model is interpreted in light of some key institutional features characterizing tax enforcement in OECD countries. Throughout the survey, findings originating from the deterrence model are contrasted with predictions which result from a simple model of criminal activity and law enforcement.

Definition

Tax evasion by individuals represents the attempt to illegally reduce the payment of taxes which

have to be remitted by an individual tax payer to below the level prescribed by law.

Introduction

The analysis of income tax evasion by economists has covered many issues, and most extant surveys focus on a selection of relevant aspects (see, e.g., the contributions by Andreoni et al. (1998), Alm (1999, 2012), Cowell (2004), Slemrod and Yitzhaki (2002), Marchese (2004), Slemrod (2007), Franzoni (2009), and Sandmo (2012)). In many of these reviews, the investigation of tax evasion is interpreted as a special case of the approach which is employed in the economic analysis of crime. In this survey, we explicitly adopt such a perspective and relate findings originating from the analysis of income tax evasion to the broader economics literature on crime. In doing so, we first take a theoretical perspective, present the basic deterrence model of tax evasion, derive its main predictions, and indicate its restrictions as well as the analytical flexibility. Second, we adopt a more institutional viewpoint and confront the theoretical predictions with basic features of real-world enforcement systems. Finally, we compare selected aspects which are discussed in both the literature on tax evasion and the public enforcement of law (as reference for the literature on the public enforcement of law, we use the article in the *Handbook of Law and Economics* (Polinsky and Shavell 2007)).

Basic Theory

We consider a representative, risk-averse individual who is endowed with an exogenously given income Y . This income represents the tax basis and is subject to a linear tax at the rate t . The individual can decide on the amount of income X he/she does not report to tax authorities. Therefore, the gain from evading taxes will equal Xt if tax evasion remains undetected. This takes place with an exogenously given probability p , $0 < p < 1$, and the individual's income then amounts to $Y^s = Y(1 - t) + Xt$. With the opposite

probability, $1 - p$, the individual or taxpayer will be audited, and tax evasion will be detected. In this case, a fine F is imposed, and the resulting income equals $Y^c = Y^s - F$. While the fine is assumed to be a function of undeclared income X in the seminal contribution by Allingham and Sandmo (1972), Yitzhaki (1974) argues that the penalty is usually based on the amount of taxes evaded, Xt . Consequently, we define the fine F as a linear combination of both determinants, $F := fX[\alpha t + (1 - \alpha)]$, where $f, f > 0$, is labeled marginal fine. The parameter α , $0 \leq \alpha \leq 1$, depicts the relative importance of the amount of taxes evaded. For $\alpha = 1$ (0), this implies that the fine is solely a function of taxes evaded (undeclared income). The specification of F reflects the fact that the penalty rates in many OECD countries vary with amount of undeclared taxes but include fixed components or change with other determinants than the underdeclaration (OECD 2009, p. 136 ff). As the final building block, we assume that utility u is increasing in disposable income at a decreasing rate, $u' > 0 > u''$, and that the individual can be described by von Neumann-Morgenstern preferences. Accordingly, expected utility $U(X)$ is given by

$$U(X) = pu \underbrace{(Y(1 - t) + Xt)}_{:=Y^s} + (1 - p)u \underbrace{(Y(1 - t) + Xt - fX(\alpha t + 1 - \alpha))}_{:=Y^c} \quad (1)$$

Maximizing U with respect to the underdeclaration, X , yields as first-order condition

$$U'(X) = pu'(Y^s)t + (1 - p)u'(Y^c)(t - f(\alpha t + 1 - \alpha)) = 0 \quad (2)$$

The first term in Eq. 2 describes the utility gain from underdeclaring an extra unit of income if tax evasion is successful, while the second term depicts the loss because income declines when being punished. The underdeclaration which results when these two effects are balanced out

is indicated by X^* . Note that there will only be an underdeclaration if the gain from evading the first euro of taxes is positive, that is, if $U'(X)$ is greater than zero for $X = 0$ and, hence, for $Y^s = Y^c$. This implies that there is an upper level for the marginal fine $f_{\max} = t/[(1 - p)(\alpha t + 1 - \alpha)]$. Furthermore, tax evasion will only be costly if disposable income Y^c shrinks with the underdeclaration in the case of detection. Accordingly, there is a minimal marginal fine $f_{\min} = (1 - p) f_{\max} = t/(\alpha t + 1 - \alpha) > t$. The setting described above focuses on tax evasion. While evasion is the illegal attempt to reduce tax payments, tax avoidance is often interpreted as its legal counterpart. By setting the detection probability, $1 - p$, equal to unity and adding a cost function which increases in the amount of taxes avoided at an increasing rate, the above framework can be amended in order to analyze tax avoidance. Furthermore, many findings derived with regard to tax evasion also hold in an avoidance setting.

Central Results

How does the optimal underdeclaration, $X^* > 0$, vary with income, Y , the parameters of the tax enforcement system, p and f , and the tax rate, t ? The respective effects often depend on whether the fine, F , varies with the tax rate, t , i.e., on the value of the parameter α and on the relationship between income and the Arrow-Pratt measure of absolute risk aversion, $R_a(Y) := -u''(Y)/u'(Y)$.

Income, Y , exerts a positive impact on the optimal underdeclaration, X^* , if the individual exhibits decreasing absolute risk aversion, R_a , that is, if the willingness to engage in risky activities rises with income. To provide an intuition, note that a higher exogenous income, Y , raises disposable income for a given underdeclaration, irrespective of whether tax evasion is detected or not. If this general increase in income makes the individual more willing to take risks, the gain from higher income shrinks by less than the costs in terms of utility. Therefore, the optimal underdeclaration, X^* , rises. Since the tax basis, Y , becomes larger, the amount of taxes paid, that is, $(Y - X^*)t$, may nevertheless increase. If, however, absolute risk aversion, R_a ,

does not vary with income, there is no income effect, and the underdeclaration remains constant, while the amount of taxes paid surely increases.

A higher marginal fine, f , and a greater detection probability, $(1 - p)$, both reduce the optimal underdeclaration, X^* . If the marginal fine, f , rises, two consequences strengthen the incentives to pay taxes. First, there is an income effect since a higher fine payment decreases disposable income, Y^c , if evasion is detected. Therefore, the marginal utility of income in this state of the world rises, and the utility loss resulting from the fall in income if penalized becomes larger. Second, the penalty on the last euro of underdeclared income rises. A higher probability of detection, $1 - p$, makes it more likely that an income loss occurs. Consequently, the individual responds by reducing the loss in disposable income if this more likely event takes place.

The consequences of a higher tax rate, t , hinge on the specification of the fine and on absolute risk aversion, R_a . Suppose, initially, that the fine, F , depends on the amount of taxes evaded ($\alpha = 1$). The optimal underdeclaration, X^* , will decline with the tax rate, t , if the individual exhibits constant or decreasing absolute risk aversion, R_a , while the impact is theoretically ambiguous otherwise. For $\alpha = 1$, F is a multiple of the tax rate, t . Accordingly, a rise in t alters the gain and costs from evasion proportionately. Therefore, the impact of the tax rate, t , on the optimal underdeclaration, X^* , is solely determined by the income effect. A higher tax rate, t , reduces disposable income and does so more if evasion is detected than if it remains unobserved. If a decline in income, in turn, raises absolute risk aversion, the optimal underdeclaration, X^* , will shrink. If, alternatively, the fine, F , depends on the underdeclaration ($\alpha = 0$), X^* will rise with the tax rate, t , if absolute risk aversion, R_a , is constant or increasing with income, while the relationship will once again be ambiguous otherwise. In this case, a higher tax rate, t , reduces the penalty relative to the gain from evasion, namely, the lower tax payment. A relative decline in the penalty induces the individual to raise the underdeclaration, X^* , *ceteris paribus*. This substitution effect will be mitigated or reversed by the

income effect which provides greater incentives to underdeclare if absolute risk aversion is declining with income.

While the above analysis has assumed a representative individual, one can easily incorporate heterogeneous taxpayers, for example, in terms of gross income, Y ; the degree of absolute risk aversion, R_a ; or the marginal tax rate, t . Thus, the analytical model can be used to predict that individuals facing a higher marginal tax rate, t , are more likely to evade and not to pay any taxes, since the maximal and the minimal fines $f_{\max}$ and $f_{\min}$ increase with t for $\alpha < 1$.

Relating the predictions derived above to the findings obtained in the analysis of crime, it may be observed that the model of criminal activity often employed is based on the assumption that a crime is either committed or not, while the extent of criminal activity per individual is constant. Higher fines and a greater detection probability reduce the incentives to undertake criminal actions, while a higher potential gain will raise them. The latter prediction may be compared to a change in the tax rate, t , derived above. The impact of a higher gross income, or of wealth, depends on whether income also increases disposable income when the criminal is penalized, *inter alia*. However, the degree of risk aversion does not play a role in the basic setup. These partial differences with respect to the effect of changes in exogenous parameters indicate that predictions can depend crucially on the underlying view of the illegal activity. Does it represent a simple portfolio choice with one safe and another risky asset, as in the case of tax evasion, or does it constitute an endeavor which can be separated from other income-generating activities?

Extensions

The basic deterrence model of income tax evasion has been expanded in numerous ways. We subsequently sketch two extensions which further clarify the sensitivity of the predictions but also the flexibility of the analytical approach. First, we incorporate the idea that individuals will generally

be able to decide on the amount of gross income they earn. This decision is likely to result from a trade-off between higher disposable income on the one hand and a greater disutility from generating this income on the other. Hence, utility may be given by $U(Y^s, Y)$ and $U(Y^c, Y)$, depending on whether evasion is detected or not. In addition, $\partial U / \partial Y < 0$ captures the disutility of generating income. In an early contribution, Pencavel (1979) showed that virtually all predictions developed above will not necessarily hold in such a setting. The reason is that any activity which makes tax evasion less attractive also reduces the incentives to generate income. This reduction in the tax base, in turn, lowers evasion activities for a given underdeclaration and, hence, strengthens the incentives to evade. The net effect is generally uncertain because of the differential changes in the marginal utility of disposable income (i.e., $\partial U / \partial Y^s$ and $\partial U / \partial Y^c$) and from generating Y (i.e., $\partial U / \partial Y$). This first extension is an impressive example of the sensitivity of predictions with regard to incorporating additional choice variables.

The individual considered above has occasionally been termed an “amoral tax payer” (Crocker and Slemrod 2005, p. 1595), because the tax evasion decision results solely from the comparison of monetary gains and losses. Therefore, secondly, the question arises how the optimal underdeclaration will be affected if there is a norm with regard to paying taxes. In a simple extension of the basic model, it can be presumed that tax evasion imposes a utility loss on individuals who evade taxes. It has, *inter alia*, been assumed that this loss (1) is constant, (2) increases in the extent of individual tax evasion, (3) depends on how tax revenues are spent, or (4) varies with an aggregate measure of tax evasion (see Alm and Torgler 2011). While the existence of a norm imposes additional costs of tax evasion in cases (1) and (2) and, therefore, mitigates such activities, the impact in cases (3) and (4) is less obvious. This can be illustrated by assuming that the utility loss from violating the norm of paying taxes varies across individuals and becomes weaker the more people evade taxes. Then, the model may have (at least) two equilibria. In the first, many or all individuals evade taxes, and the

norm does not really bite. In the other equilibrium, few individuals evade taxes. Therefore, the norm imposes substantial costs of evasion, and this helps to stabilize the equilibrium with few people underdeclaring income. In such a setting, the impact of changes in exogenous parameters can be reversed. To illustrate, suppose that higher fines weaken the societal norm of paying taxes. In this case, more severe penalties will reduce evasion, *ceteris paribus*, but weaken the norm and may induce a move from a low-evasion to a high-evasion equilibrium. In this case, the standard prediction that higher fines reduce illegal activities may no longer hold. This second extension clarifies that the standard deterrence model of tax evasion is flexible enough to be applicable to taxpayers whose preferences include non-monetary components such as norms.

An Institutional Perspective

Given the importance of the assumptions underlying the model presented above, it is instructive to view them in light of essential features characterizing real-world tax and enforcement systems. In most OECD countries, the nominal income tax rises with income, suggesting that the tax system is progressive. Moreover, the marginal tax burden on wages, taking into account exemptions and government benefits, also generally rises with income, although the marginal rate may decline at specific income levels (OECD 2013). Therefore, the tax rate depends on gross income, $t = t(Y)$, or declared income $Y - X$ in the case of successful evasion. Since it would be optimal to overdeclare income if the marginal tax rate is negative, increasing marginal tax rates have usually been analyzed. While the impact of changes in the enforcement system is generally unaffected by the nature of the tax system, the consequences of changing the marginal tax rate or the progressivity of the tax schedule can also depend on what individuals decide on, namely, the magnitude of the underdeclaration (as in this setting) or of voluntary tax payments (cf., e.g., Yitzhaki 1987; Goerke 2003). Consequently, the tax schedule on its own may affect tax evasion.

The penalty rates in many OECD countries are considerably lower than 100%, even for severe cases of tax evasion (OECD 2009, p. 136 ff). Additionally, the cursory evidence available suggests that the detection probability with regard to income tax evasion is perhaps as low as 1% (see Slemrod 2007 for corresponding information for the United States). In order to integrate this information into the analytical setup, suppose that the utility function u is given by $u(Y) = Y^{1-a}/(1-a)$ and hence features constant relative risk aversion, $R_r(Y) := -u''(Y)Y/u'(Y) = a$. Moreover, the fine is a function of the amount of taxes evaded ($\alpha = 1$); the marginal fine, f , equals 2; the detection probability is assumed to be 10% ($p = 0.9$); and the tax rate is set to $t = 1/3$. Substituting these values into the basic model (cf. Eqs. 1 and 2), the fraction of gross income, Y , which is optimally underdeclared will only be less than 100% if relative risk aversion, R_r , exceeds two. Moreover, the optimal underdeclaration shrinks with R_r , given the above specification of the utility function. If, for example, a value of $R_r = 10$ is assumed, the individual would still underdeclare about 22% of the gross income. Therefore, it has been argued that the standard model seriously overpredicts tax evasion for plausible values of relative risk aversion, R_r , such as between one and five, given the parameters of the tax enforcement system observed in most countries (cf. Alm et al. 1992; Feld and Frey 2002, *inter alia*).

The response to this criticism has been manifold: Firstly, it has been argued that the payoff of taxpayers is not only affected by the monetary gains and cost of evasion activities but also by the gain of adhering to, or the cost of violating, a social norm, as outlined above. Moreover, the gain may be altered, for example, by whether taxpayers can decide on and approve of the use of tax revenues or how they perceive tax authorities. Secondly, the use of alternative specifications of preferences has been suggested, such as rank-dependent expected utility or prospect theory (cf. Alm and Torgler 2011; Hashimzade et al. 2013). Thirdly, it has been maintained that the numerical example provided above is not an appropriate one with regard to the decision of wage earners but only with respect to self-employed or small businesses (Slemrod 2007).

Wage income is generally subject to withholding regulations. Therefore, the probability that evasion of such income will be detected may approach 100%. A detection probability of 50%, however, would eradicate all evasion incentives in the above numerical example, and the deterrence model of tax evasion can, thus, be reconciled with the data.

Tax Evasion and the Economics of Crime

As mentioned at the outset, the investigation of tax evasion is often interpreted as an application of the economic analysis of crime. However, the perspectives of the two approaches are fundamentally different. A large majority of contributions on tax evasion ask either positive or incrementally normative questions, such as how the tax structure affects evasion activities or whether a certain tax structure is to be preferred to another (see, e.g., the survey by Slemrod and Yitzhaki (2002), in which only one (long) out of eight sections deals with normative issues). The economic analysis of crime focuses strongly on the enforcement of legal rules by public institutions, and the “general problem of public law enforcement may be viewed as one of maximizing social welfare” (Polinsky and Shavell 2007, p. 406). A basic result of this approach is that in the presence of risk-neutral individuals and monetary fines, which have no direct welfare effects, the optimal expected monetary fine should equal the harm caused by a crime. If the sanction is non-monetary, such as a prison sentence, the expected penalty should be lower because a nonmonetary sanction increases enforcement costs and, thus, lowers welfare.

In the analysis of tax evasion, however, such normative issues have played a comparatively minor role. In one important exception, Slemrod and Yitzhaki (1987) inquire as to what the optimal size of a tax collection agency is in the presence of risk-averse individuals. For a given amount of tax revenues, less tax evasion mitigates income variability, and this reduction in the “excess burden of tax evasion” (Slemrod and Yitzhaki 1987, p. 187) represents the welfare gain from reducing evasion activities. Higher enforcement costs constitute the

welfare loss due to fighting tax evasion. The optimal degree of law enforcement is attained when the revenue effect of stricter enforcement still exceeds the resource costs of achieving this revenue impact. The reason is that the revenue gain is mainly distributionary and has no direct welfare impact in a setting with identical individuals, while costs of enforcement reduce welfare. This finding resembles those obtained in the economic analysis of crime.

Given the different perspectives, the literature on tax evasion and the contributions on public law enforcement have also approached many extensions of the basic settings in alternative ways. To illustrate, we consider the nature of penalties and settlements.

From an economics of crime perspective, non-monetary sanctions, such as prison sentences, can have two advantages over fines. Firstly, the financial means of a tax evader may be insufficient to pay a fine. However, imprisonment is feasible irrespective of wealth so that nonmonetary penalties may still deter illegal activities when monetary fines no longer have this effect. Secondly, imprisonment generally limits future crimes. Such an incapacitation effect is less likely to occur in the case of monetary penalties. One important disadvantage of nonmonetary penalties is the higher cost of enforcing such penalties. In most countries, the penalties for evading personal income taxes are represented by monetary fines. However, for severe cases of tax evasion, prison sentences can also be imposed (cf. OECD 2009, Table 31). Nonetheless, questions such as (1) what are the effects of monetary and nonmonetary penalties on tax evasion?, (2) when should imprisonment be used and sentences be suspended?, and (3) what is the optimal combination of fines and imprisonment? have not figured prominently in contributions on tax evasion. This is in contrast to the literature on public law enforcement.

Settlements, that is, agreements between an offender and authorities to terminate or avoid a court trial in exchange for accepting a penalty, have received substantial attention in the economic analysis of crime. Settlements can be desirable because they reduce the costs of law enforcement. Furthermore, risk-averse individuals may prefer

certain penalties to uncertain court outcomes. The main disadvantage of settlements is that they will be attractive to offenders only if they effectively imply a lower penalty. This dilutes the deterrence effect of sanctions. In addition, a settlement may hinder the detection of all illegal activities of an offender and can prevent the development of precedents. While such aspects of settlements have been discussed in the literature on the public enforcement of law (Polinsky and Shavell 2007, p. 435 f), there are few relevant contributions relating to tax evasion (Macho-Stadler and Pérez-Castrillo 2004; Franzoni 2004).

The relative infrequency of settlements may be due to the fact that trials in cases of tax evasion are much less frequent than for criminal activities such as theft, fraud, physical injury, or murder. However, the perspective can also be reversed. Often, tax authorities impose a penalty. This procedure may be interpreted as the tax authority's (pretrial) proposal of a settlement. Accordingly, the relevant question in the context of tax evasion may not be whether settlements are beneficial but why they are used so extensively.

The two above examples clarify that the investigation of topics analyzed in the public enforcement of law may also generate additional insights in the context of income tax evasion. Other such issues may relate to the self-reporting of past tax evasion activities, the treatment of repeat offenders, the employment of tax advisors, corruption among enforcement agents, and the role of marginal deterrence. The analysis of such topics will be especially rewarding if institutional features of tax evasion activities are taken into account. Such investigations would help to clarify whether or not predictions based on general models of illegal behavior carry over to the more specific settings applicable to the investigation of income tax evasion.

References

- Allingham MG, Sandmo A (1972) Income tax evasion: a theoretical analysis. *J Public Econ* 1(3–4):323–338
- Alm J (1999) Tax compliance and tax administration. In: Hildreth WB, Richardson JA (eds) *Handbook on taxation*. Marcel Dekker, New York, pp 741–768
- Alm J (2012) Measuring, explaining, and controlling tax evasion: lessons from theory, experiments and field studies. *Int Tax Public Financ* 19(1):54–77
- Alm J, Torgler B (2011) Do ethics matter? Tax compliance and morality. *J Bus Ethics* 101(4):635–651
- Alm J, McClelland GH, Schulze WD (1992) Why do people pay taxes? *J Public Econ* 48(1):21–39
- Andreoni J, Erard B, Feinstein J (1998) Tax compliance. *J Econ Lit* 36(2):818–860
- Cowell FA (2004) Carrots and sticks in enforcement. In: Aaron H, Slemrod J (eds) *The crisis in tax administration*. Brookings Institution Press, Washington, DC, pp 230–275
- Crocker KJ, Slemrod J (2005) Corporate tax evasion with agency costs. *J Public Econ* 89(9–10):1593–1610
- Feld LP, Frey BS (2002) Trust breeds trust: how taxpayers are treated. *Econ Gov* 3(2):87–99
- Franzoni LA (2004) Discretion in tax enforcement. *Economica* 71(283):369–389
- Franzoni LA (2009) Tax evasion and avoidance. In: Garoupa N (ed) *Criminal law and economics*, vol 3, 2nd edn, *Encyclopedia of law and economics*. Edward Elgar, Cheltenham/Northampton, pp 290–319
- Goerke L (2003) Tax evasion and tax progressivity. *Public Financ Rev* 31(2):189–203
- Hashimzade N, Myles GD, Tran-Nam B (2013) Applications of behavioural economics to tax evasion. *J Econ Surv* 27(5):941–977
- Macho-Stadler I, Pérez-Castrillo D (2004) Settlement in tax evasion prosecution. *Economica* 71(283):349–368
- Marchese C (2004) Taxation, black markets, and other unintended consequences, Chapter 10. In: Backhaus JG, Wagner RE (eds) *Handbook of public finance*, vol 1. Kluwer, Boston, pp 237–275
- OECD (2009) *Tax administration in OECD and selected non-OECD countries: comparative information series* (2008). OECD, Paris
- OECD (2013) *Taxing wages 2013*. OECD, Paris
- Pencavel JH (1979) A note on income tax evasion, labor supply, and nonlinear tax schedules. *J Public Econ* 12(1):115–124
- Polinsky AM, Shavell S (2007) The theory of public enforcement of law, Chapter 6. In: Polinsky AM, Shavell S (eds) *Handbook of law and economics*, vol 1. Elsevier, Amsterdam, pp 403–456
- Sandmo A (2012) An evasive topic: theorizing about the hidden economy. *Int Tax Public Financ* 19(1):5–24
- Slemrod J (2007) Cheating ourselves: the economics of tax evasion. *J Econ Perspect* 21(1):25–48
- Slemrod J, Yitzhaki S (1987) The optimal size of a tax collection agency. *Scand J Econ* 89(2):183–192
- Slemrod J, Yitzhaki S (2002) Tax avoidance, evasion, and administration, Chapter 22. In: Auerbach AJ, Feldstein M (eds) *Handbook of public economics*, vol 3. Elsevier, Amsterdam, pp 1423–1470
- Yitzhaki S (1974) A note on income tax evasion: a theoretical analysis. *J Public Econ* 3(2):201–202
- Yitzhaki S (1987) On the excess burden of tax evasion. *Public Financ Q* 15(2):123–137

Tax Structure

► Fiscal System

Tax Systems

► Fiscal System

Taxation

► Fiscal System

Telecommunications

Alden F. Abbott
Heritage Foundation, Washington, DC, USA

Abstract

Telecommunications involves the transmission of information without change in the form or content. Telecommunications networks increase in value as the number of users rises (“network effect”) and the rise of the Internet and wireless communications has bestowed huge economic benefits on countries worldwide. The development of telecommunications has been heavily influenced by regulatory regimes. Regulation in the United States has featured efforts to restrain private monopoly power and promote market allocation of spectrum, while European regulation recently has focused on the privatization of former state telecommunications monopolies and the transition to a pan-European regulatory regime. As the United States transitions away from US government stewardship of the Internet and toward a more global form of Internet governance, the International Telecommunication Union is seeking to play a greater role in Internet oversight. Emerging telecommunications policy issues include

the privatization of the radio spectrum, “net neutrality” regulation aimed at treating all data traffic equally, and the growth of “cloud computing” and “big data” compilations.

Conceptual Overview

Telecommunications may be broadly defined as “the transmission, between or among points specified by the user, of information of the user’s choosing, without change in the form or content of the information as sent and received” (47 U.S. Code § 153 [2011](#)). The information transmitted may take the form of data, text, audio, and/or visual materials.

Whether the telegraph, the telephone, or the Internet, the full value of a telecommunications medium may not be realized until there is a network of users that can transfer information among themselves. This addition in value can be attributed to what is known as a “network effect.” There is a significant difference, however, between network effects and network externalities. While the term “network effect” refers to the increase in the value of a network that corresponds with the increase in the number of participants in the network, “network externalities” occur when market participants do not fully “internalize” (obtain) that increase in value, for example, the owners of a private network (Liebowitz and Margolis [1998](#)).

The network effect is an example of a positive externality, in which the action of one individual benefits another individual without any mutual agreement to make compensation for that benefit (Easley and Kleinberg [2010](#)). A prime example is the benefit gained from additional participants in a social networking site, which raises the potential for any given individual to network through that site, even though no participant is explicitly compensated for joining (Easley and Kleinberg [2010](#)). The Internet illustrates the network externality of the social media to a grand degree, with over 200 million Americans having broadband access to the Internet (Broadband Fact Sheet Pew Research Center [2013](#)) and an estimated 2.9 billion Internet users worldwide (International

Telecommunication Union Key ICT Data 2005 to 2015), bringing major economic benefits with it. The Internet “alone accounted for 21% of the GDP growth in mature economies from [2006 to 2011]”: Brazil, Canada, China, France, Germany, India, Italy, Japan, Korea, Russia, Sweden, the UK, and the United States (Manyika and Roxburgh 2011). According to one study, among such “high-income” economies from 1980 to 2002, “a 10% increase in broadband penetration yielded an additional 1.21 percentage points of GDP growth,” while the same level of broadband penetration in “low- and middle-income economies” yielded 1.38 percentage points of GDP growth (International Telecommunications Union Impact of Broadband 2012).

This value of an established network due to an existing pool of participants has led some to worry that an earlier-created inferior product with a pre-existing network of participants would win out against a later-arriving but superior product without such a network. If two competing networks produce similar but incompatible products, the product with the greater market share will have an advantage. If the network effect does not diminish, the advantaged network could be seen as a natural monopoly (Liebowitz and Margolis 1998).

However, the network effect is not limitless. If additional participants cease to provide value to the existing network of participants, meaningful competition among networks may be possible. Subjective valuation of the network by individuals may also differ, allowing multiple competing networks to coexist (Liebowitz and Margolis 1998). Similarly, different individuals may value individual participants added to the network differently, which could provide an opening in the market for separate networks to serve different groups based on the actual makeup of the participants (Liebowitz and Margolis 1998).

The United States and Europe possess the most mature telecommunications regulatory regimes. An overview of these systems provides insight on the sorts of problems national governments face as they oversee the development of their telecommunications sectors. Following the overview, this essay briefly describes international telecommunications regulation and emerging policy issues.

Telecommunications Regulation in the United States of America (USA) and the European Union (EU)

In one of the first attempts to regulate telecommunications as a whole, the US Congress passed the Communications Act of 1934, establishing the Federal Communications Commission (FCC) to take responsibility for regulating radio and wire communications, which had previously been dealt with by separate agencies (Communications Act of 1934). Congress amended the law with the Telecommunications Act of 1996, allowing for federal preemption of local regulations that acted as barriers to entry and more competition in the long-distance market and requiring incumbents to allow access to their networks at wholesale prices (Telecommunications Act of 1996). The FCC also has oversight over wireless services and radio and television broadcasting (“What We Do,” Federal Communications Commission 2015), which are licensed to use certain portions of the electromagnetic spectrum.

In 1974, the Department of Justice sued under the Sherman Antitrust Act to rein in the dominant US national telecommunications company, in filing suit against AT&T to modify an existing 1949 consent order. The suit alleged that AT&T, the monopoly provider of local telephone service in most parts of the United States, had engaged in various anticompetitive acts to maintain monopoly power in the provision of long-distance telephone service (*U.S. v. AT&T* 1978). The suit ended in divestiture for AT&T in a modified consent decree in 1983, effectively breaking up the telecom giant (*U.S. v. AT&T* 1983). Since 1983, additional antitrust actions have been directed at major US telecommunications companies. For example, in 1998, the United States sued under the Clayton Antitrust Act to block the merger of AT&T with TCI (the merger went forward subject to a consent agreement) (*U.S. v. AT&T* 1999), and Verizon was sued under the Sherman Antitrust Act for conduct that had been found to violate the Telecommunications Act of 1996 (the antitrust suit failed) (Verizon Communications 2004).

The telecommunications regulatory regime in the United States receives policy advice and technical support from the National

Telecommunications and Information Administration (NTIA), established in 1978. The NTIA advises the President and works with executive branch agencies to develop policy on telecommunication and information issues and manages the federal use of the electromagnetic spectrum by administering grants and holding auctions to assign licenses (About NTIA 2015) (the NTIA has worked with the FCC to reassign certain spectrum frequencies from public use to private use). The NTIA also has been indirectly involved in Internet governance (and, in particular, the administration of the Internet Domain Name System) through its administration of a US government contract with the Internet Corporation for Assigned Names and Numbers, or ICANN, and the Internet Assigned Numbers Authority, or IANA (ICANN 2015).

In Europe, telecommunications was mostly provided by state-owned monopolies until the 1970s, when pushes for a smaller role for governments in the telecommunications market began to rise (Bauer 2013). In the 1980s, the European Commission began to promote a vision of a pan-European telecommunications sector (Bauer 2013). The EU successfully liberalized terminal equipment, value-added, and other services and had opened all services up to competition by 1998 (Bauer 2013). By 2012, all but one member, Luxembourg, had at least partially privatized their telecommunications sector (Bauer 2013). Since then, the European Union has moved toward facilitating regulation and standardization of telecommunications throughout EU member states as part of the *Digital Agenda for Europe*. The EU's framework is made up of five directives, the Framework Directive, the Access Directive, the Authorisation Directive, the Universal Service Directive, and the Directive on Privacy and Electronic Communications, and two regulations, the Regulation on Body of European Regulators for Electronic Communications (BEREC) and the Regulation on Roaming on Public Mobile Communications Network (Digital Agenda for Europe Telecoms Rules 2015). In 2012, the European Parliament and Council approved the first Radio Spectrum Policy Programme (RSPP) to set objectives, make recommendations, and establish principles for the administration of the radio spectrum

(Digital Agenda for Europe Radio Spectrum Policy Program 2015). The overall aim of these efforts is to move toward a pan-European approach to telecommunications regulation, in place of the nation-specific regulatory regimes that currently exist within the EU.

Other jurisdictions throughout the world employ a variety of regulatory schemes, with the trend being toward provision of telecommunications services through private operators rather than the state (Struzak 2003). The recent fast international growth of mobile wireless telecommunications, which has rapidly spread the availability of telecommunications services to new populations (especially the poor), is another feature that is expanding the telecommunications network effect and, in particular, widespread access to the Internet.

International Telecommunications Regulation

While the United States has historically acted as steward of the Internet, the NTIA has received pressure to move toward a more global model of Internet governance. As a step toward this model, the NTIA asked ICANN to turn over its role in coordinating the Internet's Domain Name System (DNS) to the international community as part of a program of privatization (NTIA Announces Intent to Transition 2014). In so doing, the NTIA, acting consistently with resolutions of the US Senate and House of Representatives (S. Con Res. 50, 112th Cong., 2012), specified that it would not support handing its responsibility to any governmental or intergovernmental. The current contract expires on September 30, 2015. Some commentators have expressed concerns about the implications of this transition for the future of Internet governance (Schaefer et al. 2015), while others support NTIA's initiative (Llansó 2015; Tennenhouse 2014).

The International Telecommunication Union (ITU), founded in 1865 as the International Telegraph Union (International Telecommunication Union History 2015), is now a United Nations Agency focused specifically on information and communication technologies (ICTs). The

organization has proposed to take up the mantle of the NTIA, resolving to “to explore ways and means for greater collaboration and coordination between ITU and relevant organizations involved in the development of IP-based networks and the future Internet, through cooperation agreements, as appropriate, in order to increase the role of ITU in Internet governance so as to ensure maximum benefits to the global community,” specifically mentioning ICANN in its 2014 Plenipotentiary Resolution 102 (ITU Plenipotentiary Resolution 2014).

Emerging Policy Issues

Proposals to increase privatization of the radio spectrum and to regulate more heavily the provision of Internet communications are the subject of considerable recent debate. While auctions have been the chosen method of allocating newly privatized spectrum blocs in the United States, the process is not without controversy. For example, after a January 2014 auction for wireless licenses, AT&T accused Dish Network Corporation of intentionally driving up the price of spectrum licenses at auction, arguing that the coordinated bidding strategy involving three entities artificially inflated the perceived demand for the licenses. Dish Network asserted that it fully complied with FCC rules when implementing the bidding strategy, but there have been calls for the FCC to intervene to prevent such behavior (Gryta and Ramachandran 2015).

More generally, some commentators have advocated in favor of “net neutrality,” the principle that Internet service providers (ISPs) should treat all data traffic equally, in the regulation of ISPs (Lessig and McChesney 2006). On February 26, 2015, The FCC adopted a rulemaking that would, among other things, classify the Internet as a “public utility” regulated under Title II of the US Telecommunications Act, invoking the cause of net neutrality. In so acting, the FCC opined that in absent regulatory action, the Internet service providers may throttle data speeds based on content or offer high-paying customers prioritization in data traffic (FCC News Release 2015). Opponents voice concern that the regulation of the

Internet as a public utility will result in the entrenchment of larger Internet service providers (ISPs) at the expense of smaller providers, a slowdown in Internet speeds (or the rate of increase in speeds), and Internet access rates (Summary of Pai Testimony 2015). The rule is likely to face challenges from opponents, and the future of Internet regulation by the FCC remains uncertain (Hughes 2015). Concerns about imposing “net neutrality” and various other constraints on the provision of Internet service may be expected in other jurisdictions as well.

Ongoing changes in telecommunications infrastructure will also drive policy debates. Commentators have predicted the increased use of cloud services, investment in telecommunications infrastructure, and replacement of telephone lines with broadband in the developed world and with the rise in multi-device telecommunication services (Lopez 2015). The problem of “big data,” or collections of information that “had grown so large that the quantity being examined no longer fit the memory that computers use for processing,” has been a focus of discussion among telecommunications experts (Pflugfelder 2013; International Telecommunications Union Press Release December 2014).

Conclusion

The importance and diffusion of telecommunications services may be expected to grow apace, with the rapid expansion of wireless services and Internet-related transactions. This development promises to bestow large and growing benefits on producers and consumers worldwide. Nevertheless, questions about the future of Internet governance and regulation in general create some uncertainty as to the manner in which telecommunications (and, in particular, Internet) service provision will grow and evolve.

References

- 47 U.S. Code § 153 (2011)
- Bauer JM (2013) The evolution of the European regulatory framework for electronic communications. Catedra Telefonica, Institut Barcelona D’Estudis Internacionals,

- pp 6–7. <http://catedratelefonica.ibe.org/wp-content/uploads/2013/06/IBEI-%C2%B7-41.pdf>. Accessed 13 Mar 2015
- “Broadband Technology Fact Sheet,” Pew Research Center (2013). <http://www.pewinternet.org/fact-sheets/broadband-technology-fact-sheet>. Accessed 2 Mar 2015
- Communications Act of 1934, Public Law 73-416
- Corwin PS If Stakeholders already control the Internet, why NETmundial and the IANA transition? CircleID. Accessed 13 Mar 2015
- Digital Agenda for Europe, “Radio Spectrum Policy Program,” (2015) <https://ec.europa.eu/digital-agenda/node/173>. Accessed 2 Mar 2015
- Digital Agenda for Europe, “Telecoms Rules,” (2015) <https://ec.europa.eu/digital-agenda/en/telecoms-rules>. Accessed 17 Mar 2015
- Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, pp 509–510. <http://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch17.pdf>. Accessed 2 Mar 2015
- FCC News Release, “FCC Adopts Strong, Sustainable Rules to Protect the Open Internet,” *Federal Communications Commission*, February 26, 2015. <http://www.fcc.gov/document/fcc-adopts-strong-sustainable-rules-protect-open-internet>. Accessed 2 Mar 2015
- Gryta T, Ramachandran S (2015, February 20) AT&T Says Dish Bidding Practices Skewed Results in Spectrum Auction. The Wall Street Journal. <http://www.wsj.com/articles/at-t-says-dish-bidding-practices-skewed-results-in-spectrum-auction-1424454394>. Accessed March 2, 2015
- Hughes S (2015, February 25) FCC’s Net Neutrality Rules Expected to Unleash Court Challenges. The Wall Street Journal. <http://www.wsj.com/articles/fccs-net-neutrality-rules-expected-to-unleash-court-challenges-1424919940>. Accessed 12 Mar 2015
- “ICANN,” *National Telecommunications and Information Administration*, 2015, <http://www.ntia.doc.gov/category/icann>. Accessed 12 Mar 2015
- International Telecommunication Union, “History,” 2015. <http://www.itu.int/en/about/Pages/history.aspx>. Accessed 2 Mar 2015
- International Telecommunications Union, Impact of Broadband on the Economy: Research to Date and Policy Issues, April 2012, p 4, http://www.itu.int/ITU-D/treg/broadband/ITU-BB-Reports_Impact-of-Broadband-on-the-Economy.pdf. Accessed 12 Mar 2015
- International Telecommunications Union, Key ICT Data for the World, 2005 to 2015, www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx. Accessed 2 Mar 2015
- International Telecommunications Union Press Release, “ITU Telecom World 2014 Highlights, Innovations, Technologies, and Ideas Shaping Future of ICTs: Interactive Debates and Showcases Focus on Future of Technology and Its Impact on Society,” International Telecommunications Union, December 10, 2014, http://www.itu.int/net/pressoffice/press_releases/2014/76.aspx#.VQil9I7F98E. Accessed 17 Mar 2015
- ITU Plenipotentiary Resolution 102 (Rev. Busan 2014), p 4, <http://www.itu.int/en/plenipotentiary/2014/Documents/final-acts/pp14-final-acts-en.pdf>. Accessed 13 Mar 2015
- Lessig, L, McChesney R (2006, June 8) No Tolls on the Internet. The Washington Post. <http://www.washingtonpost.com/wp-dyn/content/article/2006/06/07/AR2006060702108.html>. Accessed 12 Mar 2015
- Liebowitz SJ, Margolis SE (1998) Network Externalities (Effects) The New Palgrave Dictionary of Economics and the Law. <http://www.utdallas.edu/~liebowit/palgrave/network.html>. Accessed 2 Mar 2015
- Llansó E Don’t Let Domestic Politics Derail the NTIA Transition Center for Democracy and Technology. <https://cdt.org/blog/dont-let-domestic-politics-derail-the-ntia-transition/>. Accessed 12 Mar 2015
- Lopez M (2015) What mobile cloud means for your business. Fortes. www.forbes.com/sites/marite/lopez/2015/04/06/what-mobile-cloud-means-for-your-business/. Accessed 20 May 2016
- Manyika J, Roxburgh C (2011) The Great Transformer: The Impact of the Internet on Economic Growth and Prosperity (New York: McKinsey Global Institute), pp 1–2, http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_great_transformer. Accessed 16 Mar 2015
- NTIA Announces Intent to Transition Key Internet Domain Name Functions. National Telecommunications and Information Administration, Office of Public Affairs, March 14, 2014, <http://www.ntia.doc.gov/press-release/2014/ntia-announces-intent-transition-key-internet-domain-name-functions>. Accessed 2 Mar 2016
- “Overview,” *International Telecommunication Union*, <http://www.itu.int/en/about/Pages/overview.aspx>. Accessed 2 Mar 2015
- Pflugfelder EH (2013, August) Big Data, Big Questions. Communication Design Quarterly 1.4, p 18
- Schaefer BD, Rosensweig PS, Gattuso JL. Time is Running Out: The U.S. Must be Prepared to Renew the ICANN Contract. Heritage Foundation Issue Brief No. 4340, February 3, 2015, <http://www.heritage.org/research/reports/2015/02/time-is-running-out-the-us-must-be-prepared-to-renew-the-icann-contract%20-%20fn3>. Accessed 12 Mar 2015
- S.Con.Res.50, 112th Cong, 2d Sess., Sept. 22, 2012, pp 2–3, <http://thehill.com/images/stories/blogs/flooraction/jan2012/sconres50.pdf>. Accessed 13 Mar 2015
- Struzak R (2003, 2–22 February) Introduction to International Radio Regulations. Lectures given at the School on Radio Use for Information And Communication Technology Trieste, p 6, http://www.iaea.org/inis/col/lection/nclcollectionstore/_public/38/098/38098197.pdf. Accessed 12 Mar 2015
- “Summary of Commissioner Pai’s Oral Dissent on Internet Regulation,” Press Release, Federal Communications Commission, February 26, 2015, <http://www.fcc.gov/document/summary-commissioner-pais-oral-dissent-internet-regulation>. Accessed 2 Mar 2015
- Telecommunications Act of 1996, Public Law 104-104
- Tennenhause D (2014, March 3) “Microsoft Applauds US NTIA’s Transition of Key Internet Domain Name Functions,” Microsoft on the Issues. <http://blogs.microsoft>

com/on-the-issues/2014/03/17/microsoft-applauds-us-intias-transition-of-key-internet-domain-name-functions/. Accessed 12 Mar 2015

U.S. v. AT&T, 461 F.Supp. 1314, 1318 (1978)

U.S. v. AT&T, 552 F. Supp. 131, 131 (D.D.C. 1983)

U.S. v. AT&T (Complaint), 1999 WL 1211462 (D.D.C. Aug. 23, 1999)

Verizon Communications v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398 (2004)

“What We Do,” *Federal Communications Commission*, <http://www.fcc.gov/what-we-do>. Accessed 11 Mar 2015

Further Reading

Haring, J Telecommunications. In: The concise encyclopedia of law and economics (2nd ed.), <http://www.econlib.org/library/Enc/Telecommunications.html>. Accessed 17 Mar 2015

World Bank and International Telecommunications Union, Telecommunications Regulation Handbook, Colin Blackman and Lara Srivastava eds. (10th Anniversary ed. 2011)

Television Market

Paola Savini

Autorità per le garanzie nelle comunicazioni – AGCOM, Rome, Italy

Definition

The television market is made up of content producers, broadcasting network operators, packagers and TV service providers, carriers and network providers, as for the supply side, as well as the so-called “audience,” for the demand side. It is a complex TWO-SIDED market, mature and strongly regulated. It is characterized by cultural, political, and industrial elements. These different components are deeply rooted within the national television markets but, at the same time, respond to global economic dynamics and supranational laws. The television industry can be divided into three main activities, corresponding to three different economic functions. These are the programs’ production, the organization of television programs in a schedule, and signal distribution in the territory. Almost a century after its invention and standardization, television is nowadays evolving again, as the competition provided by

newer and different screens (pc, smartphone, tablet) is increasing, thus offering the audience multiple options to access linear and on demand audiovisual contents.

Television Market: An Overview

In 2016, the European Union hosts more than 4,000 television services (European Audiovisual Observatory 2017), ranging from the historical generalist channels to the thematic ones, offering sports, film and series, children’s programs, and documentary to millions of habitants. Also many Video on demand service providers operate within the European countries, and hundreds of such services are available everywhere (Ofcom 2016).

The one-century old television is, in fact, still the mass medium *par excellence*, offering entertainment and information worldwide and accounting for about 20% of the not-sleeping hours in many countries (Ofcom 2017). Most countries currently license a high number of television channels which can operate at a national or sub-national (local) level, now accessible via different platforms, such as the digital terrestrial television (DTT), the cable, the satellite, or the Internet protocol television (IPTV), giving to the public the chance, never experienced before, to watch a huge variety of audiovisual contents. This audiovisual consumption may now happen through the TV set, but also on mobile devices, personal computers, tablets, or game consoles.

All of these changings are both technology and policy driven, as more than other media, television is characterized by cultural, political, and industrial elements (see entry on ► “Media”). These different components are deeply rooted in the national contexts of each television company (hereinafter, also called “broadcasters”) but, at the same time, respond to global economic dynamics and supranational laws (Freedman 2008).

Legal Framework

As regards the legal framework, numerous forms of public authorities exercise some kind of regulation over broadcasting activities, at different levels: apart from national legislative Chambers,

which regulate particularly the governance of the Public Service Media (PSM) and the content' level, there are the Communitarian institutions within the EU framework, some governmental administrative authorities, some independent regulatory authorities (see EPRA – European Platform of Regulatory Authorities – 52 authorities from 46 countries are members – which cooperates with the European Commission, the Council of Europe, the European Audiovisual Observatory and the Office of the OSCE Representative on Freedom of the Media), as well as the national Courts for specific issues. Within some Member State, also some forms of self-regulation apply, especially for Public Service Media.

Media ownership and concentration issues have traditionally been regulated at national level, with regard to media pluralism and freedom of expression in particular. However, some common principles may be found within the “*EU Charter of Fundamental Rights*” (Article 11) and within the “*European Convention for the Protection of Human Rights and Fundamental Freedoms*” (Article 10).

Differences are huge between European countries, as concerns media ownership: in France, some limitations apply both to the ownership of a single television channel and to the ownership of the number of licenses; in Italy, limitations apply both to the cross-ownership between national and local terrestrial channels and to the revenues within the media sector; in Germany, audience market share limits apply.

Concerning other issues such as programming, advertising, sponsorship, and the protection of certain individual rights, some Council of Europe countries signed already in 1989 the “European Convention on Transfrontier Television” (ETS n.132), which created a legal framework for the free circulation of transfrontier television programs in Europe. In particular, the Convention defined for the first time “audiovisual works of European origin” as the “creative works whose production or co-production is controlled by natural or legal European producers.”

Then, at EU level and with the neighboring countries, in 1989 it became effective a

coordination of national legislations on all audiovisual media, through the so-called “*Television Without Frontiers Directive*” (Directive 89/552/EEC), which aimed to safeguard some public interest objectives, such as cultural diversity, the protection of minors, and the right of reply, and rested on two basic principles: the free movement of European television programs within the internal market and the requirement for TV channels to reserve, whenever possible, more than half of their transmission time for European works (the so-called “broadcasting quotas”).

Having been this Directive substantially amended several times during the years (it is currently again under revision, and a new legislative proposal has been adopted by the European Commission on 25 May 2016), a minimum set of common rules applying to all the 28 Member States, it is now operating for the audiovisual and new media sector in the Union, based on the country of origin principle, ensuring freedom of reception and retransmission of the TV channels as well as advertising regulation, promotion of European works, and protection of minors (AVMSD 2010). All audiovisual media services providers (both television broadcasts and contents selected by viewers on-demand over an electronic communication network) shall display clearly their name and address and need to respect the basic rules provided in the AVMSD, as regards: the prohibition of incitement to hatred based on race, sex, religion, or nationality; the accessibility for people with visual or hearing disabilities; some qualitative and quantitative requirements for commercial communications.

At an even more international level, global radio spectrum and satellite orbits as well as technical standards are regulated by the International Telecommunication Union (ITU), the United Nations specialized agency for information and communication technologies.

Industrial Innovations

As regards the industry level, since the 1970s, technological innovations strongly fostered the enlargement of the television market, generally born as state-owned, and enriched the consumers' experience.

In 1962 the first transatlantic transmission, via the Telstar communications satellite, was made, permitting satellite images to be taken from around the globe. The first videotape recorders were sold to the public in 1965. Color broadcasting began in the USA in 1964. In 1972, the first digital converter changed pictures from the US 525-line format into the 625-line European standard. In the same year, the first cable operators appeared in the USA. An audience measurement system was implemented in the USA in 1973 and the UK in 1977 (the British Audience Research Bureau was created in 1981).

At the beginning of 2000, digitalization burst onto the scene as a disruptive innovation that would be able to revolutionize the entire TV ecosystem once again, at various levels of the TV supply chain. This technological innovation has been institutionally recognized as pro-competitive and thus as having positive direct and indirect influences on media pluralism, as a result (Adda and Ottaviani 2005). For the past 30 years the replacement of analog technologies with digital ones has indeed changed the transport systems of the signal: a long process of redefinition for transmission standards has been set up more recently (2009–2015), for the final switch off of the analogue signal, in the OECD countries. Digital terrestrial television was launched in the UK in 1998, just after digital satellite television. In the United States, the cutoff date for analogue TV was established in 2009, in Germany in 2008 and in Italy and UK in 2012.

Compression systems have also facilitated an increase in the transmission capacity of the networks, resulting in the multiplication of channels and storage, production, and distribution at decreasing costs, thus removing a major bottleneck from the whole TV industry chain. Since the cultural and social benefits that have historically been found in the broadcasting activity, free or low cost access to the spectrum has been granted to the broadcasters.

Digital competition has also revitalized the electronics market through the sale of digital receivers, set-top-boxes, televisions HD ready, as well as the 3D TVs or connected TVs.

Furthermore, as regards consumption habits, traditional live viewing is slowly declining everywhere while time-shifted and nonbroadcast viewing activities (such as Video on demand services or streaming video) are rising, through services delivered over the Internet, such as Netflix, Amazon Prime Video, Hulu, iQIYI, Youku, Ditto TV (now absorbed by Zee5), Now TV, with different percentages of subscription within the European countries and the USA (Ofcom 2017).

Resources and Business Models: Advertising, Public Funds, and Subscriptions

The television industry can be divided into three main activities, corresponding to three different economic functions. These are the programs' production, the organization of television programs in a schedule, and signal distribution in the territory (Owen and Wildman 1992). The signal distribution is undertaken by the presence of economic agents who make available to broadcasters the transmission equipment (carriers or network providers). Each distribution channel – operating through analog or digital terrestrial network, cable, satellite – involves different players (manufacturers, device producers, multiple-system operators, regulatory authorities). If the television service has controlled access (as for instance, pay-TV), packagers and pay-TV service providers will also be involved in some economic activities, such as customer relationships or channels bundling.

The cultural industries of most European and other countries have, historically, focused political attention and industrial policies on the scheduling and distribution aspects, as broadcasting systems were generally born as state-owned ones, while the US cultural industry gave centrality to the productive function and to advertising-based, privately owned, broadcasting models. One fundamental component of the television ecosystem is in fact the content, whose characteristics are important to determine the audience a broadcaster or a service provider wants to achieve. The production function can be

concentrated within the same broadcasters, either directly or through in-house production unites/subsidiaries, or it may result from independent firms (producers). In order to qualify as an “independent producer,” in Europe a production company must be a company (or other business entity) which is independent of broadcasters. The historical origins of European Union policy on self-sufficient audiovisual production and TV distribution found a cornerstone with the already mentioned “Television without Frontiers Directive.” Similar objectives can be found also in some American audiovisual policies, such as the “Prime Time Access Rule” (PTAR) and the “Financial Interest and Syndication Rules” (also known as “Fin-Syn”).

As regards the economic functions, the central activity of the television industry is the exchange of “television communication,” supplied by broadcasting networks and requested by viewers. The fundamental economic agents in the television ecosystem are therefore the television companies (“broadcasters”), who organize daily, weekly, and annual schedules.

This demand and supply of television communication is only exceptionally regulated by the price. In fact, from an economic point of view, the television product is an “information good” (Shapiro and Varian 1998), which can be reproduced and distributed somewhat inexpensively. In the case of free-to-air television, be it state-owned or private, the good offered is a classic “public good.” It is immaterial, it is not destroyed when consumed, and it is not divisible.

Its cost of production is not dependent on the number of people who will use it, that is, who will watch the program. However, the cost of production may affect the number of people who might want to look at the program (the most expensive productions will probably attract the larger audiences), but the cost will be the same, whether the program is seen or not. As “information goods,” they are subject to economies of scale and scope, resulting from the high fixed costs of production of the first copy (“sunk costs”). The marginal cost of this product is equivalent to zero, since a viewer added to the program will not alter in any way the production costs to be incurred. Moreover, the

fact that a person is watching the program does not diminish the ability of another person to see it. Since it is evident that no private operator would be interested in producing such a collective good as “television communication,” unable to achieve earnings from the end user (the viewer), another market intersects with the demand and supply just described, permitting a return on investments: the advertising market.

Advertising

A second product generated by the television industry is, therefore, the “commercial,” requested by companies that produce goods and services to be advertised. The viewer is characterized in this case as a receiver of advertisements, but also as a “product,” offered by the broadcaster to its clients, such as the companies wishing to advertise on its TV channels. The price is the instrument that regulates this market, and it is influenced by variables that are related to the economic cycle, the market structure, and by regulation at national or supranational level (such as the European level). Different information about the viewers (age, gender, jobs, buying habits, and tastes), different kinds of advertising spaces (commercial breaks, sponsorship, product placement, etc.), and types of television programs (sport, news, current affairs, cartoon, sit-coms, TV serials, soaps, TV movies, reality, etc.) influence the value and price of each advertisement sold.

The television ecosystem is thus composed also by advertising agencies and companies that measure television ratings.

Within the EU, the AVMSD set a broad framework concerning some general rules applicable to all forms of commercial communications for audiovisual media services, both linear or non-linear, thus influencing the way broadcasters may collect their principal resource. First of all, all forms of audiovisual commercial communications for cigarettes and other tobacco products as well as audiovisual commercial communication for medicinal products and medical treatment available only on prescription are prohibited, and audiovisual commercial communications for alcoholic beverages shall not be aimed specifically at minors and shall not encourage

immoderate consumption. Then, more in particular, all audiovisual commercial communications shall be limited both in contents and in time; they shall be readily recognizable as such, distinguishable from editorial content (by optical and/or acoustic and/or spatial means) and subliminal techniques are prohibited all over the EU. The forms of advertising which are regulated include teleshopping, sponsorship, and product placement. Sponsorship includes any contribution made by public or private undertakings or natural persons not engaged in providing audiovisual media services or in the production of audiovisual works, to the financing of audiovisual media services or programs with a view to promoting their name, trademark, image, activities, or products. News and current affairs programs shall not be sponsored and also stricter rules may apply during documentaries and religious and children's programs. Product placement is a form of commercial where there is no payment but only the provision of certain goods or services free of charge, with a view to their inclusion in a program, and it is admissible in cinematographic works, films, and series made for audiovisual media services, sports, and light entertainment programs, with an exemption for children's programs.

Public Funded Models

However, advertising-based TV is not the only economic model in the television industry (Picard 2002). In some countries, and especially for state-owned public service broadcasters, a television license (or a broadcast receiving license) is required of the viewers for the reception of the TV service or the possession of a television set. This form of funding is typical of European countries (Josephine 2015), where Public Service Media (PSM) operate according to some national and supranational laws. For instance, in the UK, the British Broadcasting Corporation, established in 1922, has been broadcasting TV since 1936 and the 1954 Television Act established commercial television; in Italy, RAI, Radiotelevisione italiana, began a regular state-owned television service in 1954; in 1956, TVE, Televisión Española, started regular broadcasting in Spain. In 1967 the US Congress created the Public Broadcasting System,

a noncommercial, public television network, funded, however, by congressional appropriations, viewer donations, and private corporate underwriters. During last years, many PSM objectives' revisions occurred, trying to adapt the infrastructure and the contents of historical PSM to multichannel and cross-platform scenarios. At the same time, all the mandates' revisions had to comply with the public service objectives identified in the "Protocol to the Amsterdam treaty on the system of public broadcasting in the Member States" (signed on November 10, 1997), which states that the system of public broadcasting (now PSM) in the Member States is directly related to the democratic, social, and cultural needs of each society and may help to preserve media pluralism. However, economic crisis as well as changing consumption habits both influenced the structure, the governance, and the business models of such services.

In Spain, for instance, advertising and public funded mixed models have been outmatched by public funds and private broadcasters/telecom operators taxation (Ley 7/2010). The same happened in France (this form of taxation has been approved by the Court of Justice of the European Union in 2013). In Finland, instead, direct taxes apply to the citizens, instead of a license, since 2012. In United Kingdom, finally, since April 2017, the independent authority OFCOM compassed the so-called BBC Trust in supervising the implementation of broadcasting legislation over the PSM.

As regards the public funded television business model, within the EU, article 107 of the Treaty on the Functioning of the European Union (TFEU) (ex Article 87 of Treaty establishing the European Community – TEC) provides that any aid granted by a Member State or through State resources in any form, which distorts or threatens to distort competition by favoring certain undertakings or the production of certain goods, shall be incompatible with the internal market (see entry ► "State Aids and Subsidies"). However, Article 106 specifies a number of circumstances (so-called "exemptions") in which the same state aid is acceptable, when some conditions apply, such as the service is clearly of general economic

interest by the EU county concerned, the undertaking in question must be explicitly entrusted by the EU country with the provision of that service, and the ban on state aid must obstruct the performance of the particular tasks assigned to the undertaking. Among these exemptions, some have direct or indirect influence on TV market. As for the direct influence, the public broadcasting service remit, in some circumstances and granting transparency in providing that service, may be included within the exemptions, according to the Amsterdam Protocol. In 2009 the EU Commission published a “Communication on the application of State aid rules to public service broadcasting,” following a 2001 first Communication and then three rounds of public consultations (2008 and 2009), providing again for a derogation for funding granted to broadcasting organizations for the fulfillment of the public service remit so long as it does not affect trading conditions and competition in the EU to an extent that would be contrary to EU interests.

As for the indirect influence, the Commission adopted a revised General Block Exemption Regulation (GBER) in May 2014, which includes aid schemes for audiovisual works. Article 54 of the Commission Regulation lists the specific compatibility criteria applicable to aid schemes for audiovisual works: particularly, the aid must support the script-writing, development, production, distribution, and promotion of audiovisual works, assuming that those products are cultural products as defined by each Member State with a selection and a predetermined list of cultural criteria. Aid may take the form of aid to the production, pre-production, and distribution of audiovisual works. Same Regulation defines also (Article 2) what is considered a “difficult audiovisual work,” which is “identified as such by Member States on the basis of pre-defined criteria when setting up schemes or granting the aid and may include films whose sole original version is in a language of a Member State with a limited territory, population or language area, short films, films by first-time and second-time directors, documentaries, or low budget or otherwise commercially difficult works.”

Finally, another indirect influence of the EU state aid framework on the TV market is the exemption concerning public investments on broadband (see entry on ► [“Public Investments: Broadband”](#)), as the always more convergent audiovisual world is strongly shifting from terrestrial free-to-air television systems or satellite technology to a broadband-driven audiovisual consumption.

Subscription-Based Models

Finally, TV revenues may also come from channel/bundles/content subscriptions, paid by the audience to access the content, both linearly and on demand. This form of revenue accounts for more than a half of the total revenues of the TV market worldwide, in 2016 (Ofcom [2017](#)). Pay-TV revenue is a well-established business model, especially in some country such as USA and Australia, followed by UK and Germany, while in other countries public funded or advertising-based models prevail.

Mostly, recently released films or TV series and sport events (so-called “premium contents”) and back-catalogue films are the most popular types of programs watched on subscription services. These services are in charge of traditional broadcasters, who differentiate business models reorganizing the offer through different platforms and revenue models, or are provided also by Over-the-Top operators. The widening of the TV market encompasses unquestionably the enrichment of the pay-TV offer, and the competition between actors in search for the richest targets (who may pay to access content) is growing, also thanks to the convergence between technologies and platforms.

Future Directions

During the last years, the TV market has been changing completely and viewers can now access TV content directly, on Internet platforms. At the same time, more traditional business models – advertising based or public funded – persist, offering to the audience an incredible amount of audiovisual contents, linear and on demand. Despite the convergent television-Internet-

telecommunications world that has been expected for decades, integration between platforms was initially very difficult. Relations between the TV ecosystem actors have been rough. Telecoms operators (see entry on ► [“Telecommunications”](#)) have tried to enter the television business many times, but it was not always easy and the first versions of Internet protocol television (IPTV) did not obtain their estimated success, in many countries. However, nowadays the so-called Quadruple players provide content across mobile, video, and broadband platforms (Ofcom [2017](#)). Content producers, distributors, and broadcasters, finally, compete today both with telecoms operators and with players coming from the web, such as the so-called Over-the-Top players (Google, Amazon, Netflix, Hulu, . . .), offering to the audience new forms of services and a huge amount of audiovisual contents.

Cross-References

- [Media](#)
- [Public Investments: Broadband](#)
- [State Aids and Subsidies](#)
- [Telecommunications](#)

References

- Adda J, Ottaviani M (2005) The transition to digital television. *Econ Policy* 20(41):160–209
- AVMSD 2010. Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services Charter of Fundamental Rights of the European Union (2012/C 326/02)
- Commission communication on the application of state aid rules to public service broadcasting [Official Journal C 257 of 27.10.2009]
- Commission Regulation N. 651/2014 of 17 June 2014 declaring certain categories of aid compatible with the internal market in application of Articles 107 and 108 of the Treaty
- Communication from the Commission on the application of State aid rules to public service broadcasting [Official Journal C 257 of 27.10.2009]
- Consolidated version of the Treaty on the Functioning of the European Union (2010/C 083/01)
- Council Directive 89/552/EEC of 3 October 1989 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the pursuit of television broadcasting activities
- European Audiovisual Observatory (Council of Europe) (2017) Audiovisual services in Europe – focus on services targeting other countries, Strasbourg
- European Convention for the Protection of Human Rights and Fundamental Freedoms, amended (1950)
- European Parliament resolution of 21 May 2013 on the EU Charter: standard settings for media freedom across the EU (2011/2246(INI))
- Freedman D (2008) The politics of media policy. Polity, Cambridge, UK
- Josephine MA-C (2015) Public service media remit in 40 European countries. IRIS bonus 2015-3. European Audiovisual Observatory, Strasbourg
- Ley 7/2010, de 31 de marzo, General de la Comunicación. Audiovisual. Jefatura del Estado. “BOE” núm. 79, de 1 de abril de 2010
- Ofcom (2016) The International Communications Market 2016
- Ofcom (2017) The International Communications Market 2017
- Owen BM, Wildman SS (1992) Video economics. Harvard University Press, Cambridge, MA. La Editorial, UPR
- Picard R (2002) The economics and financing of media companies. No. 1. Fordham University Press, New York
- Shapiro C, Varian HR (1998) Information rules: a strategic guide to the network economy. Harvard Business Press.

Terror

- [Terrorism](#)

Terrorism

- Friedrich Schneider¹ and Daniel Meierrieks²
¹Department of Economics, Johannes Kepler University of Linz, Linz-Auhof, Austria
²Department of Economics, University of Freiburg, Freiburg, Germany

Synonyms

[Political violence](#); [Terror](#)

Definition

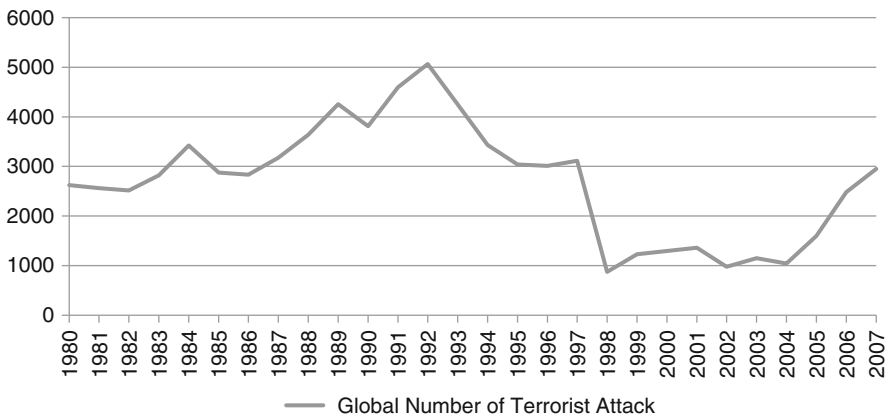
Terrorism is the premeditated use, or threat of use, of extra-normal violence by non-state outside the context of legitimate warfare with the intention to coerce and intimidate an audience larger than the immediate victims in order to obtain political, economic, religious, or other social objectives through intimidation or fear.

Introduction

The *Global Terrorism Database (GTD)* defines terrorism as any action by a non-state actor (usually, a terrorist organization) outside the context of legitimate warfare with the intention to communicate (through the use of violence) with, coerce, or intimidate an audience larger than the immediate victims of a terrorist act, where this act is associated with achieving political, economic, religious, or other social objectives. Especially the terrorist attacks on New York and Washington, D.C., on September 11, 2001, have sparked a renewed interest in the economic analysis of terrorism and counterterrorism.

Given the persistence of terrorism (illustrated by Fig. 1 which uses data from the *GTD* to illustrate the global patterns of terrorism over the past decades), we want to provide an overview of the academic literature on the *economics of terrorism and counterterrorism*. The following contribution

is a condensed version of Schneider et al. (2014) and uses material from it. In the section “[Rational-Choice Theory and the Economic Analysis of Terrorism](#)” of this contribution, we discuss the role of rational-choice theory in the economic analysis of terrorism and counterterrorism. Here, we argue that simple cost-benefit models using rational-choice representations of terrorist behavior provide a well-founded model for the study of terrorism. In the section “[Terrorism and Counterterrorism in a Rational-Choice Framework](#),” we review the fundamental strategies of counterterrorism that follow from the rational-choice framework. We discuss the implications for the design of appropriate counterterrorism efforts and the empirical evidence assessing the effectiveness of these efforts. Here, we also take stock of the literature on the causes and consequences of terrorism. However, we also hint at the limitations associated with these simple cost-benefit models, which especially relate to the failure of accounting for collective action problems linked to the phenomenon of transnational terrorism and the dynamic interaction between terrorism and counterterrorism. Thus, in the section “[Interaction Between Terrorism and Counterterrorism](#),” we discuss some of the consequences that result from the reaction of terrorists and other economic agents to distinct counterterrorism measures. The section “[Conclusion and Future Research](#)” concludes by discussing which counterterrorism strategies



Terrorism, Fig. 1 Global terrorist activity, 1980–2007

may ultimately prove most helpful in the fight against terrorism.

Rational-Choice Theory and the Economic Analysis of Terrorism

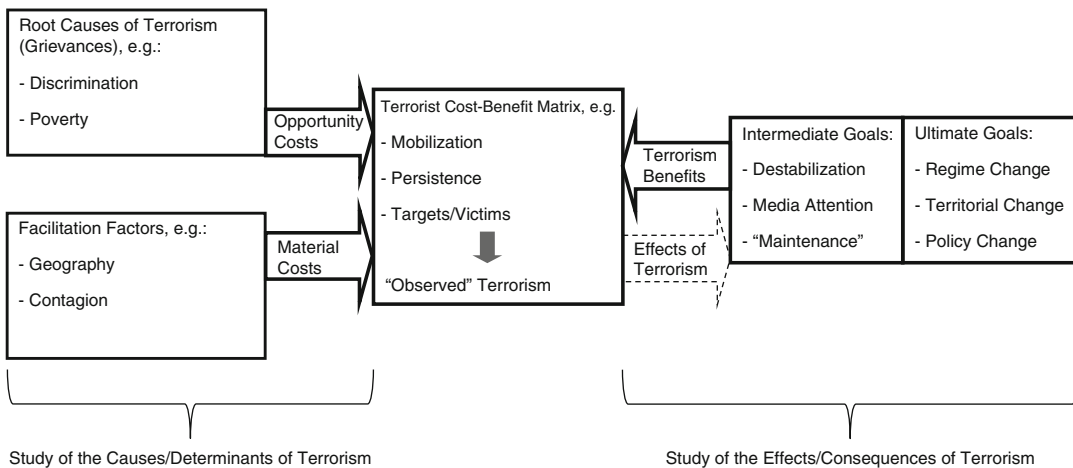
Rational-choice models are the theoretical workhorse of most economic analyses of terrorism (e.g., Landes 1978; Sandler and Enders 2004; Caplan 2006; Freytag et al. 2011). In short, in a rational-choice approach, terrorists are considered rational actors who choose the optimal (utility-maximizing) level of violence by considering the costs, benefits, and opportunity costs of terrorism, where the utility from terrorism is usually associated with achieving certain (political) goals (Sandler and Enders 2004).

The existence of a calculus of (rational) terrorists allows for an economic analysis of terrorism. For one, it informs empirical studies on the *determinants of terrorism*. For another, it informs empirical studies on the *consequences of terrorism*. Figure 2 shows the relationship between rational-choice theory and the study of terrorism and counterterrorism.

Terrorists as Rational Actors

As argued by Krueger and Maleckova (2003), rational-choice models of terrorism can be considered as an extension of models that economically

analyze criminal behavior. Economic models of crime (Becker 1968) suggest that criminals are rational actors who maximize their utility subject to a calculus that involves the costs of criminal activity (e.g., from the risk of punishment), its benefits (e.g., the “wage” a crime pays), and its opportunity costs (e.g., foregone earnings from noncriminal activity). Analogous to this, rational-choice models of terrorism assume terrorists to be rational actors – behaving more or less like *homi economici* – that try to maximize their utility, given the benefits, (opportunity) costs, and economic constraints linked to these actions. While public perception tends to view terrorist behavior as “irrational,” psychological studies of terrorist behavior provide little evidence that terrorists routinely suffer from mental incapacities (for a review, see Victoroff (2005)). As a matter of fact, Caplan (2006) provides an extensive analysis of terrorist irrationality and comes to the conclusion that the sympathizers of terrorism and most active terrorists act more or less rational, so that “the rational choice model of terrorism is not that far from the truth [and] the Beckerian analysis of crime remains useful” (Caplan 2006, p. 101). This may even extend to the (rare) case of *suicide terrorism*. While Caplan (2006) argues that suicide terrorists indeed typically violate the rationality assumption of the rational-choice model and thus should be considered as outliers, Pape (2003) argues that even



Terrorism, Fig. 2 Rational-choice theory and the study of terrorism

suicide terrorism is the result of strategic-rational decision-making.

The Terrorists' Calculus

The Direct Costs Associated with Perpetrating Terrorism

The direct or material costs of terrorism are one element of the terrorists' calculus. Usually, these costs are associated with the operations of a terrorist organization, i.e., offensive and defensive activities. For instance, they accrue from acquiring financial resources, purchasing firearms and explosives, and establishing safe houses to evade government punishment.

The Benefits of Terrorism

Benefits of terrorism are connected with the tactical and strategic goals of terrorism. As argued by Schelling (1991), the short-run (tactical) goals of terrorism include *politico-economic destabilization* (to weaken their enemy, i.e., the state) and *media attention* (to communicate the terrorists' cause). Thus, achieving these tactical goals ought to provide the member of a terrorist group with benefits. The long-run goals of terrorist groups usually involve the wish to induce political, economic, religious, or social change, most prominently (for separatist terrorist groups) gaining independence and (for religious and left-wing or right-wing terrorist groups) changing a country's politico-economic system to have it coincide with religious doctrine or political ideology. These political objectives are discussed in more detail in Shughart (2006). Any success related to achieving these long-run goals (through government concessions, territorial gains, winning political influence, etc.) also ought to constitute a benefit from terrorist activity.

The Opportunity Costs of Terrorism

The opportunity costs of terrorism refer to the foregone utility associated with non-terrorist activity. Typically, this utility comes from material rewards (i.e., wages) linked to nonviolence, e.g., participation in the ordinary economic life. It can also be understood as a monetary equivalent (derived from economic activity) of the potential

political influence associated with terrorist activity. Such lines of reasoning are at the center of many contributions that try to economically model terrorist behavior (e.g., Sandler and Enders 2004; Freytag et al. 2011).

Terrorism and Counterterrorism in a Rational-Choice Framework

The identification of an economic calculus associated with terrorist behavior allows for *three fundamental strategies* to reduce terrorist activity. Ceteris paribus, the utility-maximizing level of terrorism (i.e., the observed level of violence) ought to be lower when (1) the costs of terrorism are increased, (2) the benefits of terrorism are reduced, or (3) the opportunity costs of terrorism are raised. We discuss these options below in more detail.

Raising the Costs of Terrorism

Any counterterrorism policy that aims at raising the material costs of terrorism (e.g., by raising the penalty for terrorist offenses, direct police or military efforts) ought to make it more difficult for terrorist groups to maintain their level of activity.

The prevailing counterterrorism strategy to raise the direct costs of terrorism is *punishment and deterrence*. This strategy involves the use of direct state action by the police, military forces, and intelligence agencies to capture active terrorists and their supporters, while also deterring potential recruits (due to long prison sentences, increased probability of being captured, etc.). Indeed, counterterrorism activities by the police and intelligence services (e.g., infiltration through the use of informers and undercover agents, observation, information gathering) have repeatedly weakened the operative capacity of terrorist organizations (for an overview, see Schneider et al. (2014)). Also, empirical evidence suggests that increased punishment for specific kinds of terrorist activity leads to fewer of these acts (e.g., for the case of skyjackings, see Landes (1978)). As for a more harsh direct counterterrorism strategy, efforts in the form of decapitation – i.e., the killing of terrorist leaders – have also been shown

to prove helpful against terrorist groups (e.g., Johnston 2012).

Reducing the Benefits of Terrorism

Previously, we have established that the benefits of terrorism are directly linked to the short-run and long-run objectives of terrorist organizations. In the short run, the benefits from terrorism arise from politico-economic destabilization and media attention, both of which help the terrorists to achieve their long-run sociopolitical objectives. Indeed, many empirical studies have established that terrorism tends to negatively affect political development (e.g., in the form of respect for human rights) and political stability (e.g., Dreher et al. 2010). There is also evidence that terrorism reduces economic activity, e.g., by depressing domestic and foreign investment or tourism (e.g., Crain and Crain 2006; for a review, see Sandler and Enders (2008)).

Counterterrorism efforts may be effective when they successfully deny terrorist groups media attention and increase a country's politico-economic resiliency to terrorism's destabilizing effects, which ultimately ought to mean that political concessions associated with the long-run goals of terrorist groups become less likely. This is because less media attention and less success in producing politico-economic damage are expected to negatively affect a terrorist organization's bargaining position vis-à-vis the government it opposes (Schneider et al. 2014).

One interesting strategy to increase politico-economic robustness is *decentralization*. Political decentralization ought to make it less likely that terrorism creates a political vacuum (e.g., when prominent political figures are assassinated) that cannot be filled by other actors or levels of government (Frey and Luechinger 2004). Economic decentralization is expected to have a similar effect on terrorism. It may include avoiding concentrating power within a company in the hands of few individuals (because these individuals are attractive targets for attacks), not concentrating a company in one large headquarters (particularly when this headquarter is located in an iconic building such as the World Trade Center), and using multiple (rather than monopolistic) suppliers to immunize

a company's supply chain against disruptions (Frey and Luechinger 2004). Indeed, Dreher and Fischer (2010) find that decentralization may reduce the likelihood of terrorism.

Other means to reduce the benefits of terrorism may include, e.g., the increased protection and fortification of prospective terrorist targets and denying the terrorists media attention by means of spreading disinformation (Schneider et al. 2014).

Influencing the Opportunity Costs of Terrorism

While raising the direct/material costs of terrorism usually involves manipulating related *facilitation factors* (e.g., by restricting access to firearms and explosives), influencing the opportunity costs of terrorism typically involves policies that try to ameliorate *grievances* that underlie social conflict.

Among the factors possibly involving grievances and thereby inciting terrorism that are analyzed in cross-national studies are (1) *poor socioeconomic conditions* (e.g., Freytag et al. 2011); (2) *unfavorable political and economic institutions* which, e.g., do not sufficiently protect civil liberties and do not enable politico-economic participation (e.g., Krueger and Maleckova 2003; Abadie 2006); (3) *demographic stress*, e.g., in the form of discrimination along ethno-religious lines (e.g., Piazza 2011); (4) trends in *globalization* that may generate resentment among the "globalization losers" and traditionalist who fear that the inflow of foreign ideas threatens their local culture or religion (Zimmermann 2011); and (5) *aggressive foreign policy behavior* – in the form of, e.g., military interventions, state sponsorship of terrorism, and overall politico-military dominance – which tends to coincide with more terrorist activity directed against the proponents of such policies (e.g., Pape 2003; Dreher and Gassebner 2008). Gassebner and Luechinger (2011) and Krieger and Meierrieks (2011) summarize the empirical literature on the causes of terrorism.

Ameliorating the aforementioned social conditions ought to reduce terrorism by swaying its opportunity costs in ways that make violence

a less attractive option. However, according to the reviews by Gassebner and Luechinger (2011) and Krieger and Meierrieks (2011), there is little consensus on the importance of specific social conditions in the emergence of terrorism. For instance, while some cross-country studies find that poor socioeconomic conditions predict terrorism (e.g., Freytag et al. 2011), other studies find that political variables are more important than economic ones (e.g., Krueger and Maleckova 2003; Abadie 2006). Thus, there seems to be no obvious “policy panacea” to fight terrorism by favorably affecting its opportunity costs. What is more, influencing the social conditions that underlie terrorist conflict in favorable ways tends to be a long-run and complex counterterrorism option, particularly in comparison to options that try to affect the direct costs of terrorism.

The Case of Transnational Terrorism

Counterterrorism efforts may be further complicated when terrorism becomes transnational. By definition, this is the case when more than one country is involved in a terrorist conflict. For instance, terrorism may internationalize when terrorist groups use foreign territory to challenge another government (e.g., the use of Syria, Jordan, Lebanon, Tunisia, and other Middle Eastern countries as bases of operation for Palestinian terrorist activity against Israel in the 1970s and 1980s).

International counterterrorism strategies still aim at influencing the (opportunity) costs and benefits of terrorism in ways that make (internationalized) terrorism less attractive. However, it is precisely their transnational nature that makes these strategies even more difficult to implement. Sandler (2005) shows that internationalized terrorism usually involves *collective action failures* for countries attacked by transnational terrorist groups. While governments are able to choose an optimal counterterrorism strategy in the fight against domestic terrorism, counterterrorism against transnational terrorism may involve externalities. For instance, it may be attractive for some – often weak – states to tolerate the activities of terrorist organizations within their borders in exchange for no direct harm at the expense of other nations (*paid riding*).

Also, national governments may be tempted to overstress defensive counterterrorism measures to deter terrorist attacks, which may merely result in a relocation of terrorist attacks against the respective nation’s citizens (e.g., there were more anti-US attacks in the Middle East and Africa after 9/11 had led to stronger counterterrorism efforts within the United States). Alternatively, as suggested by Sandler (2005), countries may try to place the burden of offensive counterterrorism measures (e.g., preemptive strikes) onto the prime target of transnational terrorism, thereby benefitting from the reduction of terrorism without paying for it (*free riding*). More generally, defensive policies can be regarded as largely private goods, so that the benefits of security provision are mostly internalized by the investors, while proactive policies show the characteristics of public goods (Sandler and Siqueira 2006). Consequently, this may lead to an oversupply of defensive and an undersupply of proactive measures (Sandler and Siqueira 2006).

Collective action problems (e.g., free and paid riding, benefits of noncompliance) have so far undermined many measures directed against terrorism at an international level. For instance, Cauley and Im (1988) find that the introduction of a United Nation’s convention on preventing crime against diplomatic personnel did not help to reduce terrorism directed against diplomats. Similar problems arise when international counterterrorism efforts aim at curtailing terrorism by means of military actions (e.g., the US air strikes against Libya in 1988 in retaliation of alleged Libyan support for anti-US terrorism, which in turn led Libya to sponsor anti-American terrorism) and the use of economic and/or military aid (Schneider et al. 2014).

Interaction Between Terrorism and Counterterrorism

Previously, we have argued that rational-choice theory helps to understand how terrorism develops and how it can be countered through three fundamental strategies that aim at influencing the (opportunity) costs and benefits of terrorism in ways that make terrorism less attractive. One

shortcoming of this approach, however, is its *static* nature. In reality, terrorism and counterterrorism tend to interact. In this section, we therefore give a brief – nontechnical – summary of typical interactions between terrorism and counterterrorism. Some of these points are also discussed in Kydd and Walter (2002) and Findley and Young (2012).

Adaption, Innovation, and Substitution

Terrorist organizations may be able to adapt to specific counterterrorism measures by means of innovation and/or substitution. The former may include, e.g., the use of more advanced terrorist “technology” to attack (e.g., the use of more powerful weapons) or organizational innovation as a hierarchical group develops into a looser terrorist network with a cell structure. The latter may involve, e.g., choosing new targets that are less protected by counterterrorism measures. Arguably, substitution effects are standard elements of economic models, so that it is not surprising that such effects also matter to the economic study of terrorism and counterterrorism (Enders and Sandler 2004). As a consequence, the evidence suggests that some counterterrorism measures may merely change the face of terrorism (e.g., new attack methods, new targets) but not affect the overall level of terrorist violence (Schneider et al. 2014).

Provocation and Escalation

Findley and Young (2012) argue that terrorist groups may use terrorism to provoke a harsh, disproportionate response by the government (e.g., by means of excessive police force), which in turn is expected to fuel radicalization and social polarization, meaning that provocation and excessive counterterrorism measures by the state may easily result in cycles of violence. Indeed, it seems to be in the natural interest of terrorist groups to muster additional support by letting any social conflict escalate, given that broader conflicts (e.g., civil wars) make it easier for them to pursue and implement their agendas (Findley and Young 2012).

Spoiling the Peace

As a specific reaction to government concessions, terrorist groups may try to spoil the peace. For

instance, terrorist groups may have economic incentives (e.g., gains from illegal activity) associated with conflict and thereby oppose peace. Thus, it may be in the natural interest of extremist factions to sabotage the peace (Kydd and Walter 2002). Crucially, such sabotage may result in – as it is in the interest of the attacking groups – an end of negotiations and concessions and, instead, provoke more violent counterterrorism measures. Kydd and Walter (2002) argue that spoiling the peace – as a reaction to benevolent policy measures such as concessions – may be particularly effective when mistrust and weakness among the more moderate negotiator on the sides of the terrorists and the state abound.

Organizational Evolution

We have already discussed above that terrorist groups may innovate in the face of counterterrorism by changing their internal organization (e.g., moving from a hierarchical to a cell structure). However, counterterrorism may also result in changing the overall politico-military orientation of a terrorist group. The nature of these changes depends on the offered incentives and the effectiveness of counterterrorism. For one, positive incentives and relative counterterrorism effectiveness may lead to a situation where a group gives up its armed struggle, evolving into a political party that tries to foster change nonviolently. One example is the Colombian *Movimiento 19 de Abril*. For another, however, negative incentives and/or relative counterterrorism ineffectiveness may lead to an escalation of conflict, so that terrorist campaigns may evolve into full-scale insurgencies (Findley and Young 2012). As another possibility, perhaps in frustration over the inability to achieve certain political goals due to counterterrorism, terrorist groups may become increasingly interested in purely financial gains, evolving into criminal groups.

Conclusion and Future Research

In this entry, we provided an overview of theoretical and empirical studies on the economic

analysis of terrorism and counterterrorism. We argued that rational-choice models of terrorist behavior provide a good starting point for such analyses. Here, simple cost-benefit models that follow from rational-choice approaches imply that terrorism can be fought by affecting the terrorists' calculus which involves the (opportunity) costs and benefits of terrorism. We discussed a number of specific counterterrorism strategies (e.g., direct police or military action, decentralization) that follow from such models. One factor that complicates these models – and the strategies derived from them – are collective action problems that arise when terrorism becomes transnational (the most infamous example being the 9/11 attacks). Another complicating factor associated with these simple cost-benefit models is that they are usually static and thus miss the dynamic nature of the terrorism-counterterrorism relationship. We (non-exhaustively) discussed a number of reactions of terrorist groups to specific counterterrorism strategies (e.g., the use of innovative modes of attack in the light of target-hardening efforts). These interactions show that the relationship between terrorism and counterterrorism tends to be more complex than simple cost-benefit models – even though they provide a good starting point for any economic analysis of terrorism – can capture.

Considering the most effective counterterrorism strategy, our entry suggests the following: (1) Strategies that involve influencing the benefits of terrorism (e.g., target protection) usually do not stop terrorism but lead to adaption (new targets, new methods, etc.). (2) Raising the direct/material costs of active terrorists (e.g., through military means) may prove effective in the short run. Indeed, many terrorist groups have been negatively affected by military pressure. However, such efforts may easily backfire, creating (potentially large) unintended consequences. Here, (relative) counterterrorism ineffectiveness may contribute to the emergence of powerful insurgencies or crime networks, i.e., to the evolution rather than decline of terrorist groups. For instance, the French counterterrorism efforts during the Battle of Algiers of 1957 against the Algerian *Front de Libération*

Nationale (FLN) were a military success but – due to use of torture and other harsh counterterrorism measures – led to the international political isolation of France and growing (domestic and international) support for the *FLN*, while also triggering a political crises in France. (3) Strategies that aim at raising the (opportunity) costs of terrorism by “winning the hearts and minds” of would-be terrorists, potential terrorism supporters, and possibly even active terrorists seem to be more effective in the long run. For instance, terrorist groups may disappear due to a loss of legitimacy – e.g., as popular support for terrorism dwindles due to higher terrorism opportunity costs in the face of benevolent counterterrorism strategies – or evolve into political parties when political participation opportunities open up.

Clearly, the effectiveness of these strategies is very much context-dependent. Counterterrorism strategies that worked well against terrorism in the twentieth century (which usually had domestic goals, was hierarchically structured, and attacked primarily military targets) may prove unsuccessful against the *Al-Qaeda*-styled terrorism of the twenty-first century (which has transnational goals, is organized as a network, and also attacks civilian targets). This provides many avenues for future research. For instance, this research may try to evaluate the appropriateness of counterterrorism strategies vis-à-vis, e.g., the goals of terrorist organizations (to assess which kinds of concessions can be made), the implementation costs of specific policies (to examine their cost-efficiency), the international perspective (to factor in collective action problems), the overall level of popular support for terrorism, and the organizational structure and politico-military strategy of a terrorist group. These factors have been largely disregarded in the theoretical and empirical literature on terrorism and counterterrorism.

Cross-References

- [Crime: Organized Crime and the Law](#)
- [Violence, Conflict-Related](#)

References

- Abadie A (2006) Poverty, political freedom, and the roots of terrorism. *Am Econ Rev* 96:50–56
- Becker G (1968) Crime and punishment: an economic approach. *J Polit Econ* 76:169–217
- Caplan B (2006) Terrorism: the relevance of the rational choice model. *Public Choice* 128:91–107
- Cauley J, Im EI (1988) Intervention policy analysis of skyjackings and other terrorist incidents. *Am Econ Rev* 78:27–31
- Crain NV, Crain WM (2006) Terrorized economies. *Public Choice* 128:317–349
- Dreher A, Fischer JAV (2010) Decentralization as a disincentive for transnational terror? An empirical test. *Int Econ Rev* 51:981–1002
- Dreher A, Gassebner M (2008) Does political proximity to the U.S. cause terror? *Econ Lett* 99:27–29
- Dreher A, Gassebner M, Siemers LH-R (2010) Does terror threaten human rights? Evidence from panel data. *J Law Econ* 53:65–93
- Enders W, Sandler T (2004) What Do We Know About the Substitution Effect in Transnational Terrorism? In Silke A (ed) *Research on Terrorism: Trends, Achievements, and Failures*. London, Frank Cass, pp. 119–137
- Findley MG, Young JK (2012) Terrorism and civil war: a spatial and temporal approach to a conceptual problem. *Perspect Polit* 10:285–305
- Frey BS, Luechinger S (2004) Decentralization as a disincentive for terror. *Eur J Polit Econ* 20:509–515
- Freytag A, Krüger JJ, Meierrieks D, Schneider F (2011) The origins of terrorism: cross-country estimates of socio-economic determinants of terrorism. *Eur J Polit Econ* 27:5–16
- Gassebner M, Luechinger S (2011) Lock, stock, and barrel: a comprehensive assessment of the determinants of terror. *Public Choice* 149:235–261
- Johnston PB (2012) Does decapitation work? Assessing the effectiveness of leadership targeting in counterinsurgency campaigns. *Int Secur* 36:47–79
- Krieger T, Meierrieks D (2011) What causes terrorism? *Public Choice* 147:3–27
- Krueger AB, Maleckova J (2003) Education, poverty and terrorism: is there a causal connection? *J Econ Perspect* 17:119–144
- Kydd A, Walter BF (2002) Sabotaging the peace: the politics of extremist violence. *Int Organ* 56:263–296
- Landes WM (1978) An economic study of US aircraft skyjackings, 1961–1976. *J Law Econ* 21:1–31
- Pape RA (2003) The strategic logic of suicide terrorism. *Am Polit Sci Rev* 97:341–361
- Piazza JA (2011) Poverty, minority economic discrimination, and domestic terrorism. *J Peace Res* 48:339–353
- Sandler T (2005) Collective versus unilateral responses to terrorism. *Public Choice* 124:75–93
- Sandler T, Enders W (2004) An economic perspective on transnational terrorism. *Eur J Polit Econ* 20:301–316
- Sandler T, Enders W (2008) Economic consequences of terrorism in developed and developing countries: an overview. In: Keefer P, Loayza N (eds) *Terrorism, economic development, and political openness*. Cambridge University Press, Cambridge, pp 17–47
- Sandler T, Siqueira K (2006) Global terrorism: deterrence versus pre-emption. *Can J Econ* 39:1370–1387
- Schelling TC (1991) What purposes can “international terrorism” serve? In: Frey RG, Morris CC (eds) *Violence, terrorism, and justice*. Cambridge University Press, Cambridge, pp 18–32
- Schneider F, Brück T, Meierrieks D (2014) The economics of counter-terrorism: a survey. *J Econ Surv* (forthcoming), <https://doi.org/10.1111/joes.12060>
- Shughart WF (2006) An analytical history of terrorism, 1945–2000. *Public Choice* 128:7–39
- Victoroff J (2005) The mind of the terrorist: a review and critique of psychological approaches. *J Confl Resolut* 49:3–42
- Zimmermann E (2011) Globalization and terrorism. *Eur J Polit Econ* 27:152–161

Further Reading

- Cronin AW (2011) *How terrorism ends: understanding the decline and demise of terrorist campaigns*. Princeton University Press, Princeton
- Enders W, Sandler T (2011) *The political economy of terrorism*. Cambridge University Press, New York
- Hoffman B (2006) *Inside terrorism*. Columbia University Press, New York
- Walker C (2011) *Terrorism and the law*. Oxford University Press, Oxford

Third-Party Financed Litigation

- [Third-Party Litigation Funding](#)

Third-Party Financing of Litigation

- [Third-Party Litigation Funding](#)

Third-Party Litigation Funding

Myriam Doriat-Duban
Department of Economics, BETA UMR 7522,
Université de Lorraine, Nancy, France

Synonyms

[Third-party financed litigation](#); [Third-party financing of litigation](#)

Abstract

Third-party litigation funding allows the transfer by the plaintiff of his/her legal costs to a financial company whose sole aim is to make profit. TPLF have an impact on access to justice but also on conflict resolution and finally on social well-being. The aim is to show how law and economics scholars invest this new field of the conflict literature and study the role of this new actor in the litigation.

Definition

Third-party litigation funding (TPLF) is a financial activity that consists in an investment company specialized in the financing of litigation agreeing to cover all or part of the trial costs of a litigant, in exchange for a portion of the damages paid if the case is won. The remuneration of the funding third party can be a multiple of the initial investment (1.5–6 times), a percentage of the award obtained in a judgment or out-of-court settlement (20–40%, sometimes 50%), or a combination of both (Veljanovski 2012). The funds used are provided by wealthy individuals, families or institutions, insurance companies, or hedge funds (de Silguy 2013; Veljanovski 2012). The return/risk ratios can reach very high levels, up to 200% of the amounts invested (McLaughlin 2007), sometimes more (Abrams and Chen 2013). There are several types of TPLF: interest-bearing loans intended to provide replacement income to victims of bodily injuries or discriminatory practices (consumer legal funding), loans to firms of lawyers in a contingent fees system, and investment in commercial litigation in exchange for a percentage of the damages (Faure and De Mot 2012).

Introduction

Historically, TPLF first appeared in the Middle Ages, but the prohibition of “maintenance” and “champerty” in the United States prevented it developing (Lyon 2010). It reappeared in Australia at the beginning of the 1990s, before

spreading to other common law countries. Its development in countries with a civil law tradition came later and was more modest, with the exception of Germany and, to a lesser extent, Austria and Switzerland. TPLF concerns both litigation before the courts and arbitration procedures (in international trade in particular).

There are many arguments in favor of developing TPLF. The most immediate advantage is the removal of the financial barrier to the plaintiff’s access to justice (Jackson 2009). Other arguments relate to the easier access to third-party capital (Daughety and Reinganum 2014); better allocation of funds for companies, which can invest in their business instead of devoting funds to their defense (Rubin 2011; Veljanovski 2012); and a better division of the risks (De Mompurgo 2011) between a risk-averse plaintiff and a neutral third-party funder who can pool the risks in a portfolio of claims not correlated with each other (Rubin 2011).

The economic analysis of TPLF means studying the impact of having litigation funded by a third party on the one hand, litigants’ access to justice on the other, the method of resolving disputes, and finally the impact on social well-being.

TPLF and Access to Justice

From a theoretical point of view, TPLF is based on the idea that the right to sue is exchangeable according to the Coase theorem: by injecting funds into the case, the funder purchases the right to receive all or part of the recovery. By doing so, the funder facilitates access to justice by removing the financial constraint on the plaintiff, as do other schemes (legal aid, legal expenses insurance, contingent fees), but only in lawsuits where the stakes are high (De Fontmichel 2012) and mainly, in Europe, in the field of international arbitration. In a model inspired by Katz (1990), Deffains and Desrieux (2014) have demonstrated, however, that access to justice is not guaranteed even for meritorious claims. Indeed, the third-party funder will invest in a case if, first of all, he considers that it is sufficiently profitable once all the transaction costs have been taken into account (obligation to sign two contracts, one between the lawyer and the client and another between the third part funder and the client, cost

of negotiating the percentage of the damages or the interest rate, costs of obtaining experts' opinions, dispute solving costs, etc.) and, secondly, if it is more profitable than alternative investments. They also show that where there is asymmetry in the information on the real quality of the case, TPLF can encourage plaintiffs bringing frivolous lawsuits to bring more actions than in a no-win, no-fee arrangement or if they are self-financing. This outcome is due to the fact that the equilibrium settlement amount is lower with TPLF and therefore the likelihood of reaching an arrangement with the defendant is higher.

As well as facilitating access to justice, TPLF guarantees a better defense of plaintiffs' interests. Indeed, TPLF will orient the plaintiff toward better lawyers, whom he will be also be able to better control (Schanzenbach and Dana 2009). However, a problem of agency can arise between the third-party funder and the lawyer. Demougin and Maultzsch (2014) nevertheless show that it is possible to arrive at a combination of third-party funding and a no-win, no-fee arrangement that can overcome both these agency issues and the access to justice difficulties of potential plaintiffs with "deserving" claims.

Impact of TPLF on Methods of Conflict Resolution

TPLF has an impact on the method of conflict resolution because it increases the plaintiff's bargaining power: with the benefit of substantial financial resources thanks to the support of the third-party funder, his threat of taking legal action gains in credibility (De Mompurgo 2011). TPLF is therefore thought to be a way of enabling small- and medium-sized enterprises to fight large international disputes on an equal footing (De Fontmichel 2012).

Furthermore, TPLF breaks the monopsony power of the defendant: the third-party funder, along with the defendant who wants to come to an arrangement, can be considered as buyers of the plaintiff's right to sue (Hylton 2014). With the resources that he devotes to raising the value of the right he has thus acquired, the third-party funder enables the plaintiff to obtain a better award (Avraham and Wickelgren 2014) and

ensures a higher likelihood of judgments being enforced (De Fontmichel 2012).

The credibility of the plaintiff's threat of legal action is also reinforced by the credibility of the signal concerning the quality of the case, due to the in-depth investigations (defendant's solvency, chances of success, quality of the lawyers) carried out by the third-party funder in order to minimize his risks (Faure and De Mot 2012). For Avraham and Wickelgren (2014), it is in the third-party funder's interests to send a clear signal to the defendant and to the judge about the quality of the case he is funding, via the interest rate applied to the plaintiff.

Ultimately, few claims will be selected (those with more than a 70% likelihood of being won, according to Veljanovski 2012), so that the third-party funder's actual risk is zero. Daughety and Reinganum (2014) confirm this in the case of TPLF where the third party provides the plaintiff with funds to cover his day-to-day expenses. The authors have produced a signal model where the plaintiff has private information on the real value of his case that determines its type. They show that the optimal interest rate, understood as that which maximizes the joint payment of the plaintiff and the funder, always leads to an out-of-court settlement of the dispute, whatever the type of plaintiff. Kirstein and Rickman (2004) arrive at the same result in a configuration where there is no information asymmetry but only diverging estimates by the parties of the plaintiff's chances of winning the court case.

TPLF and the Improvement of Social Well-Being

By facilitating access to justice, TPLF generates positive externalities that tend to increase social well-being. Thus, Hylton (2012) explains that if socially desirable legal action is not taken because it is not cost-effective, then TPLF is socially desirable because the transfer of costs that it allows makes legal action cost-effective for the plaintiff from a private point of view. This implies that there are an optimum number of litigation cases. Any analysis of the social desirability of TPLF should therefore start by determining whether the number of litigation

proceedings brought is lower or higher than the socially optimal number.

Other positive external effects are also discussed in the literature. For example, TPLF is said to increase competition between lawyers (Veljanovski 2012; Costargent 2012) but also between large law firms on the one hand and third-party funders on the other (De Fontmichel 2012). Finally, according to De Morpurgo (2011), by facilitating access to justice, TPLF is said to increase equality of litigants before the law, which would therefore have the consequence of increasing the usefulness of all individuals with a concern for justice and therefore also social well-being.

In addition to these positive external effects, there are also negative external effects, which render the conclusions on the social desirability of TPLF theoretically undetermined. It is quite possible, for example, that the drop in the number of expected damages cases will not last. Hylton (2014) is interested not in existing, matured claims but in potential future, unmatured claims. The sale of *ex ante* rights is a complex issue because inefficient rights transfers can lead to an increase in the number of damages cases for potential victims. It all depends on the third party's intentions. If what he wants to achieve is the payment of damages, his interests are in line with those of the potential victim, and social well-being is therefore potentially increased. However, if his aim is to avoid proceedings by purchasing his right to reparation from the victim (interests in line with those of the defendant), then there will be fewer precautions taken and therefore an increase in the frequency of damages cases.

Rubin (2011), however, worries about the defendant's situation. The defendant's participation in the conflict is determined by the plaintiff's decision to take legal action. The American rule on the allocation of litigation costs creates an externality for the defendant, who necessarily has to bear his litigation costs. In the case of TPLF, in particular in commercial lawsuits, those litigation costs are usually high. Any measure facilitating plaintiffs' access to justice must take account of the costs borne by the opposing party. In the end, better access to justice is socially desirable if the expected benefits are higher than

all the litigation costs, as for it the damages paid are a simple transfer.

Conclusion

The question of the socially desirable nature of TPLF runs through all the articles on this form of funding access to justice. The fears most often expressed concern the risk of encouraging socially undesirable litigation; economic analysis of conflicts shows that this appears to be unfounded due to third-party funders' very strict selection of the cases they take on. Furthermore, there can be other positive effects, in particular in terms of damage prevention. However, these are counterbalanced by the risk of congestion of the courts, the development of a compensation culture rather than a culture of recognizing rights and the risks that the intervention of a third party can bring to bear on the lawyer-client relationship.

It has to be admitted that the theoretical conclusions do not provide a definitive answer to the question of the social desirability of this type of funding. Empirical studies could in this case provide some assistance with decision-making. There is only one empirical article on the subject, concerning Australia. Abrams and Chen (2013) demonstrate that third-party funding of litigation increases the frequency of legal action being brought and the number of court cases. They ascertain in particular that in the Australian States where third-party litigation funding is common, the courts are very congested, the number of cases is completed, and the elucidation rate are lower, which also has consequences for the courts' spending. And yet, the authors explain that these effects may be temporary and the conclusions in terms of social well-being ambiguous (increase in the rate of out-of-court settlements, faster recognition of rights, clarification of the law). These conclusions cannot, however, be generalized to all countries.

Cross-References

- [Legal Aid](#)
- [Litigation and Legal Expenses Insurance](#)

References

- Abrams DS, Chen DL (2013) A market for justice: a first empirical look at third party litigation funding. *Univ PA J Bus* 14(7):1075–1109. <https://doi.org/10.2139/ssrn.2404483>
- Avraham R, Wickelgren A (2014) Third-party litigation funding – a signaling model. *De Paul Law Rev* 63:233–263. <https://doi.org/10.2139/ssrn.2302801>
- Costargent J-R (2012) Le financement par un tiers comme réponse aux évolutions de l'arbitrage international, avec Guy Lepage, Versailles International Arbitration and Business Law Review. Disponible à l'adresse: http://www.lafrancaise-icfund.lu/fileadmin/redacteurs/docs/articles/article_financement_par_tiers_J-C_Costa_rgent.pdf
- Daughety AF, Reinganum JR (2014) The effect of third-party funding of plaintiffs on settlement. *Am Econ Rev* 104(8):2552–2566. <https://doi.org/10.2139/ssrn.2197526>
- Deffains B, Desrieux C (2014) To litigate or not to litigate? The impacts of third-party financing on litigation. *Int Rev Law Econ*. <https://doi.org/10.1016/j.irle.2014.08.005>. Available online: <https://doi.org/10.1016/j.irle.2014.08.005>
- De Fontmichel M (2012) Les sociétés de financement de procès dans le paysage juridique français. *Revue des sociétés* 8:279
- De Morpurgo M (2011) A comparative legal and economic approach to third-party litigation funding. *Cardozo J IntL Comp Law* 19:343–412
- Demougins D, Maultzsch F (2014) Third-party financing of litigation: legal approaches and a formal model. *CESifo Econ Stud* 60:525–553. <https://doi.org/10.1093/cesifo/ift006>
- De Silguy S (2013) Du financement de process par des hedge funds. *Revue Lamy de Droit Civil (RLDC)* 105:66–68
- Faure MG, De Mot JPB (2012) Comparing third party financing of litigation and legal expenses insurance. *J Law Econ Policy* 8(3):743–778
- Hylton KN (2012) The economics of third-party financed litigation. *J Law Econ Policy* 8:701. <https://doi.org/10.2139/ssrn.1971229>
- Hylton KN (2014) Toward a regulatory framework for third-party funding of litigation. *De Paul Law Rev* 63(2):527–546. <https://doi.org/10.2139/ssrn.2281453>
- Jackson J (2009) Review of civil litigation costs, final report, published with the permission of the Ministry of Justice on behalf of the controller of her Majesty's Stationery Office. Disponible online: www.judiciary.gov.uk/Resources/JCO/.../jackson-final-report-140110.pdf
- Katz A (1990) The effect of frivolous lawsuits on the settlement of litigation. *Int Rev Law Econ* 10:3–27. [https://doi.org/10.1016/0144-8188\(90\)90002-b](https://doi.org/10.1016/0144-8188(90)90002-b)
- Kirstein R, Rickman N (2004) “Third Party Contingency” contracts in settlement and litigation. *J Inst Theor Econ (JITE)* 160(4):555–575. <https://doi.org/10.2139/ssrn.427400>
- Lyon J (2010) Revolution in Progress: Third-Party Funding of American Litigation. *Ucla Law Review* 58:571–60. <https://doi.org/10.2139/ssrn.2034625>
- McLaughlin JH (2007) Litigation funding: charting a legal and ethical course, *Vanderbilt Law Review*, 31: 620–621
- Rubin PH (2011) Third-party financing of litigation. *Northern Kentucky Law Rev* 38:673–685
- Schanzenbach M, Dana D (2009) How would third party financing change the face of American tort litigation?, Third party financing of litigation roundtable, Searle Center, Northwestern University Law School. Available at http://www.law.northwestern.edu/searlecenter/papers/Schanzenbach_Agency%20Costs.pdf
- Veljanovski C (2012) Third-party litigation funding in Europe. *J Law Econ Policy* 8(3):405–449

Titling Systems

Benito Arruñada

Pompeu Fabra University and BGSE, Barcelona, Spain

Definition

Titling systems are the institutions used to enforce property rights as rights in rem and reduce the cost of transacting on them. To be effective in non-local markets, they require a registry, which produces information on claims or rights, thus allowing the judge to verify them, establish their relative priority, and solve conflicts between claimholders by adjudicating rights in rem and in personam to them. Since the judge relies on register evidence, access to registers also allows contractual parties to reduce their information asymmetry before transacting.

Introduction: The Tradeoff between Property Enforcement and Transaction Costs

Rights to land and many other assets can be enforced as property rights, *iura in rem*, claimable against the asset itself and therefore valid against all persons, *erga omnes*. These property rights are said to “run with the land,” meaning that they

survive unaltered through all kinds of transactions, and transformations dealing with other rights on the same parcel of land or on a neighboring parcel. For example, the mortgagee keeps the same claim on the land even after the mortgagor sells it. Property rights oblige all people: the new owner who has purchased the land is obliged to respect both the mortgage and, in particular, the right to foreclose if the guaranteed debt is not paid. Enforcement of a property right is independent of who holds other rights on the same asset. Alternatively, rights on assets can be contract rights, enforceable against a specific person, *inter partes*. To clarify the difference between property and contract rights, consider what happens in the case of a lease of land, this being a right that in many jurisdictions may be structured as either a contract or a property right. Assume that the land is sold during the life of the lease. If the lease is a contract right, the lessee loses the right of occupation and gains instead a contract right against the lessor. However, if the lease is a property right, the lessee keeps the right of occupation. It is then the land purchaser who may have a contract right against the seller, if the sale was made free of leases. The buyer is subrogated into the seller's position. There is no change to the lease, which has run with the land from the seller to the buyer and survives intact after the sale.

When the law enforces a right as a right in rem, consent of the right holder is required for the right to be affected, that is, damaged, in any way. This requirement of consent – either real or constructive – provides precious enforcement benefits for rights on durable and immovable assets. The enforcement of contract rights, on the other hand, depends on the availability, resources, and legal status of persons, who are mobile and may become unavailable, or judgment proof when obliged to pay. For durable assets, a property right is therefore much more valuable than a contract right having the same content – that is, when the only difference between them is that the latter lacks in rem enforceability.

These enforcement benefits come at a cost, however. When multiple rights exist on an asset, transactions do not convey property rights with the promised in rem extent until all affected right

holders have consented. In other words, to produce perfect property – that is, in rem – rights, some kind of explicit or implicit contracting has to take place between the transactors and each of the affected right holders, in order for the latter to give their consent. Many institutions in the field of property law are designed to make these “contracts” with affected right holders possible. Consent can be given explicitly, by private agreement, declaring to a register or in court proceedings, as well as implicitly, simply by the passing of time. Consent can also be produced at the moment the transaction takes place. Consequently, the rights resulting from the transaction will be free of uncertainty as to who the true legal right holder is and as to their precise nature. Alternatively, consent can be postponed, and the transaction then produces rights which are burdened with the survival of any property rights whose right holders have not yet consented.

In any case, without the consent of affected right holders, transactions produce a mix of property and contract rights: property – that is, in rem – effects to the extent that is compatible with the surviving property rights held by others and contract rights for the difference. The proportion of property and contract rights in the mix varies with the kind of conflicting right. In the extreme case of a fraudulent conveyance, the grantee gets only a contract right against the grantor, who is not the true legal owner. More generally, any intended property right is in fact partially contractual if an affected right holder keeps a contradictory or concurrent right against it.

Property rights thus face a trade-off with positive and negative effects (Arruñada 2012). On the one hand, they facilitate specialization by ensuring enforcement, given that right holders' consent is required to affect them. However, for the same reason, their survival after conveyance of the asset or any other transformation of rights requires costly institutions (mainly, property registers) in order to organize the process of searching, bargaining, and contracting for consent. In particular, the possibility of hidden property rights increases the information asymmetry between the conveying parties: the seller knows better

than the acquirer about hidden property rights. More generally, the need to know which conflicting property rights exist and who their holders are, and bargaining with such right holders to obtain their consent and contracting or somehow formalizing an agreement with them, all increase the costs of transforming and conveying rights. This may in turn hamper investment, trade, and specialization.

Private Titling: Privacy of Claims as the Starting Point

Under the Roman law tradition of private conveyance that was dominant in Europe until the nineteenth century, private contracts on land had in rem effects on third parties, even if they were kept secret. The baseline legal principle was that no one could deliver what they did not have (*nemo dat quod non habet*), which was closely related to the principle “first in time, first in right.” So, in a double sale of land, in which an owner O sells first to buyer B_1 and later to B_2 , the land belongs to B_1 because, when O sold to B_2 , O was not the owner. In cases of conflict, the judge will allocate property and contract rights between both claimants (B_1 and B_2) – that is, will “establish title” – on the basis of evidence on possession and past transactions, whether or not these transactions had remained hidden.

This potential enforcement of adverse hidden rights made gathering of all relevant consents close to impossible, hindering trade, and specialization. Most transactions in land therefore gave rise, totally or partially, to contract rights, and the enforcement advantage of property rights remained unfulfilled, especially with respect to abstract rights, such as mortgages. These difficulties are clear in the functioning of the two sources of evidence traditionally used to establish title under privacy: possession and the chain of title deeds.

First, the use of possession – that is, as a first approximation, the fact of controlling the asset – as the basis for establishing property rights is a poor solution for durable assets, because for such assets it is often valuable to define multiple

rights, at least separating ownership and possession. However, relying on possession to establish ownership makes it possible for possessors to fraudulently use their position to acquire ownership for themselves or to convey owners’ rights to third parties. In such cases, owners will often end up holding a mere contract right, an in personam right, against the possessor committing the fraud. Understandably, under such conditions, owners will be reluctant to cede possession impersonally, for fear of losing their property. Similarly, credit will involve contractual, personal guarantees provided either by the debtor or by the lender. This is because the only way of providing some type of in rem guarantee to the lender is by transferring ownership or possession to him, thus leaving the debtor subject to the lender’s moral hazard and safeguarded only by the lender’s contractual guarantee, which is weaker and costly to produce.

Second, some of the problems posed by possession are solved by embodying abstract rights, such as ownership and liens, and even complementary consents in the conveying contracts, which then form a series or “chain” of title documents or deeds (“chain of deeds,” for brevity) that is based on what can be labeled “documentary formalization.” This evidencing of rights with the chain of deeds facilitates some degree of separation of ownership and control because it is the content and possession of deeds that provide evidence of ownership. Therefore, title experts can examine the history of transactions going back to a “root of title,” which is proof of ownership in itself – either because it is an original grant from the state or, more often, because of the time that has lapsed beyond the period of prescription or the statute of limitations. However, relying on the chain of deeds also creates problems. Above all, new possibilities for error and fraudulent conveyance appear, giving rise to multiple chains of title, which leave acquirers with contract rights against the fraudulent grantor and the professionals involved in the transaction. Moreover, titles are less effective than possession in reducing the asymmetry of acquirers, as possession is observable but adverse chains of title remain hidden to the acquirer. Furthermore, acquirers remain fully unprotected against any hidden charges that are

not voluntarily contracted, such as judgment and property tax liens. Similarly, the chain of deeds also serves to enforce a security, by pledging the deeds with the lender. But this solution is also defective, as it subjects the debtor to the lender's moral hazards (the lender could impede a sale or even fraudulently sell) and causes switching costs that make mortgage subrogation difficult.

Despite these difficulties, transactions on unregistered land in England heavily relied on the chain of deeds up until the last decades of the twentieth century. Typically, ownership was proved by possession and the whole series of deeds, which was often kept by the owner's solicitor. And mortgages were formalized by pledging the deeds with the lender. Likewise, during the *ancien régime* notaries public in most civil law countries played a similar role to that of English solicitors, with the advantage that each notary office kept an archive with the original of all the titles it notarized. This gave notaries privileged access to information on the transactions that each office had authorized. However, they did not provide an effective substitute for mortgage registries, mainly because notaries' information about individual debtors was incomplete. These cases therefore illustrate a constant feature of privacy regimes: to contain fraud, private conveyancing services provided by solicitors and notaries tend to develop into professional monopolies.

Whatever the system of documentary formalization for private conveyance contracts, conveying parties will always try to contract relying on the evidence that will eventually be used by the judge to establish title. Under privacy, however, given that courts may enforce in rem rights that have remained hidden, examination of title quality is based on potentially incomplete evidence. Therefore, removal of title defects and contradictions, as well as any adjustments to the terms of the private contract, are informed only by the limited and hard-to-verify publicity provided by possession and by documentary formalization and are motivated by the risk the grantor faces when giving title warranties on a defective title. But acquirers have limited possibilities of knowing what they are buying. They, in fact, acquire residual property in rem rights plus a contract in

personam right against the grantor for the difference between the in rem property rights effectively granted and what the grantor had promised to deliver.

Understandably, legal systems try to counterbalance this chronic incompleteness of property rights in their in rem dimension under a privacy regime by adopting private and public means to strengthen contract in personam rights, such as granting formal guarantees, expanding the scope of criminal sanctions, and even relying on bonding and slavery. Moreover, legal systems also provide specialized judicial procedures capable of purging title – that is, establishing which rights in rem are alive and who holds them – thus producing a public reallocation of rights that should be useful at least to solve the most complex and valuable cases.

Ceremonial Publicity and Recordation of Deeds

Whatever the palliatives applied, the costs of contracting true property – that is, in rem – rights under a regime of pure privacy are so high that modern systems of property law have abandoned privacy in an effort to lower them. At a minimum, the law induces or requires the independent publicity of contracts, which makes them verifiable, as a prerequisite for them to attain in rem effects – that is, to convey property rights and not mere contract rights. If they keep their claims private, right holders lose or risk losing in rem effects. Private contracts may create obligations among the conveying parties but do not bind third parties: all other right holders and, especially, potential future buyers and lenders. Independent publicity therefore facilitates finding out which property rights are alive and which will be affected, thus making it possible to gather consents, purge titles, and reduce information asymmetries between the conveying parties. At a maximum, in addition to requiring publicity of contracts, the law also requires a complete purge of conflicting claims for each transaction. Because this purge is supervised by a public registrar acting in a quasi-judicial capacity, the registry does

not merely provide publicity of claims but also defines and publicizes rights.

Specific laws therefore vary substantially with respect to how and when any contradiction with other property rights must be purged by obtaining the consent of the holders of these affected rights. This second contractual stage may be postponed indefinitely or may take place at the time of the private contract. In the latter case, it may be either voluntary or compulsory and total, and, if voluntary, it may also be total or partial. Moreover, jurisdictions also differ in what their registries produce, as they may either simply publicize the deeds evidencing potentially contradictory claims or certify fully purged property rights. Lastly, there is also a logical adaptation of the specific mechanisms needed to produce these outcomes in each environment, the set of rights enforced as property rights, and the adjudication rules in cases of disputed title.

Physically marking the assets is perhaps the simplest way of providing publicity of claims. The symbolic nature of marking makes it especially suitable for abstract rights, such as ownership and security interests. This explains why it has been used extensively for enforcing ownership in the absence of possession, as in valuable movables such as livestock, automobiles, and books. Another simple way of providing publicity is by using conspicuous contractual procedures. For example, ancient Roman law required a formal and public conveyance, either through a collusive proceeding before the court or through a public ceremony known as *mancipatio*. This performed a titling function, as it publicized the conveyance, but it also served to gather the consents of affected right holders and thus purged conflicting claims either at once or after a certain period. Similarly, after 1066, English conveyances followed the continental practice of delivering possession through a ritual known as livery of seisin. The publicity function was even clearer in the practices followed in some European regions, where laws mandated sophisticated procedures of publicity “before the church” and “at the gate of town walls” for rural and urban land, respectively, as well as judicial registration in some cases.

These old practices for reaching consent and purging property rights were effective because transactions mainly took place between neighbors. For neighbors, it is easy to notice announcements and public deals, especially for the kinds of rights common in rural and traditional societies, many of which were linked to family matters. It is revealing that the effect of publicity “before the church” was immediate for right holders who were present but was delayed for one year and one day for those absent. Costlier knowledge was apparently balanced with longer time, suggesting that these systems could hardly support impersonal trade.

The next logical step in the provision of publicity is to lodge private transaction documents (i.e., the title deeds) for filing in a public registry so that this evidence on property claims can then be used by the courts to verify them and allocate property, *in rem*, rights in case of litigation. Moreover, by making this register publicly accessible to potential acquirers, the latter can ascertain the quality of the seller’s title, thus reducing their information asymmetry.

After many failed attempts, such as the Statute of Enrollments issued by Henry VIII in 1535 but never enforced, and the Massachusetts 1640 Recording Act, recordation of deeds eventually started to succeed in the nineteenth century and has been used in most of the United States, part of Canada, France, and some other countries, mostly those with a French legal background. The key for its success was to switch the priority rule, because other incentives had failed to convince people to record. Historically, recordation systems thus became effective only once courts, when deciding on a conflict with third parties, started to determine the priority of claims from the date of recording in the public office and not from the date of the deed. This means that, instead of the conventional “first in time, first in right” rule, courts adjudicate according to the rule “first to record, first in right.” For instance, in terms of a double sale, the judge gives the land not to the first buyer but to the first buyer to record the purchase document.

This change in the priority rule not only protects acquirers but also avoids incomplete

recording, which hampered many of the first recordation systems. The reason is that the switch in the priority rule effectively motivates acquirers to record from fear of losing title through a second double sale or any other granting of rights (e.g., a mortgage) by the former owner to an innocent acquirer (e.g., a lender) who might record first. Consequently, all relevant evidence on property rights is available in the public records. From the point of view of third parties, the record, in principle, is complete. Other claims may not be recorded and may well be binding for the parties who have conveyed them, but these hidden claims have no effect on third parties.

The inclusiveness of the record of deeds makes it possible to assess the quality of title by having experts examine all relevant deeds, that is, only those that have been recorded, and producing “title reports.” If there is sufficient demand, a whole title assurance industry will develop for examining, gathering consents, purging, and assuring title quality. This industry may take different forms. It is composed, for instance, of notaries public in France and of title insurers in many of the United States, while abstractors, attorneys, title insurance agents, and title insurance underwriters perform separate functions in other United States. Despite their different names and differing degrees of vertical integration, the industry performs similar functions in all countries, as it mainly reduces information asymmetry between the conveying parties and encourages them to voluntarily purge the title. In particular, expert search for title defects, which can thus be removed before contracting by obtaining the relevant consent. Alternatively, if not removed, the grantee will not transact or will insist on modifying the content of the private contract, reducing the price or including additional warranties, in compensation for the survival of the defect. To motivate experts’ diligence and technical innovation and to spread remaining risks, a standard close to strict liability is often applied to such examination and assurance activities. Consequently, experts are strongly motivated to find any defects on the title and a substantial part of the remaining title risk is reallocated from acquirers to title experts and their insurers.

Moreover, as under the privacy regime, both contractual and judicial procedures are used to remove title defects. Compared to privacy, deed recordation provides more possibilities for contracting the removal of defects, because defects are better known to buyers and insurers. The identification of right holders also gives greater security to the summary judicial hearings that serve to identify possible adverse claims and publicly reallocate in rem rights. These summary hearings continue to exist today in, for example, the French judicial purge and the US “quiet title” suit. In addition to purging titles directly, the existence of such a court-ordered purging possibility also reduces bargaining costs indirectly by encouraging recalcitrant claimants to reach private agreements.

However, the recording office accepts all deeds respecting certain formal requirements (mainly, the date of the contract and the names of the conveying parties), whatever their legality and their collision with preexisting property rights. In fact, the recording office is often obliged by law to file all documents fulfilling a set of formal requirements, regardless of their legal status. The public record may therefore contain three kinds of deed. First, those resulting from private transactions made without previous examination. Second, those granted after an examination but without having all defects removed. Finally, those that define purged and noncontradictory property rights.

Transactors who record clouded titles therefore produce a negative externality for all future transactors. When examining the title of a parcel, experts do not know a priori which kinds of deed are recorded concerning it. Therefore, for each transaction, they will have to examine all relevant deeds dealing with that parcel, even those which may have been perfectly purged in previous transactions. The cost of this repeated examining of deeds can be reduced with proper organization of the registry. In the short run, the easiest way to organize the information is by relying on indexes of grantors and grantees to locate the chain of transactions for a given parcel. However, this method is subject to errors, such as those caused by identical names and misspellings. This

explains the steps taken, for instance, in 1955 to create the *fichier immobilier* in the French Registry and to forbid recording a deed if the grantor's title is not recorded. Another way of reducing costs when public records are poorly organized is to build privately owned, indexed databases (known in the United States as "title plants"). These plants replicate public records in a more complete, reliable, and accessible manner by transferring and abstracting relevant documents lodged at the public registries and building tract indexes to easily locate the relevant information for each land parcel.

Registration of Rights

Registration of rights (hereafter referred to as "registration," and often confusingly called "title registration") goes one crucial step further than recordation of deeds: instead of providing information about claims, it defines the rights (what, in jurisdictions with Torrens registries, is often referred to as "title by registration"). To do this, it requires a mandatory purge of claims before registering the rights. As in deed recordation, claims stemming from private transactions gain priority when transaction documents are first lodged with the registry. They are then subject, however, to substantive review by the registrar in order to detect any potential conflict that might damage other property rights. New and reallocated rights are registered only when the registrar determines that the intended transaction does not affect any other property right or that the holders of these affected rights have consented. When these conditions are met, the change in rights caused by the transaction is registered, antedating the effects of registration to the lodging date. (In a sense, any registry of rights thus contains a recording of deeds: its "lodgment" or "presentment" book is a temporary record of claims.) Otherwise, when the consent of an affected right holder is lacking, registration is denied, and the conveying parties have to obtain the consents relevant to the originally intended transaction, restructure it to avoid damaging other rights, or desist.

Registration aims to eliminate all uncertainties and information asymmetries, as information in the register is simplified in parallel with the purge of rights. Ideally, rights defined in each new contract are registered together with all surviving rights on the same parcel of land. Extinguished rights are removed or deleted, making it easy to know which are the valid rights. Production of information is a key element. As pointed out by Baird and Jackson, "in a world where information is not perfect, we can protect a later owner's interest fully, or we can protect the earlier owner's interest fully. But we cannot do both" (1984, p. 300). The assertion is accurate but the assumption is crucial: registration intends to produce perfect information and thus protect both the earlier and the later owners. Its goal is to abide by three principles traditionally deemed desirable for a titling system, according to which: (1) the register reflects the reality of property rights, so that potential transactors do not need to look out of the register ("mirror principle"); (2) the register reflects only valid rights, so that transactors do not need to perform a title search in the chain of title ("curtain principle"); and (3) losses caused by a registry's failure are indemnified ("assurance, insurance, or guarantee principle").

To the extent that these three principles are achieved and given that any contradictions are purged before registration, the registry is able to provide "conclusive," "indefeasible" title, meaning that a good faith third party "for value" (i.e., one who pays for the property rather than receiving it as a gift) acquires a property right if the acquisition is based on the information provided by the registry. If the seller's right is later shown to be defective, the buyer keeps the property – that is, in rem – right and the original owner gets contract in personam rights against the seller and the registry. When functioning correctly, the register is thus able to provide potential transactors with a complete and updated account of the in rem rights alive on each parcel of land. Given the enforcement advantage of in rem rights, this accounting amounts to commoditizing the legal attributes of rights on real property, which makes impersonal trade much easier.

Furthermore, registration interferes little with private property, as registry intervention focuses

on the timing and completeness of the reallocation of rights implicit in any purge. Registration is controlled by registrars, but ultimate decisions are made by right holders by giving their consent. Privacy and recordation allow conveying parties more discretion on timing and heavier reliance on privately produced information. They therefore seem to rely more on private decisions, but this perception is deceptive because even recorded titles are in fact mere claims. They retain a higher contractual content, given the survival of conflicting claims in rem. Additional intervention by the court, also subject to the possibility of allocation failure, would be required to transform them into property rights at in rem level equivalent to that provided by registration. In sum, as compared to recordation, it is useful to see registration as a quasi-judicial step, which in other titling systems is also necessary to reach full in rem enforcement. This similar degree of public involvement helps explain why both registration and recordation have taken root in countries with different legal traditions and why there is little correlation between titling systems and legal traditions.

Cross-References

- ▶ [Coase and Property Rights](#)
- ▶ [Development and Property Rights](#)
- ▶ [Emissions Trading](#)
- ▶ [Externalities](#)
- ▶ [Good Faith](#)
- ▶ [Informal Sector](#)
- ▶ [Institutional Change](#)
- ▶ [Lex Mercatoria](#)
- ▶ [Limits of Contracts](#)
- ▶ [Market Failure: Analysis](#)
- ▶ [Public Enforcement](#)
- ▶ [Transaction Costs](#)

References

- Arruñada B (2012) Institutional foundations of impersonal exchange: the theory and policy of contractual registries. University of Chicago Press, Chicago

- Baird DG, Jackson TH (1984) Information, uncertainty, and the transfer of property. *J Leg Stud* 13:299–320

Further Reading

- Arruñada B (2003) Property enforcement as organized consent. *J Law Econ Org* 19:401–444
- Arruñada B (2011) Property titling and conveyancing. In: Ayotte K, Smith HE (eds) *Research handbook on the economics of property law*. Edward Elgar, Cheltenham, pp 237–256
- Arruñada B (2015) The titling role of possession. In Chang Y (ed) *The law and economics of possession*. Cambridge University Press, Cambridge
- Arruñada B, Garoupa N (2005) The choice of titling system in land. *J Law Econ* 48:709–727
- Deininger K, Feder G (2009) Land registration, governance, and development: evidence and implications for policy. *World Bank Res Obs* 24:233–266
- Ellickson RC, Thorland CD (1995) *Ancient land law: Mesopotamia, Egypt, Israel*. Chicago-Kent Law Rev 71:321–411
- Hansmann H, Kraakman R (2002) Property, contract, and verification: the numerus clausus problem and the divisibility of rights. *J Leg Stud* 31:S373–S420
- Merrill TW, Smith HE (2000) Optimal standardization in the law of property: the numerus clausus principle. *Yale Law J* 110:1–70
- Merrill TW, Smith HE (2001) The property/contract interface. *Columbia Law Rev* 101:773–852
- Merrill TW, Smith HE (2001) What happened to property in law and economics? *Yale Law J* 111:357–398

Tort Damages

Louis Visscher
Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Definition

The amount of monetary compensation a tortfeasor has to pay to the plaintiff(s) in a tort case when he is found liable.

Introduction

Liability rules in tort law determine when a tortfeasor is liable. The rules of tort damages

determine the amount the liable tortfeasor subsequently has to pay. Together these two bodies of law therefore determine in which situations the tortfeasor has to pay how many damages to the plaintiff(s). Therefore, the behavioral incentives provided by tort law depend on both sets of rules. In this entry, the most important insights from the economic analysis of tort damages are presented. The limited space does not allow a full discussion of all possible complications (see Visscher 2009 for a more complete overview), but where relevant such complications are briefly indicated. It is also not possible to include a full discussion of the liability rules themselves, but in the remainder of this introduction, a very brief account of this topic is provided.

The economic analysis assumes that actors are incentivized to take precautionary measures by the threat of liability (tortfeasors) or the prospect of having to bear one's own losses (victims). Such measures can consist of taking care and/or adapting the activity level. Optimal measures are taken when the marginal costs equal the marginal benefits (see, e.g., Schäfer and Ott 2005; Cooter and Ulen 2012). The social benefits of precautionary measures consist of the reduction in expected accident losses (i.e., the probability of an accident multiplied by the magnitude of the losses if an accident happens). The private benefits of a party consist of the reduction in the losses they themselves have to bear. In order to provide socially desirable incentives, tort damages therefore should equal the social losses, which in principle calls for full compensation (*restitutio in integrum*), and negligence should be defined as taking less than socially optimal care.

An essential difference between strict liability and negligence is that under the first rule the tortfeasor is liable irrespective of his care level and under the latter rule he is only liable if he took less than due care. This implies that whereas under strict liability tort damages indeed should be full, under negligence tort damages can be lower, because they only have to make taking due care (and hence escaping liability) more attractive than taking low care (and hence being liable). However, if a negligent tortfeasor is only held liable for the losses which are caused *by his negligence*, this

difference disappears and also negligence requires full compensation.

In bilateral accident situations, where both the tortfeasor and the victim can influence the accident probability, both parties should receive incentives for optimal precautions. Negligence provides optimal care incentives to both parties (the tortfeasor takes optimal care to avoid liability and the victim to minimize the costs he himself has to bear), but strict liability needs a defense of comparative or contributory negligence to achieve this. No liability rule can give optimal activity incentives to both parties, because only the party who ultimately bears the accident losses will choose the correct activity level.

Pecuniary losses are either monetary losses or losses of replaceable goods, where the replacement costs are a good measure of the losses. Nonpecuniary losses consist of damage to irreplaceable things such as family portraits but also health and emotional well-being (Shavell 2004, p. 242). For pecuniary losses, the concept of full compensation is relevant, because damages can make the victim indifferent between the situation without the tort on the one hand and the situation with the tort and with damages on the other hand (also see section “[What is Full Compensation?](#)”). With nonpecuniary losses, it is problematic that money cannot truly compensate such losses, that they cannot be observed directly, and that there is no market so that one needs to apply indirect methods of assessing them (see section “[Non-pecuniary Losses](#)”). In order to provide the correct incentives to the injurer, tort damages should equal the sum of pecuniary and nonpecuniary losses (Shavell 2004, p. 242).

What is Full Compensation?

In many European countries, the *difference hypothesis* of Mommsen is important in assessing the losses which the liable tortfeasor has to compensate. In this approach, the loss consists of the difference between the wealth as it is at a certain moment and the wealth as it would have been at the relevant moment if the loss causing event would not have happened. In economic terms,

this can be expressed as follows: before the accident, the victim had a certain level of utility. After the accident, he has a lower level of utility, and the difference in both utility levels is his loss. Even though Mommsen defined the loss as a difference between two wealth levels, in many jurisdictions also nonpecuniary losses are compensable. Again in economic terms, the decrease in utility can not only be caused by a reduction in wealth but also by a reduction in other sources of utility, which may have a nonpecuniary nature.

As explained in the Introduction, in order to provide correct incentives, tort damages in principle should fully compensate the victim for his losses (Posner 2003, p. 192; Cooter and Ulen 2012; Shavell 2004, p. 236) and should be based on the social losses. This way, the tortfeasor internalizes the negative externality he has created. By receiving an adequate amount of money, the victim can be brought back to his original utility level, or better, to the utility level he would have reached without the tort. Under negligence, if the tortfeasor would have taken due care, the victim also would run a certain risk of being harmed, so that he should not be brought to a position in which he would run no risk at all. This implies that the victim should not be compensated for losses which also would have happened at due care (Van Wijck and Winters 2001).

Several remarks can be made regarding the principle of full compensation:

- If victims can mitigate their losses by taking measures which cost less than the reduction in accident losses they yield, they should be incentivized to do so. Limiting damages to optimally mitigated losses plus mitigation costs provides the proper incentives (Shavell 2004, p. 248ff).
- If repair is more expensive than replacement, damages should be based on replacement. If repair or replacement is not (completely) possible, monetary damages should be high enough to bring the victim back to his original utility curve.
- Costs which the victim would have incurred anyway should not be compensated (e.g., gasoline costs of a rental car which is needed during repair of the victim's car).
- Damages should take interest, taxes, and inflation into account. For future losses, there is a preference for lump sum rather than periodic contingent payments. A victim who prefers periodic payments can convert the lump-sum payment into periodic payments, lump-sum payment avoids the uncertainties of periodic contingent payment as well as possible moral hazard on the side of the victim, and the victim is not confronted with ongoing monitoring to assess the extent of the losses which would be required under periodic contingent payment (Rea 1981, p. 132).
- The injurer has to compensate the losses of the victim, also if they are higher than normal. Such losses should not be limited on the basis of foreseeability or adequacy but rather the tortfeasor should take the victim as he finds him. After all, limiting damages if they are higher than normal would result, on average, in too few incentives (Shavell 2004, p. 239).
- Judicial moderation and limitation of damages are generally critically assessed by law and economics scholars, because they result in less than full compensation. The often-heard argument of uninsurability of the full losses is not convincing, because it is not primarily the magnitude of losses that results in uninsurability but uncertainty regarding the accident probability. Restricting liability does not solve that issue and might even result in higher losses by diminishing tort law's care incentives (Cooter and Ulen 2012; Shavell 2004, p. 230ff). Restricted liability may induce a potential tortfeasor to engage in a socially desirable activity which he, due to risk aversion, might not engage in otherwise. But it could also induce a potential tortfeasor to engage in an activity which is socially *undesirable* due to the high costs, of which he now only has to bear a fraction.
- One could base damages not on the true losses in individual cases but on the average loss. Such an abstract, objective method of damage assessment instead of a concrete, subjective method can greatly reduce the administrative

costs of tort law. For example, tort damages for a damaged car can be based on the costs of repair by a competent mechanic, even if in the actual case the victim did not have the car repaired or did it himself. The reduction in administrative costs in such often-occurring losses outweighs the possible disadvantage of less fine-tuned deterrence incentives. It is even the question if there is a reduction in deterrent incentives in the first place, because a better damage assessment *ex post* does not result in better incentives if the tortfeasor *ex ante* does not know the exact losses he may cause. As long as the abstract method assesses the average damages correctly, the injurer receives the correct incentives (Kaplow and Shavell 1996). If losses are systematically over (under) estimated, the tortfeasor is generally assumed to choose a too high (low) care level and a too low (high) activity level.

- Harm to the victim generally speaking is a better basis for tort damages than gain to the injurer. In the latter case, if courts would underestimate the gain, the incentives would be inadequate (Polinsky and Shavell 1994, p. 431ff). Furthermore, basing damages on the loss results in internalization of the externality and allows an activity which yields more benefits than costs to be continued. However, if harm to the victim is difficult to assess, e.g., due to its subjective nature, basing damages on the gain may provide better deterrent incentives. Also if one wants to fully deter the behavior, rather than “merely” inducing the tortfeasor to take optimal precautionary measures, basing damages on gain is preferable.
- If litigation costs are taken into account, optimal damages may differ from full compensation of harm (Polinsky and Shavell 2014).

Nonpecuniary Losses

As explained in the Introduction, nonpecuniary losses are difficult to assess because they cannot be observed directly (see e.g. Arlen 2013, p. 439ff). This could induce victims to overstate such losses, and therefore it is sometimes

suggested not to compensate them at all when they are likely to be small. This would greatly reduce administrative costs and only have a limited impact on deterrence. Adams provides a different argument why such losses should not be compensated. Victims who have to bear these losses themselves receive desirable behavioral incentives (Adams 1989, p. 215ff). The problem with this view is, of course, that tortfeasors then receive too few incentives because they only compensate part of the losses. If nonpecuniary losses are too large to ignore, one could use tables or formulas to determine damages, which is a form of an abstract method as discussed in section “[What is Full Compensation?](#)” In order to avoid structural over- or undercompensation, it is important to try to find a method which assesses such losses correctly on average (if at all possible). This idea will be further developed below.

Given that nonpecuniary losses cannot be observed directly, an indirect way to assess them is required. The most well-known indirect method in this respect is provided by the *insurance theory*. This approach argues that a victim should only receive compensation for losses against which he is willing to insure himself. Otherwise, tort law would force coverage upon him which he does not want, but for which he might have to pay via higher prices for goods or services produced by the liable tortfeasor. According to the insurance theory, nonpecuniary losses do not increase marginal utility of wealth, and hence rational victims are not willing to insure against such losses, because the premium would cost more utility than the expected coverage yields (Friedman 2000, p. 95ff; Shavell 2004, p. 270ff). The fact that in practice there is no demand for such insurance is regarded as corroboration of the line of reasoning. The insurance theory therefore claims that victims should not receive compensation for nonpecuniary losses. The injurer, however, should pay for them, so that liability should be decoupled from compensation (Geistfeld 1995, p. 799ff).

Various critiques regarding the insurance theory have been expressed. First, even if marginal utility of wealth would not increase, potential victims might still want to insure against such

losses because the money they would receive after suffering these losses would mitigate the decrease in the *level* of utility, irrespective of the *marginal* utility of wealth at that point. Second, the fact that there is no demand for insurance against nonpecuniary losses may be caused by imperfect information regarding the extent of nonpecuniary losses, the probability of their occurrence, and the compensation required in respect of them or by countervailing social norms in the form of societal hostility to putting a price on pain and sorrow and legal restrictions such as the indemnity principle (Croley and Hanson 1995, p. 1845ff). Third, the argument that nonpecuniary losses do not result in an increased marginal utility of wealth may be flawed to start with. Empirical research which suggests that marginal utility of wealth decreases after suffering nonpecuniary losses is mostly based on personal injury situations where non-disabled people are asked how they think personal injuries would affect them if they would suffer such losses. It is doubtful whether nondisabled people can accurately assess the impact of such injuries on the marginal utility of money, and insights regarding adaptation suggest that this research results in an underestimation of marginal utility which may in fact increase after the nonpecuniary loss is suffered (Pryor 1993, p. 116ff).

But even if marginal utility would not increase, and even if people would not self-insure against nonpecuniary losses, it is questionable whether this should lead to the conclusion that victims should not receive compensation for such losses. Insurance decisions provide information on the degree of risk aversion of the actor involved but not on his willingness to pay to avoid the loss. If the victim would be willing to forego resources ex ante in order to avoid the nonpecuniary loss or at least to lower the chances of suffering them, this would imply that he does regard these losses as undesirable and that they do lower his utility level. The resources the victim himself is willing to spend on accident avoidance therefore provide information of how the victim himself assesses the nonpecuniary loss. This is a better indirect way of assessing damages for nonpecuniary losses than the insurance decision, which – again – regards risk aversion rather than

loss assessment. By basing damages on the resources victims themselves would have been willing to spend on accident avoidance, the victims are not overcompensated against their will, nor are potential injurers overdeterred because, in essence, they bear the costs of avoidance measures which were worthwhile ex ante.

For fatal accidents, the so-called *Value of a Statistical Life* (VSL) provides information on the willingness to pay to lower the chance of fatal accidents. The VSL is derived from all kinds of decisions people take which affect health and safety, such as wearing a helmet when riding a (motor)bike, installing a smoke detector at home, asking a premium for dangerous work conditions, etc. Such decisions contain an implicit trade-off between money and risks. If, for example, 1,000 people are each willing to spend € 1,000 on a measure which reduces the probability of a fatal accident by one pro mille, one statistical life is saved by spending € 1,000,000. The VSL in this example then is at least € 1,000,000 because this population was willing to spend at least that amount. In 2004, Sunstein assessed the VSL at about \$6.1 million (Sunstein 2004, p. 205; Posner and Sunstein 2005, p. 563). Expressed in 2014 euros this would amount to about €6.4 million. This American VSL is comparable to that in other developed countries (Viscusi and Aldy 2003, pp. 24, 35, and 63). If one wants to distinguish between victims on the basis of their age, one should use the *Value of a Statistical Life-Year* (VSLY) instead. Posner and Sunstein argue that damages for fatal accident which are based on the VSLY should be around \$6 million or higher and including the emotional loss of surviving relatives could result in an increase of several millions of dollars as well (Posner and Sunstein 2005, p. 586ff). These amounts are much higher than currently awarded in European countries, where damages are often limited to funeral costs and loss of maintenance. This implies that from an economic point of view, damages for fatal accidents are structurally undercompensated.

Schäfer and Ott argue that for nonfatal injuries, damages should be “some fraction” of the value attached to the willingness to pay to prevent death (Schäfer and Ott 2005, p. 373), but they do not

suggest how the appropriate fractions could be determined. The concept of the *Quality-Adjusted Life-Year* (QALY) could play a good role here (Miller 2000; Karapanou and Visscher 2010). QALYs express the value of living 1 year in a certain health condition, which is a proxy of the quality of life during that period. By combining information regarding the QALY value of that health condition and the duration thereof on the one hand and a monetary value of a QALY on the other hand, it is possible to express in monetary terms the loss of utility due to suffering personal injuries. QALYs are used to evaluate health programs and medical treatments and techniques and hence express how much society is willing to prevent or cure the health conditions under consideration. They therefore can provide information on the societal willingness to pay to avoid or cure such losses and are therefore suitable as a basis for the assessment of pain and suffering damages for personal injuries. It turns out that applying this method, even with conservative monetary values, results in higher pain and suffering damages than currently awarded in most European countries, so that here as well, currently these losses are structurally undercompensated.

Pure Economic Loss

In the Introduction, it became clear that tort damages should in principle be based on the social loss the tortfeasor has caused. In cases of pure economic loss, the private losses of the victim often do not coincide with the social loss. The private losses might be offset by private gains elsewhere, so that the tort resulted in a redistribution of social welfare rather than in a reduction thereof. If firm A cannot produce due to a power supply interruption caused by a tort, but firm B now produces and sells more, there might not be a social loss at all, and tort liability then makes no economic sense. This redistribution argument is regarded as a reason not to include pure economic loss in tort damages (Bishop 1982; Gómez and Ruiz 2004; Dari-Mattiacci and Schäfer 2007, p. 10). However,

several reasons exist why there might be a social loss after all.

First, the products of firm B might not be a perfect substitute so that there is a loss of consumer surplus.

Second, firm B had to have overcapacity in order to meet the additional demand, and overcapacity in itself can be regarded as inefficient. Compensating pure economic loss provides better behavioral incentives to potential tortfeasors, resulting in fewer accidents and hence a reduced need for overcapacity (Rizzo 1982, p. 202).

Third, there will be many cases of pure economic loss where there is not only a redistribution of wealth but also a decrease in social welfare. For example, if an accountant negligently approves the balance sheet of a firm, besides the redistribution of welfare between buyers and sellers of overpriced stock, such faults may decrease trust in information provided by accountants and induce parties to do additional research and result in suboptimal investment decisions. These are all examples of social losses.

Gómez and Ruiz present two other reasons why pure economic loss should sometimes be included in tort damages. In situations where the victim in essence “insured” someone else against losses by covering them, it is desirable that he can recover his pure economic loss from the tortfeasor due to the same reasons why subrogation in real insurance contracts is desirable. An example of this is an employer who pays wages to his employee also during the period where the latter cannot work because he is injured in a tort. It is also possible that a third party suffers pure economic loss due to breach of contract. The third party might not be able to recover these losses under contract law because he is no party to the contract. Tort law could then serve as a surrogate for contractual liability (Gómez and Ruiz 2004).

Punitive Damages

Punitive damages are damages which exceed compensatory damages and are used to punish wrongdoers in cases of, e.g., malice, gross

negligence, or intent. Continental European jurisdictions generally do not award punitive damages. In debates regarding punitive damages, a fear for “American situations” in which enormous amounts of damages are awarded is often expressed. However, such large amounts are often reduced by courts, and empirical research shows that punitive damages are not often awarded and generally speaking are not very high, and they are correlated with compensatory damages (Cohen and Harbacek 2011; Vidmar and Wolfe 2009, p. 189, However, also see Polinsky 1997). The “American situations” are therefore more myth than reality, and fear for them should not frustrate the debate on introducing punitive damages in continental Europe. This holds even stronger now there are good economic arguments in favor of punitive damages (also see Arlen 2013, p. 488ff).

First, punitive damages can ameliorate the problem that the probability that a tortfeasor will actually be held liable is below 100%. By increasing damages, the *expected* damages can again equal social harm. For example, if the probability of being held liable for causing a loss of 100 is 25%, but total damages (consisting of compensatory and punitive damages) are 400, the expected liability is again 100 ($25\% \cdot 400$) (Cooter 1989; Polinsky and Shavell 1998). This idea is consistent with the fact that punitive damages are often awarded in cases of intentional torts, because there the injurer might try to avoid being caught.

Second, if compensatory damages do not cover all (types of) losses of the victim, damages are too low from an economic viewpoint because they do not cover the full externality. Such undercompensatory damages can result from losses which are difficult to assess, e.g., because they are nonpecuniary or subjective. In principle, punitive damages could tackle this problem. The fact that in cases of defamation (with indeed subjective and nonpecuniary losses) punitive damages are often possible fits this idea. However, in order to assess the correct magnitude of punitive damages, one has to assess the extent of undercompensation of normal damages, and if one would be able to assess that, one could better

increase compensatory damages by this amount (Polinsky and Shavell 1998, p. 940).

Third, if the utility a tortfeasor yields from his activity is regarded as socially illicit so that we do not want to include it into the social weighing of costs and benefits, compensatory damages may not adequately deter the undesirable behavior because the tortfeasor still yields his utility. Punitive damages could result in more deterrent incentives (Cooter 1989, p. 79ff; Shavell 2004, p. 245). The value of this third rationale is limited, because (1) many of the acts which yield socially illicit utility will be subject to criminal liability and (2) labeling certain benefits as socially illicit *assumes* the conclusion that these acts are socially undesirable rather than that this is *proven* on the basis of a weighing of costs and benefits (Friedman 2000, p. 229ff).

Fourth, the threat of punitive damages could induce a potential tortfeasor to seek a voluntary transaction with the potential victim, rather than to commit the tort. Law and economics scholars generally prefer such voluntary transactions because then also subjective elements in the valuation are included in the decision whether or not to transfer an entitlement, whereas tort damages do not include these elements, so that they might not fully compensate the victim. In cases where a voluntary transaction was possible, the potential tortfeasor should be incentivized not to take away the entitlement without consent and to pay the objective damages afterward, exactly because of the possible subjective elements. Punitive damages can help by making the involuntary transfer more expensive (Shavell 2004, p. 245ff).

Fifth, victims who start a tort claim in essence serve the social goal of deterrence, yet they bear all the costs. If the costs outweigh the expected private benefits, they may stay rationally apathetic. Punitive damages, such as treble damages in antitrust cases, may ameliorate this situation.

Sixth and final, in as far as independent value is attached to the punishment goal (the desire of individuals to see blameworthy parties appropriately punished), punitive damages can ensure that more reprehensible behavior results in a more severe sanction (Polinsky and Shavell 1998,

p. 948ff). This is consistent with the fact that punitive damages are often connected to malice, gross negligence, or intent.

Conclusion

As a starting point, tort damages should fully encompass the externality caused by the tortfeasor. Under negligence, lower damages are possible as long as they make taking optimal care more attractive than being negligent. Victims should receive incentives to take optimal precautionary and mitigation measures. Tort damages should encompass both pecuniary and nonpecuniary losses, unless the administrative costs of assessing the latter outweigh the behavioral benefits of including them. In principle, especially for difficult classes of losses, an objective abstract assessment of damages suffices, as long as systematic over- or undercompensation is avoided. It seems that actual nonpecuniary damages for fatal accidents and personal injuries are (much) lower than economically desirable. Pure economic loss should only be compensated if it also entails a social loss. Punitive damages serve several economic goals, and the European debate should not be shut down by a simple reference to “American situations,” which are more myth than reality.

References

- Adams M (1989) Warum kein Ersatz von Nichtvermögensschäden? In: Ott C, Schäfer HB (eds) *Allokationseffizienz in der Rechtsordnung*. Springer, Berlin, pp 210–217
- Arlen J (ed) (2013) *Research handbook on the economics of Torts*. Edward Elgar, Cheltenham
- Bishop W (1982) Economic loss in Tort. *Oxford J Legal Stud* 2:1–29
- Cohen TH, Harbacek K (2011) Punitive damage awards in state courts, 2005. US Department of Justice Special Report March 2011
- Cooter RD (1989) Punitive damages for deterrence: when and how much? *Ala Law Rev* 40:1143–1196
- Cooter RD, Ulen TS (2012) *Law and economics*, 6th edn. Addison Wesley, Boston
- Croley SP, Hanson JD (1995) The nonpecuniary costs of accidents: pain and suffering damages in Tort law. *Harv Law Rev* 108:1785–1917
- Dari-Mattiacci G, Schäfer HB (2007) The core of pure economic loss. *Int Rev Law Econ* 27:8–28
- Friedman DD (2000) *Law's order. What economics has to do with law and why it matters*. Princeton University Press, Princeton
- Geistfeld M (1995) Placing a price on pain and suffering: a method for helping juries determine Tort damages for nonmonetary injuries. *Calif Law Rev* 83:773–852
- Gómez F, Ruiz JA (2004) The plural – and misleading – notion of economic loss in Tort: a law and economics perspective. *Z Eur Priv* 12:908–931
- Kaplow L, Shavell S (1996) Accuracy in the assessment of damages. *J Law Econ* 39:191–210
- Karapanou V, Visscher LT (2010) Towards a better assessment of pain and suffering damages. *J Eur Tort Law* 1:48–74
- Miller TR (2000) Valuing nonfatal quality of life losses with quality-adjusted life years: the health economist's meow. *J Forensic Econ* 13:145–167
- Polinsky AM (1997) Are punitive damages really insignificant, predictable and rational? A comment on Eisenberg et al. *J Legal Stud* 26:663–677
- Polinsky AM, Shavell S (1994) Should liability be based on the harm to the victim or the gain to the injurer? *J Law Econ Organ* 10:427–437
- Polinsky AM, Shavell S (1998) Punitive damages: an economic analysis. *Harv Law Rev* 111:869–962
- Polinsky AM, Shavell S (2014) Costly litigation and optimal damages. *Int Rev Law Econ* 37:86–89
- Posner RA (2003) *Economic analysis of law*, 6th edn. Aspen Publishers, New York
- Posner EA, Sunstein CR (2005) Dollars and death. *Univ Chic Law Rev* 72:537–589
- Pryor ES (1993) The Tort law debate, efficiency, and the kingdom of ill: a critique of the insurance theory of compensation. *Virg Law Rev* 79:91–152
- Rea S (1981) Lump sum versus periodic damage awards. *J Legal Stud* 10:131–154
- Rizzo MJ (1982) The economic loss problem: a comment on bishop. *Oxf J Legal Stud* 2:197–206
- Schäfer HB, Ott C (2005) *Lehrbuch der ökonomischen Analyse des Zivilrechts*, 4th edn. Springer, Berlin
- Shavell S (2004) *Foundations of economic analysis of law*. The Belknap Press of Harvard University Press, Cambridge, MA
- Sunstein CR (2004) Lives, life-years, and willingness to pay. *Columbia Law Rev* 104:205–252
- Van Wijck P, Winters JK (2001) The principle of full compensation in Tort law. *Eur J Law Econ* 11:319–332
- Vidmar N, Wolfe MW (2009) Punitive damages. *Ann Rev Law Soc Sci* 5:179–199
- Viscusi WK, Aldy JE (2003) The value of a statistical life: a critical review of market estimates throughout the world. *J Risk Uncertain* 27:5–76
- Visscher LT (2009) Tort damages. In: Faure MG (ed) *Tort law and economics*, vol I, 2nd edn, *Encyclopedia of law and economics*. Edward Elgar, Cheltenham, pp 153–200

Tourism

Marianna Succurro

Department of Economics, Statistics and Finance,
University of Calabria, Rende (CS), Italy

Abstract

The term tourism indicates all the heterogeneous activities and services referring to the temporary transfer of people from the habitual residence to other destinations for diverse reasons. It is a social, cultural, and economic phenomenon which is very important, and, in some cases vital, for many countries. Over the past decades, tourism has experienced continued expansion and diversification, becoming one of the largest and fastest-growing economic sectors in the world. It generates also direct effects on the social, cultural and educational sectors of national societies and on their international relations. Due to these multiple impacts, there is a need for a holistic approach to tourism analysis, development, management and monitoring in order to formulate and implement national and local tourism policies and international agreements.

Definition

The term tourism indicates all the heterogeneous activities and services referring to the temporary transfer of people from the habitual residence to other destinations for leisure, amusement, entertainment, culture, medical treatment, business, sport, and other purposes.

These people are called visitors (which may be either tourists or excursionists, residents or nonresidents), and tourism has to do with their activities, some of which imply tourism expenditure.

In 1936, the League of Nations defined a foreign tourist as “someone traveling abroad for at least 24 h.” Its successor the United Nations amended this definition in 1945, by including a maximum stay of 6 months.

Tourism and its Multiple Impacts

Tourism is a social, cultural, and economic phenomenon which is very important, and, in some cases vital, for many countries.

Many researchers reckon it is a key driver of socioeconomic progress through export revenues, the creation of jobs and enterprises, and infrastructure development. As such, tourism has implications on the economy, on the natural and built environment, on the local population at the destination, and on the tourists themselves.

In the *Manila Declaration on World Tourism of 1980*, tourism is recognized as “an activity essential to the life of nations because of its direct effects on the social, cultural, educational, and economic sectors of national societies and on their international relations.” Due to these multiple impacts, there is a need for a holistic approach to tourism analysis, development, management, and monitoring in order to formulate and implement national and local tourism policies and international agreements.

The Increasing Importance of Tourism in the World as Consumer Good

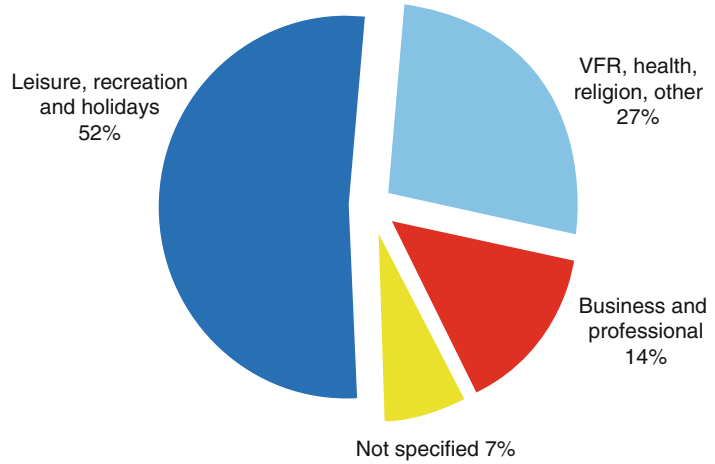
The origins of modern tourism can be traced back to the traditional trip of Europe undertaken by mainly upper-class European young men. The custom flourishes from about 1660 until the advent of large-scale rail transit in the 1840s, usually associated with a standard itinerary.

Mass tourism develops with the improvements in infrastructures and technology, allowing the transport of large numbers of people in a short space of time to places of leisure interest, so that greater numbers of people can begin to enjoy the benefits of leisure time.

Over the past six decades, indeed, tourism has experienced continued expansion and diversification, becoming one of the largest and fastest-growing economic sectors in the world. Many new destinations have emerged in addition to the traditional favorites of Europe and North America.

Despite occasional shocks, international tourist arrivals have shown virtually uninterrupted

Tourism, Fig. 1 Inbound tourism by purpose of visit, 2013 (share) (Source: World Tourism Organization)



growth from 25 million in 1950 to 278 million in 1980, 528 million in 1995, and a record of 1087 million arrivals in 2013 (UNWTO 2015).

Nowadays, tourism is considered a popular global leisure activity.

Most travel is for leisure purposes (52%), followed by health and religion purposes (27%), business (14%), and others (Fig. 1).

According to the UNWTO's long-term forecast *Tourism Towards 2030*, international tourist arrivals worldwide are expected to increase by 3.3% a year from 2014 to 2030 to reach 1.8 billion by 2030.

The developments in technology and in transport infrastructure, such as low-cost airlines and more accessible airports, have also made many types of tourism more affordable. Internet sales of tourist services have facilitated changes in lifestyle. Some sites have started to offer dynamic packaging, in which an inclusive price is offered for an appropriate package requested by the customer (see paragraph 4.1 on ICT and tourism distribution).

Tourism Data and Statistics

Secondary Data and Statistics

Worldwide Tourism Statistics

The United Nations World Tourism Organization (UNWTO) *Tourism Highlights* is a very useful source for a long-term outlook on tourism. The

UNWTO is the United Nations' agency responsible for the promotion of responsible, sustainable, and universally accessible tourism. As the leading international organization in the field of tourism, UNWTO promotes tourism as a driver of economic growth, inclusive development, and environmental sustainability and offers leadership and support to the sector in advancing knowledge and tourism policies worldwide. It is the leading organization collecting and disseminating the most up-to-date and comprehensive tourism data, short- and long-term forecasts, and knowledge on specific and source markets; thus it serves as a global forum for tourism policy and a source of tourism know-how.

The *UNWTO Tourism Towards 2030* is UNWTO's long-term outlook and assessment of the development of tourism for the two decades from 2010 to 2030, and it has become a worldwide reference for international tourism forecasts.

Understanding, for each country, where its inbound tourism is generated is essential for analyzing international tourism flows and devising marketing strategies, such as those related to the positioning of national markets abroad.

UNWTO's main dataset and publication on annual tourism statistics include:

- *Yearbook of Tourism Statistics*, which, deriving from the most comprehensive statistical database available on the tourism sector, focuses on data related to inbound tourism

(total arrivals and overnight stays), broken down by the country of origin

– *Compendium of Tourism Statistics*

Complete data and statistics for European countries are also available from the Eurostat database.

National Institutes of Statistics

Additional statistics can be found in publications on tourism made available by the National Institute of Statistics in each country.

The United Nations General Assembly has endorsed the *Fundamental Principles of Official Statistics* (last 29 January 2014). These principles are considered a basic framework which all statistical activities developed by national and international organizations must follow in recognizing official statistics as a public good.

Tourism Satellite Account (TSA)

The *Tourism Satellite Account* (described in the *Tourism Satellite Account: Recommended Methodological Framework 2008*) is, besides the *International Recommendations for Tourism Statistics 2008*, the second international recommendation on tourism statistics that has been developed in a framework of consistency with the System of National Accounts. Both recommendations are mutually consistent and provide the conceptual framework for measuring and analyzing tourism as an economic activity in each country.

As a statistical tool for the economic accounting of tourism, the TSA can be seen as a set of ten summary tables, each with their underlying data and representing a different aspect of the economic data relative to tourism: inbound domestic tourism and outbound tourism expenditure, internal tourism expenditure, production accounts of tourism industries, and the gross value added (GVA) and gross domestic product (GDP) attributable to tourism demand, employment, investment, government consumption, and non-monetary indicators.

The purpose of a Tourism Satellite Account is to analyze in detail all the aspects of demand for goods and services associated with the activity of visitors, to observe the operational interface with

the supply of such goods and services within the economy, and to describe how the supply interacts with other economic activities. It should permit greater internal consistency of tourism statistics with the rest of the statistical system of a country as well as increased international comparability of these data. TSA is highly recommended to evaluate the role that tourism sector performs in the entire economy as well as to allow processing and comparison at an international level.

Primary Data: Surveys

Where the required secondary data support is unavailable for an economic impact analysis, primary data collection from survey sampling is probably the only way forward. This approach can provide useful information where more sophisticated methods are not applicable, but the results should be used with caution due to sampling biases. Careful statistical treatments with the raw data are also needed.

International Tourism: An Overview

In 1994, the United Nations classified three forms of tourism in its *Recommendations on Tourism Statistics*:

- Domestic tourism, involving residents of the given country traveling only within this country
- Inbound tourism, involving nonresidents traveling in the given country
- Outbound tourism, involving residents traveling in another country

International tourist arrivals (overnight visitors) grew by 5% in 2013, reaching a record 1087 million arrivals worldwide, up from 1035 million in 2012.

International tourism receipts are the earnings generated in destination countries from expenditure on accommodation, food and drink, local transport, entertainment, shopping, and other services and goods. In macroeconomic terms, expenditure by international visitors counts as exports for the destination country and as imports for the

country of residency of the visitor. In 2013, international tourism receipts in destinations around the world grew 5% in real terms (taking into account exchange rate fluctuations and inflation) to reach US\$ 1159 billion (euro 873 bn). The growth in receipts mirrored the growth in international arrivals (also +5%), confirming the strong correlation between these two key indicators of international tourism.

Table 1 reports the most visited countries in 2013 in terms of the number of international tourist arrivals. Table 2 reports the top ten tourism earner countries for the year 2013.

Tourism, Table 1 International tourist arrivals

Rank	Series ¹	Million		Change (%)	
		2012	2013*	12/11	13*/12
1 France	TF	83.0	..	1.8	..
2 United States	TF	66.7	69.8	6.3	4.7
3 Spain	TF	57.5	60.7	2.3	5.6
4 China	TF	57.7	55.7	0.3	−3.5
5 Italy	TF	46.4	47.7	0.5	2.9
6 Turkey	TF	35.7	37.8	3.0	5.9
7 Germany	TCE	30.4	31.5	7.3	3.7
8 United Kingdom	TF	29.3	31.2	−0.1	6.4
9 Russian Federation	TF	25.7	28.4	13.5	10.2
10 Thailand	TF	22.4	26.5	16.2	18.8

¹TF International tourist arrivals at frontiers (excluding same-day visitors); TCE International tourist arrivals at collective tourism establishments.

Source: UNWTO (2014) Tourism highlights

In the ranking by arrivals, Europe leads the growth in absolute terms, while Asia and the Pacific record the fastest relative growth across all the regions. France continues to top the ranking of international tourist arrivals and is third in international tourism receipts. The United States ranks first in receipts and second in arrivals. Spain is still the second largest earner worldwide and the first in Europe and ranks third in arrivals. China moves to fourth in arrivals and remains fourth in receipts. Italy consolidates its fifth place in arrivals and sixth in receipts. Turkey remains sixth in arrivals and 12th in receipts. Thailand enters the top ten arrivals ranking at number ten, climbing amazing five positions, while it moves up two places to seventh in the ranking by tourism receipts. Germany and the United Kingdom remain, respectively, seventh and eighth in arrivals, but move down one place each in terms of earnings to eighth and ninth places, respectively. The Russian Federation completes the top ten ranking by arrivals in ninth place, while the two Chinese Special Administrative Regions Macao and Hong Kong rank, respectively, fifth and tenth in receipts (UNWTO *Tourism Highlights* 2014).

World's Top Tourism Destinations

When ranking the world's top international tourism destinations, it is preferable to take more than a single indicator into account. Ranked according to the two key tourism indicators – international tourist arrivals (Table 1) and international tourism

Tourism, Table 2 International tourist receipts

Rank	US\$				Local currencies	
	Billion		Change (%)		Change (%)	
	2012	2013*	12/11	13*/12	12/11	13*/12
1 United States	126.2	139.6	9.2	10.6	9.2	10.6
2 Spain	56.3	60.4	−6.3	7.4	1.5	3.9
3 France	53.6	56.1	−2.2	4.8	6.0	1.3
4 China	50.0	51.7	3.2	3.3	0.8	1.4
5 Macao (China)	43.7	51.6	13.7	18.1	13.2	18.1
6 Italy	41.2	43.9	−4.2	6.6	3.8	3.1
7 Thailand	33.8	42.1	24.4	24.4	26.7	23.1
8 Germany	38.1	41.2	−1.9	8.1	6.3	4.5
9 United Kingdom	36.2	40.6	3.3	12.1	4.8	13.2
10 Hong Kong (China)	33.1	38.9	16.2	17.7	15.8	17.7

Source: UNWTO (2014) Tourism highlights

receipts (Table 2) – it is interesting to note that eight of the top ten destinations appear on both lists, despite showing marked differences in terms of the type of tourists they attract and in average length of stay and spending per trip and per night. In the case of international tourism receipts, changes not only reflect relative performance but also (to a considerable extent) exchange rate fluctuations between national currencies and the US dollar.

The World Tourism Organization also ranks countries on the base of their total expenditure on international tourism for the year 2013. Table 3 reports the top ten countries.

Tourism Distribution Channels

Two of the most important factors influencing the competitiveness and success of a tourist destination are the efficiency and effectiveness of distribution channels (Buhalis and Laws 2001).

Tour operators play a significant role in the promotion and marketing of tourist products. This is particularly true for small and medium accommodation structures and emerging regions which do not have sufficient expertise and financial resources to invest in advertising and in the widespread distribution of their products (Buhalis 2000; Succurro 2008).

The negotiation between tourist firms and intermediaries, primarily aimed at increasing

visibility and competitiveness of services supplied, is relevant to the success of both the tourist firms and the tourist destinations as a whole. Moreover, an appropriate promotion and distribution of products have direct consequences on a balanced use of accommodation establishments. High seasonality, indeed, is one of the most problematic aspects in the tourist sector; many destinations suffer from this phenomenon every year.

Both the seasonality problem and the relative importance of traditional tour operators have been strongly affected by the new technologies. Reduced search costs and direct online organization of the trip are two prominent aspects of the tourism industry that have been deeply affected by recent technological advances. Tourism providers, indeed, can now sell their services directly through the web by avoiding intermediaries. Recent studies explore the capacity management issue under time-varying demand (i.e., the seasonality issue), the tourist information acquisition process, and, finally, the impact of online booking on tourism flows and seasonality (Jang 2004; Boffa and Succurro 2012).

ICT and Tourism Distribution

Over the last 20 years, the Internet has changed various facets of social life, creating many social

Tourism, Table 3 Top ten biggest spenders, 2013

Rank	International tourism expenditure (US\$ billion)		Local currencies change (%)		Market share (%)	Population (million)	Expenditure per capita (US\$)
	2012	2013*	12/11	13*/12	2013*	2013	2013*
1 China	102.0	128.6	37.3	23.8	11.1	1,361	94
2 United States	83.5	86.2	6.7	3.3	7.4	316	273
3 Germany	81.3	85.9	2.5	2.3	7.4	81	1,063
4 Russian Federation	42.8	53.5	36.5	28.9	4.6	143	374
5 United Kingdom	51.3	52.6	2.1	3.5	4.5	64	821
6 France	39.1	42.4	−5.8	4.9	3.7	64	665
7 Canada	35.0	35.2	6.2	3.2	3.0	35	1,002
8 Australia	28.0	28.4	2.1	8.8	2.4	23	1,223
9 Italy	26.4	27.0	−0.3	−1.0	2.3	60	452
10 Brazil	22.2	25.1	4.6	12.9	2.2	198	127

Source: UNWTO (2014) Tourism highlights

concerns (Kim 2010). A large diffusion of the ICT in the tourism sector has improved its social and economic impacts, from which many consumers and organizations can benefit (Minghetti and Buhalis 2010). Indeed, the Internet has grown to be one of the most effective means for tourists to seek information and purchase tourism-related products (Pan and Fesenmaier 2006).

The Internet plays a key role in the development of the tourism industry since it encourages people to travel both by improving access to the destinations and by reducing search costs (Boffa and Succurro 2012). With the advent of e-commerce, indeed, tourism products have become one of the most traded online items.

Technological progress, coupled with regulatory changes, has modified the nature of tourists' search process in at least two directions. First, it has expanded consumption opportunities (e.g., by decreasing the cost of reaching relatively far destination) and, as a result, the expected benefit of searching for a tourist destination. Second, it has decreased the costs of direct search, for example, through the release of faster and more reliable search tools, thanks to the Internet. The two effects contribute to making the direct online search process less time-consuming and more valuable, hence more productive.

Tourism Research

Despite the debate about its definition over the past decades, tourism is commonly recognized as a human activity which can be analyzed from different perspectives. As a field of study, tourism is gradually evolving from a multidisciplinary endeavor into an interdisciplinary stage of research (Tribe and Xiao 2011). Thus, in order to advance the understanding of tourism, it is necessary to integrate economics with other social sciences including law, psychology, sociology, and political science and to seek new holistic approaches and tools.

Numerous disciplinary contributions in diverse areas of research have supported the emergence of tourism as a field of academic study or an autonomous discipline. The epistemology of tourism

research, however, is still the subject of ongoing discussion and debate (Benckendorff and Zehrer 2013).

Economic Analysis

In the wider context of tourism knowledge creation, economics has played a significant role.

Tourism generates directly and indirectly an increase in economic activity in the destinations visited, mainly due to demand for goods and services that need to be produced and provided.

In the economic analysis of tourism, one may distinguish between tourism's economic *contribution* which refers to the direct effect of tourism and is measurable through the Tourism Satellite Account (TSA) and tourism's economic *impact* which is a much broader concept encapsulating the direct, indirect, and induced effects of tourism which must be estimated by applying models to provide the simulations (Dwyer et al. 2007; Song et al. 2012). Thus, economic impact studies aim to quantify economic benefits, that is, the net increase in the wealth of residents resulting from tourism, measured in monetary terms, over and above the levels that would prevail in its absence.

Because of the evolution of tourism as an economic activity over the past 50 years, there has been a significant growth of publications in specialized journals (*Annals of Tourism Research*, *Tourism Economics*, *Tourism Management*, *Journal of Travel Research*, to cite some of them) and several key texts on tourism economics (more recently, Dwyer et al. 2010; Stabler et al. 2010; Tribe 2011). Several scientific articles have also attempted to provide an overview of the developments in tourism economics (see, as relatively more recent works, Eadington and Redman 1991; Sinclair 1998; Tremblay 1998; Sinclair et al. 2003; Dwyer et al. 2011).

Macroeconomic Analysis

From a macroeconomic perspective, tourism contributes to both destination competitiveness – defined in different ways and measured by different methodologies – and local, national, and international economic development. Over the last few decades, indeed, a popular topic in tourism research has been the evaluation of the

economic, social, and environmental impact of tourism and its policy implications.

With reference to the economic impact, proponents of the tourism-led growth hypothesis focus on the relationship between tourism and economic growth and emphasize the role of tourism in spurring local investments, exploiting economies of scale, increasing employment, and diffusing technical knowledge (Schubert et al. 2011). The employment effect of tourism, the quality and structure of employment, and the gender wage gap are other well-established and interdisciplinary research areas which also draw on insights from sociology and political science.

With reference to the tourism-led theory, many studies confirm a unidirectional causality from international tourism to real GDP in specific countries or regions, while other studies find evidence of bidirectional relationships. The main criticism faced by the tourism-led growth studies relates to their reliance on the use of a methodology – the Granger causality test – which does not necessarily suggest the real cause-effect relationship (Song et al. 2012).

Note that recent research stresses that tourism does not always increase economic welfare. In fact, a tourism boom may lead to “deindustrialization” in other sectors due to a phenomenon known as the “Dutch Disease effect.”

Moreover, as a significant form of international trade flows, tourism also lies within the scope of international economics studies. A number of studies find supportive evidence of the bidirectional causality between international tourism and international trade. Also these studies usually rely on the Granger causality test with the subsequent aforementioned criticism.

With reference to the environmental issues, tourism production and consumption generate environmental consequences, and at the same time, tourism activities are strongly affected by the quality of the environmental resources. In the tourism industry, and differently from the manufacturing industries, the environment is not only an input factor but also a key component of its output (Razumova et al. 2009). For this reason, an increasing attention has been paid to sustainable tourism and climate change topics – both at

the micro- and macrolevel of analysis – and to discussions about the appropriate instruments for environmental governance. Several price-based instruments, semi-price instruments such as quotas, and non-price instruments, such as government regulations and industry voluntary management, have been proposed and discussed theoretically in the tourism literature.

Microeconomic Analysis

From a microeconomic perspective, economic studies are complex and cover several topics. In the recent tourism economic literature, departing from the old debate about whether tourism is an industry or a market, it has been commonly recognized that tourism cannot be defined as either an industry or a market. Clarification of this confusion has important implications for economic analysis in this field (Wilson 1998). More specifically, “. . . tourism is a composite product that involves a combination of a variety of goods and services provided by different sectors, such as transport, accommodation, tour operators, travel agencies, visitor attractions, and retailing. Moreover, tourism products are serviced and transacted in different markets” (Song et al. 2012). For this reason, diverse topics have been developed from both the demand and supply perspectives.

Demand Analysis The dominant position of demand analysis and its determinants is still observable in the latest developments in both theoretical and empirical tourism economic studies. Tourism demand is predominantly measured by the number of arrivals, by the level of tourist expenditure (receipts), and, more recently, by the number of tourist nights (length of stay). The key research tasks in tourism demand studies include the selection of the best specified models for modeling and forecasting tourism demand, the identification of the key economic determinants of tourism demand, and the computation of demand elasticities and, associated with globalization, market interdependence.

Supply Analysis The tourism supply analysis usually follows that of industrial economics. The well-known structure-conduct-performance (SCP)

paradigm has provided a useful framework for studying tourism supply from a market perspective. The SCP paradigm suggests that the type of the market structure within which a firm operates (e.g., monopoly, monopolistic competition, oligopoly) determines a firm's conduct (e.g., pricing strategies, marketing investment, innovation) which ultimately affects its overall performance (mainly productivity, efficiency, and long-term growth). A number of empirical studies have tested the SCP paradigm in tourism, but the findings are inconclusive due largely to the different empirical settings and methods used. Newer approaches take into account the dynamic nature of the market – on the base of game theory approach – and its institutional arrangements.

In addition to the above topics, industry agglomeration and clustering, with the economic importance of geographic location, have become an emerging topic in recent tourism supply studies. The new geographical economics, indeed, provides useful perspective for interfirm relationships.

Adjectival Tourism

Many niche or specialty travel forms of tourism have come into a common use by the tourism industry and academics. Among the various forms of tourism: individual tourism, collective tourism, organized tourism, educational tourism, young tourism, third-age tourism, business tourism, sustainable tourism, and ecotourism.

Other forms of “adjectival tourism” include agritourism, birth tourism, culinary tourism, cultural tourism, extreme tourism, geotourism, heritage tourism, medical tourism, nautical tourism, religious tourism, sex tourism, and wildlife tourism.

Sustainable Tourism

Tourism's rapid growth calls for a greater commitment to the principles of sustainability to amplify tourism's benefits and mitigate its possibly negative impacts on the environment and on societies.

The key issues in sustainable tourism, indeed, are defined by the fundamentals of sustainability, external to the literature of tourism research, and linked to science, environment, resource management, global change, human health, economics, and development policy (Buckley 2012).

From a more general perspective, indeed, and as reported by the World Commission on Environment and Development (officially dissolved in December 1987), sustainable development implies “meeting the needs of the present without compromising the ability of future generations to meet their own needs.” Thus, sustainable development would at least maintain ecological integrity and diversity to meet human needs.

Tourism researchers turned their attention to social and environmental issues almost four decades ago. Research using the specific term *sustainable tourism* started around two decades ago.

Specifically, “sustainable tourism is envisaged as leading to management of all resources in such a way that economic, social and aesthetic needs can be fulfilled while maintaining cultural integrity, essential ecological processes, biological diversity and life support systems”.

Sustainable tourism can be seen as having regard to ecological and sociocultural carrying capacities and includes involving the community of the destination in tourism development planning. It also involves integrating tourism to match current economic and growth policies so as to mitigate some of the negative economic and social impacts of mass tourism.

There is a myriad of definitions for sustainable tourism, including ecotourism, green travel, environmentally and culturally responsible tourism, fair trade, and ethical travel.

Ecotourism

Ecotourism, also known as ecological tourism, is responsible travel to fragile and usually protected areas that strives to be low impact and (often) small scale. It helps educate the traveler, provides funds for conservation, directly benefits the economic development and political empowerment of local communities, and fosters respect for different cultures and for human rights.

Summary/Conclusion/Future Directions

Economics has played a significant role in studying tourism, considered as a key driver of socio-economic progress. In the wider context of tourism knowledge creation, however, there is a need for a holistic approach to tourism analysis. Thus, it would be desirable to integrate economics with other social science disciplines.

References

- Benckendorff P, Zehrer A (2013) A network analysis of tourism research. *Ann Tour Res* 43:121–149
- Boffa F, Succurro M (2012) The impact of search cost reduction on seasonality. *Ann Tour Res* 39:1176–1198
- Buckley R (2012) Sustainable tourism: research and reality. *Ann Tour Res* 39:528–546
- Buhalis D (2000) Relationships in the distribution channel of tourism: conflicts between hoteliers and tour operators in the Mediterranean region. *Int J Hosp Tour Adm* 1:113–139
- Buhalis D, Laws E (2001) *Tourism distribution channels. Practices, issues and transformations*. Continuum, London
- Dwyer L, Forsyth P, Spurr R (2007) Contrasting the uses of TSAs and CGE models: measuring tourism yield and productivity. *Tour Econ* 13:537–551
- Dwyer L, Forsyth P, Dwyer W (2010) *Tourism economics and policy*. Channel View Publications, Bristol
- Dwyer L, Forsyth P, Papatheodorou A (2011) Economics of tourism. In: Cooper C (ed) *Contemporary tourism reviews*. Goodfellow Publishers, Oxford, pp 1–29
- Eadington WR, Redman M (1991) Economics and tourism. *Ann Tour Res* 18:41–56
- Jang SS (2004) Mitigating tourism seasonality. A quantitative approach. *Ann Tour Res* 31:819–836
- Kim S (2010) The diffusion of the internet: trend and causes. *Soc Sci Res* 40:602–613
- Manila Declaration on World Tourism. The world tourism conference, Manila, 27 Sept–10 Oct 1980, pp 1–4
- Minghetti V, Buhalis D (2010) Digital divide in tourism. *J Travel Res* 49:267–281
- Pan B, Fesenmaier DR (2006) Online information search. Vacation planning process. *Ann Tour Res* 33:809–832
- Razumova M, Lozano J, Rey-Maqueira J (2009) Is environmental regulation harmful for competitiveness? The applicability of the Porter hypothesis to tourism. *Tour Anal* 14:387–400
- Recommendations on Tourism Statistics (1994) Statistical papers, no.83, M. United Nations, New York, p 5. Retrieved 12 July 2010
- Schubert SF, Brida JG, Risso WA (2011) The impacts of international tourism demand on economic growth of small economies dependent on tourism. *Tour Manage* 32:377–385
- Sinclair MT (1998) Tourism and economic development: a survey. *J Dev Stud* 34:1–15
- Sinclair MT, Blake A, Sugiyarto G (2003) The economics of tourism. In: Cooper C (ed) *Classic reviews in tourism*. Channel View Publications, Clevedon, pp 22–54
- Song H, Dwyer L, Li G, Cao Z (2012) Tourism economics research: a review and assessment. *Ann Tour Res* 39:1653–1682
- Stabler MJ, Papatheodorou A, Sinclair MT (2010) *The economics of tourism*, 2nd edn. Routledge, Abingdon
- Succurro M (2008) The role of intermediaries in the growth of a lesser developed region: some empirical evidence from Calabria, Italy. *Tour Econ* 14:393–407
- Tremblay P (1998) The economic organization of tourism. *Ann Tour Res* 25:837–859
- Tribe J (2011) *The economics of recreation, leisure and tourism*. Butterworth-Heinemann, Oxford
- Tribe J, Xiao H (2011) Developments in tourism social science. *Ann Tour Res* 38:7–26
- UNWTO (2015) *Tourism highlights, 2015 Edition*. Madrid, Spain. Available at www.unwto.org/pub
- Wilson K (1998) Market/Industry confusion in tourism economic analyses. *Ann Tour Res* 25:803–817

Tournament Theory

Martin Schneider

Faculty for Economics and Business
Administration, University of Paderborn,
Paderborn, Germany

Abstract

In tournament theory the effects of competitions in which the best performers are awarded a fixed prize are studied. The tournament idea has been used to explain career patterns in large US law firms and in European judicial hierarchies. It has also been suggested in a prescriptive way as a method to select judges for the US Supreme Court. Tournaments theory helps to understand under which conditions lawyers and judges engage in a rate race to achieve promotion. But important assumptions of the formal tournament models are not met in practice, so real tournaments are unlikely occur in practice. The theory should therefore not be interpreted as an exact descriptive or prescriptive model of behavior but rather as a useful metaphor to help understand empirical patterns.

Synonyms

Contest theory

Fundamentals

Tournament theory utilizes the metaphor of a sport tournament to analyze certain types of competition between players more generally. In a tournament, the prize of a competition is awarded to a fraction of all competitors with the best performance. Hence, a tournament is distinct from other types of competition by two characteristics. First, the winner prize and the fraction of winners are specified in advance of the competition. Second, the prize is awarded not on the basis of some absolute performance standard but on the basis of relative performance (rank-order tournament).

The idea of a tournament has been suggested for the first time within labor economics and human resource management to describe competition among employees for promotion in an organizational hierarchy (Lazear and Rosen 1981; Rosenbaum 1979, 1984). Since then the tournament idea has been most widely applied to explain the pay structure within companies. It has also been adopted in a range of disciplines including law (more on this below), ecology, finance, and psychology (Connelly et al. 2014).

The basic formal model (Lazear and Rosen 1981) is a game with two identical risk-neutral agents and one principal. The principal seeks to elicit effort from the two agents by setting a tournament prize for the winner, i.e., for the agent with the highest output (output can be measured, effort cannot). Each agent maximizes utility by selecting the level of effort, given the impact of effort on the probability of winning, the prize, and the disutility from effort. At the utility-maximizing level of effort, the marginal disutility from effort is equal to the marginal increase in the probability of winning times the prize. Since agents are identical and the probability of winning depends on the other agent's effort, the tournament incites a rat race among the agents. None of the agents has an incentive to reduce effort because effort reduction implies that the other agent will win. The principal

can implement the optimal amount of effort by manipulating the prize, i.e., the difference between the winning and the losing pay level.

In terms of incentives effects, this basic model has two main testable hypotheses (Connelly et al. 2014). First, agents' effort increases with the prize, i.e., with the difference between the winning and losing pay level. Second, it is only this prize not the absolute winning pay level that matters for agents' effort.

The basic formal model has been extended in various directions (Connelly et al. 2014). There may be a succession of tournaments at different levels of an organization rather than a single competition. More than two players may compete. Agents may be heterogeneous. They may be able to influence each other's output (e.g., through sabotage). And output may be influenced by luck to varying degrees.

While the basic formal model focuses on deriving a prize that provides an optimal level of incentives, tournaments may also be implemented for other reasons, namely, to select certain players and to commit the principal to a certain policy (pay, promotion) in advance.

Tournaments of Lawyers and Judges

From a law-and-economics perspective, applications of the tournament idea on judiciaries and law firms are particularly interesting. It has been suggested that certain ("elite") law firms in the USA have implemented more or less rigid tournament models in the early twentieth century (Galanter 1994; Galanter and Palay 1990). In essence, these firms have committed to promoting a certain percentage of associates to partners each year, thus incentivizing younger associates and building up pressure in the firm as a whole to maintain revenues. This tournament, it has been argued, can explain the growth of these law firms in the twentieth century. In the big law firms, competition for partnerships is now restricted to the higher ranks ("elastic tournaments") (Galanter and Henderson 2008). Tournament theory and the idea of up-or-out competition have become the main theoretical perspective to

explain the labor market for lawyers in the large law firms in the USA and elsewhere (Wilkins and Gulati 1998).

Tournament theory has also been used to examine competition among judges for assignment or promotion. Judiciaries in civil law countries such as Germany resemble internal labor markets. Judges enter at the bottom of the hierarchy, they are tenured, and they may be promoted to higher judicial positions (Schneider 2004, 2005). The way judges are selected, the fact that pay is attached to judicial positions and observed data on promotions are in line with the idea that a succession of rank-order tournaments takes place between judges. While the tournament model was used here in a descriptive and explanatory way, another application is prescriptive. In order to avoid politicized nomination, it has been suggested that a tournament might be an appropriate method to select judges for the US Supreme Court (Choi and Gulati 2004a). As a possible criterion for promotion, a rank order of judges has been compiled based on various output measures (Choi and Gulati 2004b).

Why Tournaments?

Tournaments may be considered an efficient incentive mechanism. In contrast to an obvious alternative, namely, paying workers directly according to productivity, the measurement of output is easier. This is because, to determine the winner in a tournament, output needs not be measured continuously; an ordinal measure – a rank order – is sufficient. For both lawyers in law firms and for judges, the provision of incentives has been suggested a reason for the practice of tournament-like promotions (Schneider 2005; Choi and Gulati 2004a; Galanter and Palay 1990).

In addition, tournaments may exert favorable selection effects. If productivity in the current job predicts productivity in a job higher up in the hierarchy, then tournaments may help to identify the best lawyers and judges for more important positions in the law firm and the judiciary. This

selection effect of a tournament was considered particularly attractive for judicial positions. Here, a tournament promises to be transparent and objective, while alternative mechanisms are plagued with political bickering and tend to bring to office judges based on their political leanings (Choi and Gulati 2004a).

Tournaments, finally, are also a commitment device. Because prize, the rate of winners, and the output measure need to be specified in advance and are observed easily, workers may trust that effort will be rewarded and the employer will not renege on her promise. This argument has been suggested as an important reason for the big law firms to employ tournaments (Galanter and Palay 1990).

Problems of Tournaments

These potentially positive effects of tournaments are diluted by the various problems that tournaments meet (or would meet) in practice. Even in the large private firms, the first application of tournament theory, promotions are governed in many heterogeneous ways that do not fully comply with the tournament model (Gibbs 1994). A similar statement has been made for large US law firms (Wilkins and Gulati 1998).

Many of the assumptions of the tournament idea are not and cannot be fully met. For example, the rate of winners cannot be specified in advance in judicial hierarchies because the vacancies are often determined by the case load, organizational demography, and public budgets. Hence, there is clearly no commitment to a tournament in judicial hierarchies. Similarly, the assumption that competitors do not know each other's performance is hardly met in practice, which in turn reduces the incentive effect.

A number of important problems concern the measurement of output. Competition for promotions can only incentivize and select efficiently if output is measured appropriately. For example, in response to suggesting a tournament of judges for the Supreme Court, it has been argued that output measures referring to the publication and influence of decisions do not fully capture the "virtue" of a good judge (Solum 2004) and are plagued

with inaccuracy (Levy et al. 2010). Similar concerns have been raised with regard to measuring the productivity of associates in law firms (Wilkins and Gulati 1998).

Conclusion

Tournament theory offers an interesting perspective on the competitions between lawyers in large law firms and between judges in judicial hierarchies. The theory predicts that lawyers and judges may sometimes engage in a rat race and that the winners tend to be better than the losers in measured dimensions of productivity. However, the assumptions of a pure rank-order tournament are never fully met in real law firms and real judiciaries. Therefore, tournament theory should not be taken as a descriptive or prescriptive “model” but rather a “metaphor” to help understand empirical patterns (Wilkins and Gulati 1998).

Cross-References

- [German Law System](#)
- [Intrinsic and Extrinsic Motivation](#)

References

- Choi S, Gulati M (2004a) A tournament of judges? *Calif Law Rev* 92:299–322
- Choi SJ, Gulati GM (2004b) Choosing the next Supreme Court justice: an empirical ranking of judge performance. *South Calif Law Rev* 78:23–117
- Connelly BL, Tihanyi L, Crook TR, Gangloff KA (2014) Tournament theory: thirty years of contests and competitions. *J Manag* 40:16–47
- Galanter M (1994) *Tournament of lawyers: the transformation of the big law firm*. University of Chicago Press, Chicago
- Galanter M, Henderson W (2008) The elastic tournament: a second transformation of the big law firm. *Stanford Law Rev* 60:1867–1929
- Galanter M, Palay TM (1990) Why the big get bigger: the promotion-to-partner tournament and the growth of large law firms. *Va Law Rev* 76:747–811
- Gibbs M (1994) Testing tournaments? An appraisal of the theory and evidence. *Labor Law J* 45:493–500
- Lazear EP, Rosen S (1981) Rank-order tournaments as optimum labor contracts. *J Polit Econ* 89:841–864
- Levy MK, Stith K, Cabranes JA (2010) The costs of judging judges by the numbers. *Yale Law Policy Rev* 28:313–323
- Rosenbaum JE (1979) Tournament mobility: career patterns in a corporation. *Adm Sci Q* 24:220–241
- Rosenbaum JE (1984) *Career mobility in a corporate hierarchy*. Academic, New York
- Schneider M (2004) Careers in a judicial hierarchy. *Int J Manpow* 25:431–446
- Schneider MR (2005) Judicial career incentives and court performance: an empirical study of the German labour courts of appeal. *Eur J Law Econ* 20:127–144
- Solum LB (2004) A tournament of virtue. *Florida State Univ Rev* 32:1365–1400
- Wilkins DB, Gulati GM (1998) Reconceiving the tournament of lawyers: tracking, seeding, and information control in the internal labor markets of elite law firms. *Va Law Rev* 84:1581–1681

Traceability

Christophe Charlier

Department of Economics, Université Côte d’Azur CNRS, GREDEG, Nice, France

Definition

“The ability to trace the history, application or location of an entity by means of recorded identifications” in “ISO 8402:1994 Quality management and quality assurance – Vocabulary.”

Introduction

A general definition of traceability is found in “ISO 8402:1994 Quality management and quality assurance – Vocabulary” as “The ability to trace the history, application or location of an entity by means of recorded identifications.” Traceability is a typical example of a voluntary management practice that predated food safety regulation that ends up entering sanitary public policy. Voluntarily used at the beginning by few operators in high-quality food chains only, traceability has become a mandatory risk management practice for all food operators in Europe. However, legal forms (voluntary vs. mandatory), as well as exigencies of the various traceability systems implemented

around the world, remain very different, even for a same product (e.g., see Schroeder and Tonsor 2012 for bovine traceability).

Traceability is considered from two broad perspectives in economic literature. First, it has been analyzed from firms' standpoint. The determinants of traceability adoption have been explored in this field. The link between traceability adoption and firm characteristics such as their complexity, their hierarchical structure, and the kind of relations they have with downstream suppliers has been put forward (Galliano and Orozco 2011, 2013). The aim of operating under a private standard (that requires traceability) or under a quality label such as a geographical indication has also been presented as an important adoption determinant (Souza Monteiro and Caswell 2009).

Second, traceability has been analyzed from the social standpoint. This part of the literature is driven by the aim of establishing economic rationale for traceability. Whether mandatory or not, traceability is justified by information asymmetry among food business operators along the food chain and among food producers and consumers. This information problem arises because of the "credence good" attribute of food safety or quality. As the nature of this information problem is multidimensional, traceability may be more or less demanding. The economic literature agrees with the idea that traceability may be defined according to three characteristics (Golan et al. 2004): its breadth, defined as the amount of information delivered by the system (i.e., the variety of the items recorded); its depth, characterized by how far back in the supply chain information record is made; and its precision, generated by the tracking unit used. Depending on the "values" chosen for these characteristics, different traceability systems are used.

Whatever its form, traceability cannot in itself insure food safety or food quality. It only gives information. However, by delivering information, traceability performs three functions that impact on food safety (Hobbs 2004) and the way food safety crisis may be managed. First, the possibility of tracking back the origin of food through the supply chain allows for ex post efficient management of food safety crisis (withdrawal of unsafe

products from the supply chain) and implies cost reduction as a consequence (health cost, lost productivity, lower product sales, etc.). Second, breaking points along the supply chain and responsible operators are more easily identifiable with traceability, so that liability can be established. Operators are therefore more inclined to comply with regulatory standards. Finally, when coupled with labeling, traceability may convey information to consumers on quality or safety attributes. These links between traceability and food safety may seem obvious. However, in the aftermath of the bovine spongiform encephalopathy crisis, when traceability emerged as a risk management tool, they were debated in the Committee on General Principles of the Codex Alimentarius. Unsurprisingly, traceability has not entered national food safety legislations in the same way. For example, traceability has remained limited to certain products in the USA, whereas it has become mandatory for all food and foodstuff in the EU.

This entry considers the literature dealing with the "social standpoint" on traceability. It first focuses on the European Regulation making traceability mandatory. It then examines voluntary traceability through the two main economic reasons previously underlined: incentives and liability concerns.

Mandatory Versus Voluntary Traceability

The Regulation (EC) No. 178/2002 considers traceability as a risk management tool enabling "accurate and targeted withdrawal" of products from the food chain. The Regulation has a broad scope of application since every stage of the food chain is concerned, demanding that all food products and foodstuffs must be tracked. Breadth and depth of such a system are therefore maximal. However, the Regulation only requires operators to be able to identify "the business from which the food, feed, animal or substance that may be incorporated into food or feed has been supplied." Furthermore, no requirement on batch formation appears. The text implicitly favors a formation of

batches according to the identity to whom the products are sold without taking into account their homogeneity in terms of their inputs' origins. The precision of the mandatory traceability implemented is thus very weak. Charlier and Valceschini (2008) show that with these characteristics, the traceability is constructed "step by step" without requiring a compilation of data that would grow along the food supply chain until the good is sold in the market. The authors show that this traceability system alone may hardly reach the goal assigned by the regulation of a targeted withdrawal of products mainly because of the lack of discipline on the products' batch formation.

Together with traceability, the Regulation (EC) No. 178/2002 stipulates operator's responsibilities on food safety procedures concerning food withdrawal and obligations to inform and cooperate with public authorities in case of a sanitary disruption. Interestingly, these obligations and the proactive behavior required on product withdrawal call for a more stringent traceability than the mandatory one. The implementation of this traceability is therefore implicitly left to operators that have to coordinate their efforts (adopting common rules on the constitution of batches) so that the traceability system adopted by downstream operators does not scramble information produced by traceability efforts of upstream operators (Charlier and Valceschini 2008). This individual choice to implement stricter traceability responds to incentives created by regulatory disposition on operator's responsibilities. Not surprisingly, incentives and liability have been found in the economic literature as the two main drivers for voluntary adoption of traceability.

Incentives and Liability

A significant part of the literature has dealt with traceability through the lens of incentives. In this field, an important characteristic of traceability is that this information system allows allocating the cost of food safety failures to the responsible parties. To do so, traceability is always implemented jointly with another device (e.g., inspection procedures) allowing detecting

food safety failures (Starbird and Amanor-Boadu 2006, 2007). Relations among suppliers and buyers of foodstuff are seen as taking place in context of asymmetric information and are generally represented using a principal-agent framework (Resende-Filho and Buhr 2008). The models developed delineate the conditions under which traceability provides incentives to suppliers to deliver safe products. The results depend crucially on various variables such as the food safety failure cost, the production cost of a safe product, the inspection cost, and the cost of allocating failure responsibility. Traceability in this context may be modeled differently. It may be characterized by a "cost allocation factor," i.e., "the proportion of food safety failure costs that can be allocated to the producer responsible for unsafe food" (Starbird and Amanor-Boadu 2007), or by the "probability of finding the source of a problem" (Resende-Filho and Hurley 2012). This cost (or this probability) is an exogenous variable. When traceability is mandatory, it thus depends on policy makers' decisions.

The two main messages of this literature are first that traceability is not unambiguous for food safety since precision in tracking foodstuff and incentive payments in contracts are substitute (Resende-Filho and Hurley 2012) or because incentives to choose contracts selecting safe producers not only depend on the cost allocation factor but also on the importance of the failure costs (Starbird and Amanor-Boadu 2007). Second, depending on the values observed for the various other variables influencing incentives to deliver safe products, policy makers should choose adequate values for traceability.

This ambiguity is also found when incentives come from industry reputation preservation rather than the allocation of the cost of food safety failures. Indeed, when traceability enhances food safety along the food chain, food safety reputation for the industry is created. However, this reputation is a fragile asset and is highly sensible to food safety failures and publicized product recalls. A very localized disruption may disturb the entire food chain if the product withdrawal cannot be targeted. In such a context, traceability allowing targeted and rapid product withdrawal may be

seen as protecting the reputation of the food chain and its profits (Pouliot and Sumners 2013). However, this result on profits presupposes that demand reaction to food safety failures is strong. If this assumption is relaxed, targeted withdrawals of product have the effect of increasing price to which products remaining on the market are sold. This situation benefits to some suppliers thus less inclined to collectively engage in traceability. At an individual level, the expected revenue may be a decreasing function of food safety reached at the industry level (Pouliot and Sumners 2013).

The consideration of liability appears in articles in line with the previous ones dealing with incentives. In a food chain composed of farms, marketers, and consumers, liability incentives are enhanced by traceability and result in higher food safety level (Pouliot and Sumners 2008). Traceability from marketers to farms does not alter marketer's liability but allows the former to impose costs of food safety failures on the latter. The incentives for delivering safe products thus created by traceability increase consumers' willingness to pay for the foodstuff. Indeed, they consume safer food and are more likely compensated in case of a food safety problem. This behavior results in higher incentives to produce safe food for the entire food chain. However, as the industry size (i.e., the number of operators) increases, free-riding behaviors appear. Operators realize that the impact of their investment in food safety on the global food safety is decreasing. As a corollary, the "concentration" of the supply chain would positively impact on food safety. For a given size of a food chain, the larger are the different operators (i.e., the less numerous they are at each step), the less likely is free riding in food safety investment (Pouliot and Sumners 2008).

Discussion

In light of the preceding developments, voluntary traceability clearly raises a coordination issue among the different operators of a same food chain that has the potential to reframe the supply chain. Very few studies have tackled this problem

that may constitute an important theme for future research. Banterle and Stranieri (2008) show, for example, that voluntary traceability increases vertical coordination among firms. As both product quality and characteristics of the processes may be tracked (especially within a private standard framework), diverging views among upstream and downstream operators on the kind of traceability that should be implemented through the chain may emerge and have to be resolved. Therefore, the weight of operators, their strategic position within the supply chain, and their capacity to coordinate at a given step of the food chain to face other operators may be seen as central elements for the kind of traceability finally implemented. If food safety objective remains central in this coordination, other economic motives may be at stake and need to be analyzed.

Cross-References

- [Food Safety](#)
- [Geographical Indications](#)
- [Labeling](#)
- [Risk Management, Optimal](#)

References

- Banterle A, Stranieri S (2008) The consequences of voluntary traceability system for supply chain relationships. An application of transaction cost economics. *Food Policy* 33:560–569
- Charlier C, Valceschini E (2008) Coordination for traceability in the food chain. A critical appraisal of European regulation. *Eur J Law Econ* 25:1–15
- Galliano D, Orozco L (2011) The determinants of electronic traceability adoption: a firm-level analysis of French agribusiness. *Agribusiness* 27:379–397
- Galliano D, Orozco L (2013) New technologies and firm organization: the case of electronic traceability systems in French agribusiness. *Ind Innov* 20:22–47
- Golan E, Krissoff B, Kuchler F, Calvin L, Nelson K, Price G (2004) Traceability in the US food supply: economic theory and industry studies. *Agricultural Economic Report 830*, USDA, Economic Research Service
- Hobbs JE (2004) Information asymmetry and the role of traceability systems. *Agribusiness* 20:397–415
- Pouliot S, Sumner DA (2008) Traceability, liability, and incentives for food safety and quality. *Am J Agric Econ* 90:15–27

- Pouliot S, Sumner DA (2013) Traceability, product recalls, industry reputation and food safety. *Eur Rev Agric Econ* 40:121–142
- Resende-Filho MA, Buhr BL (2008) A principal-agent model for evaluating the economic value of a traceability system: a case study with injection-site lesion control in fed cattle. *Am J Agric Econ* 90:1091–1102
- Resende-Filho MA, Hurley TM (2012) Information asymmetry and traceability incentives for food safety. *Int J Prod Econ* 139:596–603
- Schroeder TC, Tonsor GT (2012) International cattle ID and traceability: competitive implications for the US. *Food Policy* 37:31–40
- Souza Monteiro DM, Caswell JA (2009) Traceability adoption at the farm level: an empirical analysis of the Portuguese pear industry. *Food Policy* 34:94–101
- Starbird SA, Amanor-Boadu V (2006) Do inspection and traceability provide incentives for food safety? *J Agric Resour Econ* 31:14–26
- Starbird SA, Amanor-Boadu V (2007) Contract selectivity, food safety, and traceability. *J Agric Food Ind Organ* 5, Article 2

Tradable Emission Rights

► Emissions Trading

Trade Secrets Law

Luigi Alberto Franzoni
University of Bologna, Bologna, Italy

Abstract

Standardization of trade secrets protection was one of the goals of the TRIPs Agreement of 1998. Still, substantial differences across jurisdictions remain. In defining the optimal scope of trade secrets law, lawmakers should consider that strong protection is likely to promote inventiveness but also to retard the diffusion of knowledge and stymie competition.

Definition

Trade secrets law protects firms from unauthorized disclosure of valuable information.

The misappropriation of trade secrets generally constitutes an act of unfair competition that triggers civil liability and, possibly, criminal penalties. Standard examples of misappropriation include espionage, breach of nondisclosure agreements, and unauthorized revelation to third parties.

Historically, trade secrets law has its roots in the Middle Ages, in a time when craft guilds jealously protected their specialized knowledge (the “mysteries” of the arts). Within the guild, as well as in master-apprentice relationships, secrecy was the standard, and its violation could be sanctioned with the capital penalty (see Epstein 1998). In modern times, violation of secrecy has been regarded as an act of unfair competition contrasting with the honest practices that should prevail in the business community. Unfair competition was mentioned in the *Paris Convention for the Protection of Industrial Property* of 1883 (art. 10bis), although with no direct reference to the misappropriation of trade secrets. In spite of the convention, huge differences across countries characterize the field of unfair competition law (see De Vrey 2006; Henning-Bodewig 2013).

Commercial and technological information can leak out of firms in many ways:

It can be stolen by employees or third parties (as in the case of information embodied in documents, files, or technological items); it can be obtained by means of subtle espionage techniques (tapping, dumpster diving, etc.); it can be disclosed to third parties by unfaithful employees; it can be memorized and taken away by former employees who start their own business; it can be indirectly devised by competitors by means of reverse engineering; it can be obtained from the scrutiny of documents submitted to regulatory agencies; it can be obtained by rivals by communication with parties related to the information owner (e.g., buyers and suppliers).

The main function of trade secrets law is to clearly tell methods for the acquisition of information that are admitted from those that are not. Those that are not constitute acts of “unfair

competition” and lead to civil and criminal liability. Given the multitude of ways in which commercial and technological information can spill from firms to the others, in most countries trade secrets law provisions are scattered across several branches of the law, including tort law, contract law, intellectual property (IP) law, labor law, and criminal law. On this account, substantial variations exist across legal systems (see European Commission 2013).

At international level, an important definition of trade secrets is provided by the TRIPs Agreement (art. 39.2), which postulates that lawfully acquired business information qualifies as a trade secret only if it (a) is secret, (b) has commercial value because it is secret, and (c) has been subject to reasonable steps to keep it secret. From this definition, we learn that publicly available information and everyday knowledge are not eligible for legal protection; valueless information and information not subject to reasonable protection do not qualify as trade secrets. All countries belonging to the WTO should make sure that legal protection is granted to trade secrets against acts of misappropriation which include “breach of contract, breach of confidence, and inducement to breach” (TRIPs, footnote 10). The task to define the precise set of activities falling under the “misappropriation” category lies with individual countries, which might be more or less strict on this account. In turn, misappropriation leads to remedies that usually include injunctive relief and damage awards. The latter are typically commensurate with the actual loss to the trade secret owner or the unjust enrichment of the party that has misappropriated the secret. In most countries, courts can also set a reasonable royalty for the use of the secret (for a limited time span).

From a policy perspective, the main issue raised by trade secrets law is the optimal scope of the protection granted to the owners of undisclosed information (which conducts are forbidden and which are not). Some polar conducts can easily be categorized: the theft of documents is undoubtedly unlawful, while reverse engineering tends to be lawful everywhere. With respect to other conducts, courts and lawmakers offer a variety of positions. For example, courts and

lawmakers can be more or less lenient with respect to key employees who leave their company to work for a competitor. In some jurisdictions, this conduct can lead to an unfair competition suit, under the doctrine of inevitable disclosure. In some other jurisdictions, e.g., in California, workers’ mobility finds little impediments in trade secrets law (see Gilson 1999). In deciding the strength of trade secrets protection, lawmakers need to keep in mind that strong protection comes at the cost of reduced labor mobility and lower diffusion of technological knowledge (see Fosfuri and Rønde 2004).

More generally, trade secrets law has been credited with the following beneficial effect (see Lemley 2011). First, by preventing unwarranted diffusion of valuable information, trade secrets law provides a competitive edge to the original producer of the information. With respect to innovative knowledge, for instance, the head start advantage provided by secrecy is regarded as important by most companies (see Cohen et al. 2000). Second, trade secrets law allows firms to reduce the self-protection expenditure: thanks to the legal obstacles against unwanted disclosure, firms can more freely organize their units and share information across their members (see Rønde 2001). In the absence of legal protection, costly measures would have to be taken to reduce the probability of leakage. In fact, reduction of self-protection expenditure is the main function that Landes and Posner (2003) credit to IP law. Facilitated information sharing can also concern contracting parties outside the firm. Effective enforcement of nondisclosure agreements facilitates the transmission and sale of information from the producer to third parties. In this sense, nondisclosure agreements represent a partial solution to Arrow’s information paradox (Arrow 1962), which postulates the unavailability of restitutory remedies for unwarranted information disclosure (once shared, information cannot be returned).

At the same time, trade secrets law also produces social costs. First, by limiting the circulation of information, trade secrets law may retard imitation and stymie technological progress. It has been noted, in fact, that strong technological

spillover characterizes some fast-evolving technological districts like the Silicon Valley, where high levels of labor mobility speed up the diffusion of innovative knowledge (Saxenian 1996). Information sharing facilitates the expansion of the stock of public knowledge, giving birth to new forms of collective invention (Allen 1983; von Hippel and von Krogh 2011).

A further cost brought about by trade secrets law concerns the relationship between secrecy and patent protection (explored, more generally, by Hall et al. 2014). If trade secrets law is strong, inventors have weaker incentives to rely on the patent system. Hence, fewer inventions are disclosed in the patent applications and the stock of public knowledge may advance less rapidly. Here, the issue is whether patents or trade secrecy are better protection tools from a social point of view. Patents have limited duration (normally 20 years) and require disclosure of the invention. Trade secrets can potentially last forever and, by definition, are not disclosed. This implies that either nobody has access to the information – and the original owner retains market power forever – or that third parties have to waste resources to rediscover that original invention (for the purpose to market it directly or to improve upon it). The comparison between the two forms of protection hinges on the nature of the innovation process (how many firms have the capacity to come up with the original invention, the extent of the research spillovers) and the nature of competition across firms upon duplication (under trade secrecy) (see Denicolò and Franzoni 2012).

The impact of trade secrets law on the propensity to patent, however, should not be overestimated. In fact, the subject matter of trade secrets law is infinitely broader than that of patent law. This is because nearly any type of commercial and technological information is eligible for trade secrets protection, while only inventions that meet the originality and non-obviousness requirements qualify for patent protection. A recent investigation by Hall et al. (2013) reveals that only 5% of innovative UK firms rely on the patent system. All of them, one way or another, rely on secrecy.

Cross-References

- [Competitive Neutrality](#)
- [Innovation](#)
- [Intellectual Property: Economic Justification](#)

References

- Allen RC (1983) Collective invention. *J Econ Behav Organ* 4(1):1–24
- Arrow K (1962) Economic welfare and the allocation of resources for invention. In: *The rate and direction of inventive activity: economic and social factors*. Princeton University Press, Princeton
- Cohen WM, Nelson RR, Walsh JP (2000) Protecting their intellectual assets: appropriability conditions and why US manufacturing firms patent (or not) (No. w7552), National Bureau of Economic Research
- De Vrey R (2006) Towards a European unfair competition law: a clash between legal families. Martinus Nijhoff Publishers, Leiden
- Denicolò V, Franzoni LA (2012) Weak intellectual property rights, research spillovers, and the incentive to innovate. *Am Law Econ Rev* 14(1):111–140
- Epstein SR (1998) Craft guilds, apprenticeship, and technological change in preindustrial Europe. *J Econ Hist* 58(03):684–713
- European Commission (2013) Study on trade secrets and confidential business information in the internal market. Downloadable at http://ec.europa.eu/internal_market/iprenforcement/trade_secrets/
- Fosfuri A, Rønde T (2004) High-tech clusters, technology spillovers, and trade secret laws. *Int J Ind Organ* 22(1):45–65
- Gilson RJ (1999) Legal infrastructure of high technology industrial districts: Silicon Valley, route 128, and covenants not to compete. *NY Univ Law Rev* 74(3): 575–629
- Hall B, Helmers C, Rogers M, Sena V (2013) The importance (or not) of patents to UK firms. *Oxf Econ Pap* 65(3):603–629
- Hall B, Helmers C, Rogers M, Sena V (2014) The choice between formal and informal intellectual property: a review. *J Econ Lit* 52(2):375–423
- Henning-Bodewig F (ed) (2013) *International handbook of unfair competition*. Beck, Hart, Oxford and Nomos, Baden Baden
- Landes W, Posner R (2003) *The economic structure of intellectual property law*. Harvard University Press, Harvard
- Lemley MA (2011) The surprising virtues of treating trade secrets as IP rights. In: Dreyfuss R, Strandburg K (eds) *The law and theory of trade secrecy: a handbook of contemporary research*. Edward Elgar Publishing, Cheltenham
- Rønde T (2001) Trade secrets and information sharing. *J Econ Manag Strateg* 10(3):391–417

- Saxenian A (1996) *Regional advantage: culture and competition in Silicon Valley and route 128*. Harvard University Press, Harvard
- Von Hippel E, von Krogh G (2011) Open innovation and the private-collective model for innovation incentives. In: Dreyfuss R, Strandburg K (eds) *The law and theory of trade secrecy: a handbook of contemporary research*. Edward Elgar Publishing, Cheltenham

Tradeable Discharge Permits

► Transferable Discharge Permits

Trademark Dilution

Giovanni Battista Ramello
DiGSPES, University of Eastern Piedmont,
Alessandria, Italy
IEL, Torino, Italy

Definition

A trademark dilution claim is raised whenever a new trademark, albeit it does not confuse consumers, produces detriment to the distinctive character of another trademark. Then the use of new trademark, although used in a noncompeting market, can still be forbidden because it weakens the distinctiveness of the already existing trademark.

Main Characteristics

Trademark dilution is a special kind of infringement that does not involve a direct violation, but it is rather a theoretically legitimate behavior that can nevertheless compromise the distinguishing effect of a given trademark (Economides 1998). It is forbidden by recently amended laws, and it is generally raised in the case of so-called famous or strong trademarks. Antidilution measures essentially enshrine the accomplished transformation of

trademark into a property right over a sign and its semantic sphere (Beebe 2004; Ramello 2006). More precisely the goal of antidilution clauses is no longer to protect the informational value of the trademark and avoid consumer confusion, for which the existing regulations were sufficient, but rather to punish behaviors that may indirectly encroach upon this newly forged semantic sphere and its economic exploitation (Landes and Posner 2003). The dilution claim is generally raised whenever a trademark – even if it does not confuse consumers – produces “detriment to the distinctive character” of another famous mark (*Adidas-Salomon AG v. Fitnessworld Trading Ltd.*, 2003, 1 C.M.L.R. 14) according to European law, or results in “lessening of the capacity of a famous mark to identify goods and services” according to US law (Lanham Act, Section 43c, 15 USC Section 112). The main idea is that the use of a sign somehow linkable to a famous one in a noncompeting market can still be forbidden because it weakens the distinctiveness (and most importantly the differentiation effect) of the famous mark. This can happen by “tarnishment” if applied to inferior-quality goods that might lower consumers’ opinion of the famous mark, or by “blurring” when there is a sort of semantic free-riding, so that the new sign does not confuse the consumer but still indirectly exploits, and thereby impoverishes, the distinctiveness of the famous mark (Lunney 1999).

The general legislative principle that emerges is thus to preserve the semantic capital of a given trademark, in the market and the minds of consumers, by the difficult route of defining property boundaries in the semantic sphere (Ramello and Silva 2006). Such a solution requires directly protecting the signified through a new intellectual property right that extends beyond the tangible market to cover the market for meanings (Ramello 2013). By so doing, the law protects firms’ investments in creating new semiotic products and confirms the altered function of the trademark, which now assigns a legal and economic monopoly over a broad relation between signifier and signified (Ramello and Silva 2006).

The enactment of antidilution measures, after being pushed back several times, was ultimately achieved through protracted lobbying efforts although some scholars expressively speak of “doctrinal puzzlement” (McCarthy 2004) and of “the death of common sense” (Lemley 1999). However, besides these criticisms, the antidilutions measures mark the metamorphosis of trademark, born as an informational tool originally intended as an adjunct to the market for safeguarding consumers, into a property right over a sign and its semantic sphere (McClure 1996). In the meanwhile, it marks the “divorce” of trademarks from the goods they are supposed to represent, which in fact heralds their own “modification” (Lemley 1999).

Cross-References

- [Intellectual Property: Economic Justification](#)
- [Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond](#)

References

- Beebe B (2004) The semiotics of trademark law. *UCLA Law Rev* 51:621–704
- Economides N (1998) Trademarks. In: Newman P (ed) *The new Palgrave dictionary of economics and the law*. Macmillan, London, pp 601–603
- Landes WM, Posner RA (2003) *The economic structure of intellectual property law*. Harvard University Press, Cambridge, MA
- Lemley MA (1999) The modern Lanham act and the death of common sense. *Yale Law J* 108:1687–1715
- Lunney GS Jr (1999) Trademark monopolies. *Emory Law J* 48:367–487
- McCarthy JT (2004) Dilution of a trademark: European and United States law compared. *Trademark Rep* 94:1163–1181
- McClure DM (1996) Trademarks and competition: the recent history. *Law Contemp Probl* 59:13–43
- Ramello GB (2006) What’s in a sign? Trademark law and economic theory. *J Econ Surv* 20:547–565
- Ramello GB (2013) The multi-layered action of trademark: meaning, law and market. In: Miceli TJ, Baker MJ (eds) *Research handbook on economic models of law*. Edward Elgar, Cheltenham
- Ramello GB, Silva F (2006) Appropriating signs and meaning: the elusive economics of trademark. *Ind Corp Chang* 15:937–963

Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond

P. Sean Morris

Faculty of Law, University of Helsinki, Helsinki, Finland

Abstract

The economic analysis of trademark law continues to draw a number of commentaries, yet more and more, the courts are not factoring concrete economic analysis of trademark law and trademark protection in their decisions. In this entry I give an overview and status of trademarks from a law and economic perspective and suggest that trademark laws need to respond to the economic dimension that occurs on the market and consumer economic behaviour.

Introduction

Trademarks are signs that communicate the economic interest of goods manufacturers or service providers to customers with valuable information to influence their purchasing behavior. Signs which constitutes a trademark are capable of graphical representation, in particular words, letters, numbers, shape of goods, or their packaging, in as much, such signs can distinguish the goods and services of different providers. Thus, one of the underlying functions of a trademark is to provide information about source and origin of goods and services. A trademark is granted and regulated under the trademark laws applicable to the territory, national or *supra-federal*, for which it is applied, and examples of supra-federal trademark legislations are the *Trade Mark Directive* (TMD 2008) in Europe or the *Community Trade Mark Regulation* (CTMR 2009) both of which are for trademark regulation and harmonization in the European Union (EU).

A Federal trademark legislation is, for example, the *Lanham Act* in the United States America

(USA), while a national trademark legislation is, for example, the *Trade Marks Act 1994* in the United Kingdom (UK) or the Swedish Trademark Act (2010: 1877). These acts together constitute the legal regulation for trademarks in their respective territories, although the EU's *supra-federal* TMD and national trademark laws coexist for the harmonization of trademark law throughout the EU's internal market. In the USA, the passage of the *Lanham Act* in 1946 saw the US trademark law as "harmonized" so to speak. The situation in the EU is somewhat different given that the road to harmonization of European trademark law has only been a recent phenomenon with the first TMD of 1989 and revisions in 2008 and 2014. A new trademark directive and trademark regulation in the European Union came into force in January and March 2016 respectively.

Trademarks in Law and Economics

In cases such as *Intel v CPM*, the European Union Court of Justice (EUCJ) introduced the notion of consumer *economic behavior* in the context of European trademark jurisprudence. The recognition by the Court of a change in the economic behavior of consumers was arguably an overt acknowledgement of the role of economic analysis of trademarks. Trademarks as a branch of intellectual property rights have always been the gatekeeper of how businesses function in a competitive economy, and, as the face of goods, products, and services, trademarks' economic functions and regulation are essential to both the rights holder and how competitors enter the market.

Because trademarks are signs that represent the economic identity of goods and services and, in this regard, they serve as the bridge that links customers' economic behaviors to the proprietor of goods and services, then trademarks are arguably the most important aspect of intellectual property rights in a market economy because unlike other regimes such as patents and copyrights, a trademark, albeit with renewals, is essentially perpetual and generally outlives the initial rights holder(s). This perpetual nature of

trademarks demonstrates that they are important in how goods, products, and services are placed on the market to guarantee fair trade among the many players that operate in a dynamic market. A dynamic market will always allow flexibility in trademark use through licensing or other methods for trademark owners to encourage fair competition.

But another dynamism of the market is the economic behavior of trademarks, and perhaps it was no surprise that the EUCJ in *Intel v CPM* overtly calls for the economic analysis of trademarks when trademark infringement claims are made. That decision concerned changes in the economic behavior of the average European consumer and how such changes could cause damage to a reputable trademark. This concern by the EUCJ highlights the need for more understanding of the economic dimensions of trademark law and the *real* functions of trademarks in a market economy and its impact on the economic welfare of consumers.

Although the courts sporadically address the law and economic approaches to trademarks, there has long been a growing body of literature that discusses trademarks in a law and economic context. Some of the early scholars that applied law and economics approaches to trademarks include Chamberlin (1933) and Robinson (1933) when their theories of imperfect competition are factored in and, in particular, Chamberlin's reference to "product differentiation," while other scholars (Papandreou 1956; Brown 1948; Mueller 1968) also offered their perspectives on the economics of trademarks. But some of the major works on the economic analysis of trademark law emerged in the late 1980s with those by Landes and Posner (1987) and Economides (1988) and then other contributions by scholars such as Lunney (1999), Ramello (2006), Barnes (2006), and Griffiths (2011).

These approaches to the economic dimension of trademark law incorporate the economic analysis of law. Arguably, the study of law and economics of trademark law began with the great transformation in the study of law and economics as a whole that had roots in the USA at the turn of the twentieth century. Legal luminaries such as

Holmes (1897) helped to sound the trumpet for the law and economic approach when he observed that “the man of statistics and the master of economics” were ideally good tools for the law.

The Chicago School has long been seen as the driving force behind the study of law and economics in the twentieth century with renowned economists in the 1940s and 1950s that gradually built upon *Marshallian* (1890) price theory or other *canons* such as Knight (1933). But the true form of law and economics only took off in the 1960s where economists such as Stigler (1961), Director (1964), and Coase (1960) oversaw the ascent of Chicago approaches to law and economics. Stigler (1961) played an essential role in the development of law and economics of trademarks when he developed a model of optimal consumer search behavior under which advertising reduces consumers’ search costs.

There are critics of the 1960s development of law and economics under the auspices of the Chicago School, and some have even argued that the study of law and economics is rather old given that “the study of the interrelations between legal and economic progress is as old as economics itself” (Medema 2010, p. 160). But even if the origins of law and economics can be disputed, there is without a doubt that the modern phenomena that we understand today as law and economics are essentially a phenomenon that gradually developed as the Chicago School of thought in the USA blossomed in the 1960s and 1970s, even to the point where “lemons” (problems with post-sale used cars) were used to analogize information symmetry (Akerlof 1970) or, in trademark terms, the rights holder have a better idea of the quality his mark represents.

By the 1970s, other Chicagoans such as Posner had also emerged fully in his own right applying economic analyses of the law (Posner 1973), and in 1987 he produced with a colleague a seminal article that reflects similar arguments raised in Stigler’s information advertising paper (1961). The economics of advertising, as per Stigler and the search theory, essentially served as the launch pad for economic analyses of trademarks and arguably because of information that advertising portrays in particular communicating the

trademark to the consumer, even if contemporary scholars believe that “advertising is patently uninformative” (Jordan and Rubin 2008, p. 19). As a result of the broader study of law and economics, the Chicago approach was transposed to a number of other areas such as tax law, antitrust law, criminal law, contract law, and tort law, among others (Cooter and Ulen 2012; Miceli 1997). The law and economics approach gradually moved on to the study of intellectual property rights which covered mainly patents, copyrights, and much later trademarks, and in a 1987 joint paper by Landes and Posner, the law and economics approach was applied to trademark law and to revolutionize the study of trademarks and trademark law in an economic context.

Landes and Posner (1987) in their seminal article which nowadays stands as the cornerstone on the economic analysis of trademark law have argued that a trademark’s essential function from an economic perspective is to reduce consumer search costs. Their arguments are steeped in the Chicago School of economic tradition, in particular, what I think was a reflection of Stigler’s 1961 paper on the economics of information.

An earlier paper dealing with the economics of information, albeit from a legal perspective, also discussed advertising, which broadly encompasses trademarks (Brown 1948); however, this was prior to full development of the Chicago School. Economides (1988) who offered a purely positive economic theory on the economic analysis of trademarks noted in particular that trademarks can serve as a barrier to entry since the trademark “can also serve to increase welfare through the reduction of an excess number of brands” (Economides, p. 536). This approach, and like many others, that offers an economic perspective falls within microeconomic theory that concerns how individuals make rational decisions (behavioral theory), and it is largely responsible for economic analysis of law and by extent trademark law. The approach in the literature that covers the economic analysis of trademark law has essentially still kept at the forefront the problems of trademarks in the competitive economic space of society in their dual role to promote competition and to reduce consumer search costs.

Trademark Monopolies and Economic Effects

In regulatory instruments such as the CTMR, it is explicitly noted in its recitals the desirability to promote “harmonious development of economic activities” using trademarks. In this context, there is an economic rationale for trademark law since quality and origin suggest an economic incentive for the trademark owner while giving assurances to consumers. In this regard, trademark law is the principal arbiter that protects consumers in their economic decision-making when shopping for goods that are protected under trademark law. Yet, it is those very same trademark laws that also give the trademark owner a monopoly lease to use a specific mark to designate his goods or services. The guarantees that the law provides for trademark owners include a right to prevent others from using the mark and also a *right* to enhance the creative and innovative process in a broad sense.

The argument that trademarks are monopolies is not entirely new and has been around for ages, for example, in the English case *Blanchard v Hill* (1742), trademarks were identified as “one of those monopolies.” What this means, even in contemporary times, is that the exclusivity and absoluteness of trademarks make them one of those monopolies that are subjected to being abused in an anticompetitive way and affect how consumer spending decisions are made. As monopolies, the welfare effect on consumers are negative since their spending powers are dictated by the trademark that they are locked onto, which in turn creates monopoly profits for the trademark owner who has the ability to influence the purchasing habit of the consumer. Furthermore, as a branch of intellectual property rights, trademarks are generally seen through the lens of state-granted monopolies, and the literature that emerged over the last six decades or so generally paints trademarks as monopolies (Lunney 1999).

In a number of ways, trademarks create problems on both the market and for the consumer. For consumers the problem of trademark monopolies is that, contrary to reducing search costs, trademarks lock consumers in *perpetuity* – whereas consumers are loyal only to the manufacturers of

the goods the trademark represents. Arguably then, consumer search costs are then limited since they do not engage in the option of the alternative and or at times (naively) assume the mark to which they are loyal represents quality. However, on occasions, consumers may find alternatives that allow them to be better off. For the market, the monopoly in trademarks erects barriers to entry for new goods on the market. This barrier to entry, in particular contemporary times, can arise from an abundance of registered trademarks that are not in use or competitors unwilling to license a mark to new entrant that will compete with them on the market. These observations highlight the gray zone in which trademarks must operate catering to the consumer and the rights holder.

However, the legal reality of trademarks’ economic effects is perhaps those that are articulated by the Courts since the Courts themselves help to shape the economic activity of markets. In *Qualitex Co. v Jacobson*, a US court was perhaps eager to formally recognize the economic dimension of trademarks and the law when it noted that consumers were essentially rational decision-makers because trademarks allow them to make better purchasing decisions and reduce their search costs (*Qualitex Co. v Jacobson*, 163–164). In Europe, where the harmonization of trademark law is an integral part of the EU’s internal market, the EUCJ has often echoed the economic dimension of trademarks, such as in *Intel v CPM* where it observed the possibility of change in economic behavior due to trademark use.

Perhaps the EUCJ’s more broad reading of the economic dimension of trademark protection and use was in *L’Oreal v Bellure* where the Court spoke of trademark’s investment function. This (additional) interpretation of trademark functions suggests the wide economic value of trademarks and the role trademarks play in a fully integrated and free market. The value of a company’s trademark is often the most important piece of asset that a company holds and can leverage with either for investments or loans in the company. The *L’Oreal v Bellure* Court was indeed right when it pointed out the investment function of trademarks

because the value in a trademark is built over time due to the actual investments in a trademark such as advertising or the reputation that the mark earned. As a valuable asset within a company's portfolio trademarks plays an integral investment function (even if *L'Oreal v Bellure*) did not have this reasoning in mind. The implication here is that the economic function of a trademark goes beyond quality and source function but to that of economic value represented by, a form of *inter-* and *intra-*investments which are acquired over time, thereby creating a market for "goodwill."

But despite the inherent monopoly in trademarks, it is the consumer that benefits the most from trademarking activities since the investment function that trademark performs contains a positive spillover when there is increase in trademarking activities. This spillover is the promotion of more competition and the ability of consumers to rationally select goods that are based on their purchasing power. The trademark would normally signal this and influences how the consumer behaves. In dynamic markets such as the EU where national trademarks operate alongside *European* trademarks, consumers respond differently to *external* trademarks that are not well known or where such trademark does not send quality signal. Therefore, consumers can sometimes be suspicious of goods with unknown reputation that originates in a different country. Another factor that also influences consumer economic behavior toward internal and external trademarks in the EU is that of national consciousness and pricing. Even where a trademark signals quality in a reputable good and a new entrant uses a similar trademark to compete, the Courts have cautioned that only a change in consumer's economic behavior could harm the reputable mark. In *Intel v CPM*, the EUCJ introduced the concept of economic behavior in the trademark *lingua*, suggesting the wider role of economic analyses of trademark law and behavioral patterns of consumers:

the use of the later mark is or would be detrimental to the distinctive character of the earlier mark requires evidence of a change in the economic behaviour of the average consumer of the goods or services for which the earlier mark was registered

and consequent on the use of the later mark, or a serious likelihood that such a change will occur in the future. (*Intel v CPM*, para. 77)

From an economic point of view, the Court cautiously warns that economic changes can affect trademark use and protection. The reasoning by the EUCJ that only a change in the economic behavior of the average consumer could threaten Intel's – INTEL mark if those consumers flock to CPM's – INTELMARK trademark broadens the scope for economic analyses of trademark use and protection and how economic evidence is assessed within the confines of the law. Intel did not provide any (economic) evidence that CPM's INTELMARK was causing harm to its economic activities, and therefore the Court found no reason that Intel's mark was being threatened.

Perhaps one of the wider implications of the *Intel v CPM* ruling by the EUCJ and its notion of change in economic behavior in the average consumer is to look to the broader economic theory of public goods and how that is related to trademark law. Given that Intel's mark is a reputable mark, there is no doubt that it creates positive externalities and therefore a magnet for competitors to free ride on. Trademarks are in a sense - public goods - and therefore an opportunity for competitors to free ride on the good will of reputable marks. In other words, information is a public good, even when embedded in a trademark. Barnes (2006) has similarly argued that trademark owners are information creators and such information is widely available for "all people to use" and hence, arguably is a public good. For trademark law, the implication is even grave, given that, as Barnes (2007) observed, both dynamic allocative efficiency and static allocative efficiency may emerge.

Because trademark protection in Europe and globally is expanding, there is an urgent need to factor in empirical evidence based on the economic behavioral patterns of consumers in trademark litigations that can be used to help the Courts arrive at conclusions that are based on economic theories. Such empirical evidence should go beyond the mere surveys that are often used in

trademark litigation given that excessive trademarks are “goods in their own right” (*Plasticolor v Ford*, 1989, 1332). It is also useful, when discussing economic approaches of trademark law to consider other areas competition law. This is because empirical evidence suggests that trademarks are anticompetitive, for example, by serving as tying products (Smirti 1976) or due to their inherent exclusivity or market power (Morris 2012, 2013), and whether such exclusivity is being abused by trademark owners.

The market economies in which trademarks exist are expanding, and the movement of global commerce has put into sharp focus trademark use and protection and the scope of trademark law. In trademark legislations, whether international instruments such as the Trade-Related Aspects of Intellectual Property Rights (TRIPS), regional harmonization efforts at the EU, and national trademark laws, there is the need to factor in “economic effects” of those laws and how they relate to other areas of regulation in the free market.

Trademarks are no longer seen as *ought* to be protected but rather how to assess the economic impact of trademarks that are currently protected and are in use and what are the economic causes of trademark nonuse. The economic impact of trademarks that are in use drives consumer economic behavior and how the markets respond to any perceivable changes in economic behavior, and this warrants fresh approaches to the economic dimension of trademark law.

Such economic dimension should move into a new direction beyond the predominant function (s) of trademark theory such as the search cost theory or the origin source theory. This new direction may take on, for example, the reduction of barriers to entry by allowing greater flexibility in the use of trademarks by competitors, or how the economic impact of product differentiation in trademarks affects the normative process in which trademarks must operate.

Ultimately, it is the trademark laws that can respond to the normative process, and perhaps, similar to the EUCJ’s recognition of change of economic behavior by the average consumer (also echoed in *Environmental Manufacturing v OHIM*, 2013), trademark laws, both in Europe and

beyond, can begin to reform by responding to changes in the market economy which ultimately leads to greater efficiency in the competitive process for the creation of wealth. Nevertheless, the change of economic behavior test that the European Courts advocate holds interesting interpretation for the economic analyses of trademark law both in Europe and beyond, since this new direction in trademark law interpretation has raised the threshold for trademark infringement analysis.

Cross-References

- [Behavioral Law and Economics](#)
- [Economic Analysis of Law](#)
- [Efficiency](#)
- [Innovation](#)
- [Law and Economics](#)
- [Trademark Dilution](#)

References

- Akerlof G (1970) The market for lemons: quality uncertainty and the market for mechanisms. *Q J Econ* 84:488–500
- Barnes DW (2006) A new economics of trademarks. *Northwest J Technol Intellect Prop Law* 5:22–67
- Barnes DW (2007) Trademark externalities. *Yale J Law Technol* 10:1–44
- Brown R Jr (1948) Advertising and the public interest: legal protection of trade symbols. *Yale Law J* 57:1165–1206
- Chamberlin E (1933) The theory of monopolistic competition: a re-orientation of the theory of value. Harvard University Press, Cambridge
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cooter C, Ulen T (2012) Law and economics, 6th edn. Pearson, Boston
- Director A (1964) The parity of the economic market place. *J Law Econ* 7:1–10
- Economides N (1988) The economics of trademarks. *Trademark Rep* 78:523–539
- Griffiths A (2011) An economic perspective on trade mark law. Edward Elgar, Cheltenham
- Holmes O (1987) The path of law. *Harv Law Rev* 10:457–478
- Jordan E, Rubin P (2008) An economic analysis of the law of false advertising. In: Mialon HM, Rubin PH (eds) Economics, law and individual rights. Routledge, London, pp 18–43

- Knight F (1933) The economics of organization. University of Chicago, Chicago
- Landes W, Posner R (1987) Trademark law: an economic perspective. *J Law Econ* 30:265–309
- Lunney G Jr (1999) Trademark monopolies. *Emory Law J* 48:367–487
- Marshall A (1890) Principles of economics. Macmillan, London
- Medema S (2010) Chicago law and economics. In: Ross E (ed) The Elgar companion to the Chicago school of economics. Edward Elgar, Cheltenham, pp 160–174
- Miceli T (1997) Economics of the law: torts, contracts, property, litigation. Oxford University Press, New York
- Morris PS (2012) The economics of distinctiveness: the road to monopolization in trademark law. *Loyola Int Comp Law Rev* 33:321–386
- Morris PS (2013) Trademarks as sources of market power. Working Paper
- Mueller C (1968) Sources of monopoly power: a phenomenon called ‘Product Differentiation’. *Am Univ Law Rev* 18:1–42
- Papandreou A (1956) The economic effect of trademarks. *Calif Law Rev* 44:503–510
- Posner R (1973) Economic analysis of law, 1st edn. University of Chicago Press, Chicago
- Ramello G (2006) What is a sign? Trademark law and economic theory. *J Econ Surv* 20:547–565
- Robinson J (1933) The economics of imperfect information. Macmillan, London
- Smirti S (1976) Trademarks as tying products: the presumption of economic power. *St Johns Law Rev* 50:689–724
- Stigler G (1961) The economics of information. *J Pol Econ* 69:213–225

Cases and Legislations

- Blanchard v Hill, 26 Eng Rep 692 [1742]
- Case C-252/07, Intel Corporation Inc v CPM United Kingdom Ltd [2008], ECR I – 8823
- Case C-383/12 P, Environmental Manufacturing LLP v OHIM [2013], ECR I – 0000
- Case C-487/07, L’Oreal SA v Bellure NV [2009] ECR I – 05185
- Council Regulation (EC) No 207/2009 of 26 February 2009 on the Community Trade Mark, OJ L 78/1, 24.3.2009 (CTMR)
- Directive 2008/95/EC of the European Parliament and of the Council of 22 October 2008 to approximate the laws of the Member States Relating to Trade Marks, OJ L 299/25, 8.11.2008 (TMD)
- Joined Cases C-236 to C-238/08, Google France v Louis Vuitton Malletier [2010], ECR I – 02417
- Lanham (Trademark) Act, 15 U.S.C. s. 1051, 60 Stat 427 [1946]
- Plasticolor Molded Products v Ford Motor Co., 713 F. Supp 1329 [1989]
- Qualitex Co. v Jacobson Products Co., 514 US S. Ct. 159 [1995]
- Trade Marks Act (UK) [1994]
- Trademark Act (Sweden) [2010]

Trading of Allowances

► Transferable Discharge Permits

Traffic Light Violation

► Traffic Lights Violations

Traffic Lights Violations

Laurent Carnis

University Paris Est and IFSTTAR, Noisy-le-Grand, France

Abstract

Traffic light running is a violation of Highway Code that endangers the offender as well as other road users. The consequences of crash risk induced by this reckless behavior can be interpreted as a social cost and call for public intervention. Several tools are available for policy makers to enforce traffic light obedience. However, the cost of public intervention must be in line with the social cost of violations.

Although there is a sizable literature dealing with traffic light running, researchers generally focus on the predictors of such behavior and the impact of the countermeasures. This entry presents a literature overview from an economic perspective and proposes an economic analysis of traffic light violations and their regulation.

Synonyms

Traffic light violation

The author thanks E. Kemel for comments and suggestions on a previous version. The usual caveat applies.

Introduction

Red light running (RLR) is a violation of traffic rules that endangers the offender as well as other road users. A red light isolation is established when a driver fails to stop at a red signal indication. The crash risk induced by this reckless behavior calls for public intervention. Indeed, RLR can be considered as an external effect imposed upon road crash victims. A social cost is associated with this illegal behavior. The existence of this external effect finds its origin in the fact that the road is used as a common. Several tools are available for policy makers to internalize those costly consequences and enforce traffic light obedience.

Although there is a sizable scientific literature dealing with RLR, researchers generally focus on the predictors of such behavior and the impact of the countermeasures from an engineering and psychological perspective. The economic approach is quite inexistent, so that it is impossible to determine the appropriate intervention and be insured the public policy is efficient. It neglects also the economic dimension of driver's choice. This entry proposes an economic approach which provides a new perspective for this issue and gives an account of the specialized literature.

An economic approach is required because red light violation as an external effect calls for a public intervention. However, the cost of intervention has to be proportional to the social cost of violations. In other words, the cost element constitutes the crucial dimension not only for framing countermeasures but also for understanding the offender's choice.

The first section provides an economic explanation of the need for traffic light at intersection through a game theory approach emphasizing upon the needs of cooperation and fairness. The second section deals with an analysis of RLR from the driver's choice perspective. The consequences of RLR are identified in section "Consequences of Traffic Light Violations". Section "Regulation of Red Light Violations" reviews the available possibilities for enforcing this safety rule and regulating efficiently the RLRs through the economist's lens.

Signalized Intersections

Cooperation and Safety at Intersections

Without traffic regulation, the situation of two drivers reaching the intersection from opposite directions can be represented as a noncooperative game (Table 1). If both players stop, they suffer a time loss ($-t$); if they both go, they suffer a large loss related to a road crash ($-a$), with $a > t > 0$. If they make different choices, the stopping driver suffers the time loss, while the other driver breaks even. In this game, the Pareto optimum is reached when drivers take different strategies.

In that case, there is no dominant strategy. If the adverse driver is expected to stop (resp. go), the best decision is to go (resp. stop). The absence of unique equilibrium can lead to a suboptimal situation (the crash or the mutual stop). By forcing one driver to stop, traffic light can be considered as an exogenous intervention that imposes one Pareto optimal equilibrium. Then traffic light can find a justification from the economic perspective by making cooperation possible between drivers. Here a better cooperation is related with traffic safety.

Traffic Management: Fairness and Sustainable Cooperation

Different types of traffic regulation are possible for solving the coordination problem at intersection such as putting a stop sign or giving priority to the right. However, the two aforementioned solutions systematically give priority to the same users. For the sake of fairness and efficiency of traffic regulation, traffic light alternates the two equilibria. Drivers from the non-priority roads are not systematically penalized, and waiting time is shared between drivers coming from opposite direction. Thus, traffic light ensures fairness and reciprocity among users. It is an important feature.

Traffic Lights Violations, Table 1 Game matrix of drivers at intersection

Driver 1	Driver 2	
	Go	Stop
Go	$-a, -a$	$0, -t$
Stop	$-t, 0$	$-t, -t$

Indeed, fairness and reciprocity can motivate cooperative behavior (Fehr and Schmidt 2006) and can therefore contribute to driver obedience to this traffic rule. It also makes it a sustainable one.

Traffic management is another main objective of traffic light regulation. Traffic lights are sometimes used for ramp metering, in order to make new entrants wait during congested periods. Others are used for regulating speed in urban areas. Such regulation aims at producing a smoother and calmer traffic among the road users and avoiding congestion. A driver pays a limited period of waiting time in order to enjoy larger gains with a reduced total driving time. Although such traffic management tools aim at reducing road crashes and time wastes, and ultimately the related social costs, not all drivers always abide to these rules. In a sense, RL violation can be considered as a free riding activity. The free rider would like to benefit from traffic safety and management without participating to its funding by obeying the rule. It is also a departure from the Pareto optimal equilibrium defined previously.

Traffic Light Violations

Despite of the related crash risk, red light running (RLR) is not an uncommon event. Retting et al. (1998) observed an average of three violations per hour on two urban intersections in Arlington. Carnis and Kemel (2012) reported very similar results from a field investigation for 24 traffic lights in Nantes (France) and showed that RLR characteristics vary with the type of sites, users, and contexts.

Individual Choice

A RLR is mainly an individual choice. Indeed, the classical school of economics of crime assumes that traffic offenders choose whether to violate the law or not by following utility maximization rules (Becker 1968). For red light violation, the elements of this type of decision can be presented in a decision matrix (Table 2). Stopping at the traffic light results in a sure time loss ($-Ct$),

Traffic Lights Violations, Table 2 Matrix of go/stop individual decision at red light

	No enforcement	Enforcement	
	No crossing vehicle	No crossing vehicle	Crossing vehicle
Stop	$-Ct$	$-Ct$	$-Ct$
Go	0	Cf	$-Ca$

whereas the consequence of a red light violation is uncertain and event contingent. It depends on the presence of another vehicle crossing the intersection and enforcement (a police officer or camera). If one of these events occurs while the driver decided to run the light, a cost is suffered: ($-Ca$) for the crash cost and ($-Cf$) for the cost of a fine, with $Ca > Cf > Ct > 0$. Ca can also include a penalty for causing the crash by red light violation.

There is no dominant strategy and the decision depends on users' beliefs and preferences. Decision under uncertainty is classically modeled by expected utility. This model assumes that decision makers assign subjective probabilities to events and subjective utilities to consequences and choose the alternative that maximizes the mathematical expectation of their utility. Normalizing the utility with $U(0) = 0$, a driver is expected to run the light if $U(-Ct) < pf \times U(-Cf) + pa \times U(-Ca)$, where pf (resp. pa) is the subjective probability of being fined (resp. responsible of a road crash). RLR is thus expected to vary across individuals and contexts depending on attitudes and perceived risks. This framework predicts also that RLR decreases when Pf , Pa , Cf , or Ca increases or when Ct decreases.

Empirical studies bring evidence for most of these predictions. The literature shows that increasing detection probability (by the mean of red light cameras, for instance) reduces RLR (Council et al. 2005). Moreover Carnis and Kemel (2012) show that violation rates are higher during night and low-traffic time periods when crash risk is lower.

The impact of red light duration on RLR was highlighted by Retting et al. (2008). When waiting time is too long, drivers fail to respect it. Guidelines recommend not having red light durations that exceed 2 min (CERTU 2010).

Carnis et al. (2012) report field data showing that most RLRs occur during the very first seconds of the red phase, when the violation is the most profitable in terms of avoided waiting time.

Heterogeneity is also expected between users, because of the diversity of individual preferences. Retting et al. (1999) compare authors of RLR accidents to those of other accident types. Males are overrepresented among this population. Red light violators are also younger and more likely to be intoxicated. Porter and England (2000) observed a relationship between RLR and safety-belt use. Propensity to abide to red light also depends on the vehicle type (Carnis and Kemel 2012).

Coping with Dilemma and Interactions Between Drivers

The decision to respect the rule must be taken in a very short amount of time. Indeed, drivers must make the go/stop decision within a few seconds. Because of the urgency dimensions of the decision and drivers' cognitive limits, illegal actions can sometimes be taken by mistake (Depken and Sonora 2009). Decision to go or stop at light also requires the driver to analyze the situation because the presence of closely following vehicles must be checked. If the decision to stop is likely to trigger a rear-end collision, decision to run the light must be taken. Consequently, illegal decision can be followed in particular situation for avoiding harm and costly consequences. Those aspects are not generally accounted for by the economics of crime framework that assumes that decision makers have time to choose, face clear-cut situations, and feature perfect cognitive capabilities.

The dilemma that faces the driver approaching light received an important attention in the literature (Elmitiny et al. 2010; Papaioannou 2007). The situation in which the driver is unable to stop safely or crossing the intersection at the green light is called the dilemma zone. The dilemma zone is related to the duration of amber light and the approaching drivers' speed. Shortening this time increases RLR because drivers are not averted that the light will switch. Increasing this time increases the dilemma zone and may increase the number of drivers running amber

light. Drivers exceeding speed limits are more likely to run amber and red light. Therefore, the dilemma zone does not only puzzle drivers but network managers as well.

Decision to run or not the red light is not only individual, but it is also impacted by other drivers' behavior. The choice to commit a RLR takes into consideration the presence of other (preceding or following) drivers. For instance, drivers are more likely to run a light when a preceding driver did so (Elmitiny et al. 2010, p. 110). Observing that the preceding user runs the light may provide valuable information for decision that enforcement is low or nonexistent. Even though following behavior can be rational, it also creates risk of rear-end crash if the preceding driver decides to stop at the traffic light.

Consequences of Traffic Light Violations

Safety Consequences

Red light violation is a major concern for the policy makers because of the number of road crashes and victims involved (McGee and Eccles 2003). From the economist standpoint, road crashes are interpreted as an external effect related to the common use of the road network.

Moreover, the urban intersection implies mainly the involvement of vulnerable users (pedestrians, bicyclers, and motorcyclists). It means also the collision is characterized by a true asymmetrical dimension in terms of vulnerability between the involved users in a traffic collision (for instance, vehicle vs. pedestrian).

Large-scale studies evaluating the prevalence of RLR are not common. Retting et al. (1999) report that accidents occurring at intersections represent 27% of all injury crashes in the United States. Accidents due to RLR are however less frequent. Over the 1992–1996 period, RLR crashes represented 3% of all fatal crashes and 7% of injury crashes on urban roads. Compared to the prevalence of violations, these figures suggest that the collision probability in case of RLR is much lower than one could expect. According to the economic approach, violators may also decide to run red light when the traffic conditions and the

visibility minimize crash risk. Carnis et al. (2012) observed that 90% of violations occur in the first two seconds of the red phase, when all lights of the intersection are red.

Paradoxically, a sizable part of intersection crashes derive from red or amber light stopping. Rear-end accidents are indeed not infrequent at signalized intersection. Their number has been found to increase after traffic light camera deployment (Erke 2009). Another frequent type of accident occurring at signalized intersection relates to left turns. Wang and Abdel-Aty (2006) estimate that these accidents rank third after rear-end and angle crashes for 1531 intersections in the state of Florida.

Traffic Regulation Consequences

Traffic regulation is another major objective for installing traffic lights. Therefore, consequences in terms of generated congestion and the related time losses have to be assessed. Time losses can be generated by two types of traffic light violations. First, when the traffic light regulates road access, failure to respect the red light disturbs traffic flow and increases congestion. In this case, the contribution of the marginal violator to congestion is small, but the overall effect can be important when violations are numerous. Second, when RLR occurs during a dense traffic condition, the RL runner can be stuck in the middle of the intersection and can totally freeze traffic on all junctions.

A better respect of red light can help in limiting congestion and save environmental costs related to air pollution. We are not aware of any study evaluating the impact of RLR on time losses, nor environmental costs, due to increased congestion, even if they have to be taken into consideration from an economic perspective.

Regulation of Red Light Violations

Red light violation is a source of external effects. It generates a social cost (mainly associated with the crash costs (material damages and injuries)), which requires internalization. Internalization of this external effect calls for an intervention aiming at the reduction of costs borne by the victims. To mitigate the consequences related to those illegal

behaviors, the policy maker defines and implements a public policy. This social regulation intervention can be achieved by two different categories of policies: enforcement and other interventions.

The Enforcement Policy

Enforcing the Highway Code

Becker's seminal works on the economics of crime show that illegal behavior can be mitigated by implementing an efficient policy of control and punishment (Becker 1968). Both the enforcer and the enforcement authority are concerned by the economic approach to crime. Efficiency of this enforcement policy requires taking into consideration the cost of intervention (respectively the relative costs of detection and punishment) and the social loss related to the harmful consequences of red light violations which could be reduced by deterrence. At the society level, it then becomes possible to determine an optimal deterrent policy associated with an optimal punishment (in terms of intensity of detection and severity of sanction) and an optimal number of violations. Consequently, it is rational from the economic perspective not to enforce all RLRs.

Different Techniques of Production

Different ways exist to enforce traffic light regulation. The traditional approach rests upon the manual detection of offenders by police officers, who monitor and intercept the offenders. This procedure is very costly in terms of time, because it requires a permanent supervision and numerous police officers to be able to catch the offender. In economic terms it is a labor-intensive technique of production. In practice, red light regulation was not especially enforced, because of its high unitary enforcement costs.

Since the mid-1980s, red light cameras (RLCs) have been replacing progressively the traditional enforcement method. This technique of detection can be considered as capital intensive and makes possible a systematic supervision of all the drivers, while minimizing the costs of labor intervention. Automation of traffic safety enforcement is a major trend of those last years, which has to be considered for understanding the spreading of such public programs.

When compared with the traditional approach, the RLC program appears as an efficient way for enforcing the regulation. It presents twofold economic advantages. It reduces substantially the cost of detection and punishment at a given level of traffic. The picture of the offender is automatically processed. The offender is identified through his license plate and receives his traffic infringement notice at home. It permits also to increase substantially the level of detection and punishment. Thousands or millions of tickets can be processed according to the limits of the computer system. In France, the number of RLR tickets was multiplied by 8 after the introduction of RLC. Introduction of RLC programs can be conceived as an innovation lowering the average cost of deterrence and making possible a stricter enforcement of red light regulation by generating scale economies. This cost killing effect explains probably why so many jurisdictions implemented such programs for securing the signalized road intersection (Carnis 2010).

Do RLCs Reduce Crashes?

While several contributions conclude to a positive contribution of RLC by reducing road injuries (Council et al. 2005; McGee and Eccles 2003), others show more debatable effects and question their impact. RLC would yield positive side effects with potential spillover impacts of RLC for other intersections and negative ones by increasing rear-end crashes and all category crashes (Hallmark et al. 2010; Vanlaar et al. 2014). However the gains associated with the reduction of right-angle injury crashes would largely compensate the costs related to the increase of rear-end crashes. More problematic are the recent conclusions of several contributions showing the insignificant impact of RLCs for reducing road crashes (Erke 2009; Høye 2013), contributions which were nevertheless criticized by other scholars putting in question their meta-analysis approach (Lund et al. 2009).

Cost-Benefit Evaluation is Needed

An economic approach to red light violation and regulation becomes particularly necessary when such a public intervention yields opposite and

potential adverse side effects. It constitutes a prerequisite for concluding about the economic efficiency of such programs for reducing road injuries at signalized intersection. Proceeding to the economic assessment of RLC programs requires a comparison between advantages and costs. However, only few studies investigated the economic side of red light violations. More problematic is the finding of a careful literature review showing the quasi-generalized absence of economic assessment of RLC programs and rigorous evaluation of safety impacts, so that it is impossible to conclude that such programs are efficient and to determine the scope of the internalization policy (Langland-Orban et al. 2014). In fact, the present evaluative practices of RLC programs reflect both the complexity of evaluation process (non-replication of experiences in controlled laboratory conditions) and the costs of collecting and analyzing the data. It seems also to reflect that policy makers sometimes look for intervention whatever may be their impact or cost, when facing the risk of human injuries.

Public-Private Partnership for RLR Enforcement

The economic approach is particularly relevant when programs are not directly managed by governments. There are several procurement alternatives. Some of them could associate private operator, while some governments outsource the operation of the program (FHA 2005). The total or partial outsourcing of such social regulation activity raised some new issues concerning the possibility for contractor to manipulate the control activity and illegal use of the collected data (CSA 2002). Such situation is typically a principal-agent situation with asymmetrical information. Indeed, one agent is usually more informed than the other and can modulate its efforts. This contractual dimension emphasizes the necessity of a well-designed contract to be insured that private and public interests are aligned (Travis and Baxandall 2011). Indeed, while governments are more interested in maximizing their return in terms of safety impact (public safety hypothesis), the private firms are more concerned by the maximization of profit. Those considerations are quite important, because

it could influence the location of radars and their impact for public safety.

Another interesting issue is related to the different payment options for the contractor. Fixed-price payments, fixed monthly payments, per citations payments, payments depending on time worked and materials used, and mixed payments are alternative possibilities. However, it is not clearly determined which type of payment is the most advantageous for the government and the most efficient in terms of welfare for the society. Nevertheless there are strong incentive implications for the different agents, especially here for the policy maker and the contractor. Unfortunately this dimension was not investigated.

RLC programs have to be considered also from the institutional perspective (Carnis 2010). Outsourced programs, contractual dimensions, and financial considerations are important characteristics. A more recent trend fires on the RLC programs. More and more governments turned off their RLC because of uncertain impacts in terms of traffic safety already mentioned. Between 2011 and 2013, it is estimated that 200 RLC programs were turned off in the United States (Slone 2014). The policy maker is reluctant to let a program continue while it could increase rear-end injury accidents and potentially jeopardize human lives. Moreover the court system becomes less supportive of such control system: the judge dismisses charges more often because of erroneous readings and identification of some license plates by the program, which questions its reliability. Another consideration concerns the reduction of revenues associated with the RLC while the costs are increasing. Moreover some citizens assimilate RCL fines as a new tax imposed upon the road user. Garrett and Wagner (2009) concluded that sustained municipalities obey revenue motives concerning traffic enforcement and tickets. RLC would not be exclusively concerned with public safety.

Other Policies

Providing Drivers With Better Information

The drivers' decision of stopping or running the red light depends on beliefs and preferences.

However most of the time the driver is not perfectly informed about the risks involved. Consequently, education and awareness campaign can play a useful role in providing accurate information related with the risk the driver faces when taking an adverse decision (FHA 2005). While education and awareness campaigns are conceived here as a provision for helping him in taking a correct decision, it presents a cost. Such campaigns have to be calibrated so that the costs and returns are in a same magnitude.

Nudging Drivers

Behavior change can also be achieved through nudging. This policy consists of framing the decision context. More precisely, the policy maker can design an environment that induces a promoted behavior. Thaler and Sunstein (2008) provide an illustrative example of such policy in the road safety field. The use of stripes can reinforce the visibility of a potential danger of a portion of a curve for instance (Thaler and Sunstein 2008, pp. 37–39). A review of possible applications of nudges to traffic safety policies is proposed by Avineri (2014). Regarding RLR, installing countdown systems can impact drivers' behavior and appears as a typical nudging intervention.

No Traffic Light, No Traffic Light Violation

Red light violation depends also on some Highway Code adaptations and road infrastructure context. For instance, in the United States, users are generally allowed to a right turn during red light as long as they leave priority to other vehicles. During low-traffic hours (e.g., night hours), traffic light can be turned out and the right of way applies, avoiding the unnecessary waiting time at the green light phase. In France, an experiment tests the impact of giving to the bicycle user the possibility to cross the intersection at a red light provided that the priority is given to the opposite coming vehicle.

Reframing the context of the driver decision can also require bringing some modifications to the road infrastructure. Engineer investigations showed the influence of the average daily traffic

volume, the number of traffic lanes, the left-turn lanes, and the speed limit regulation (Langland-Orban et al. 2014). A radical solution for regulating red light violation would consist of removing signalized intersection. In that case, RLR would disappear because the road infrastructure design makes them impossible. In some ways, it could constitute a kind of situational crime prevention approach. Concretely, it would consist of modifying the access to the road section through ramps or implementing roundabouts.

However such interventions can be very costly, especially in an urban context. Again a reasonable economic approach would consist of comparing costs and gains of different alternatives. This approach permits also to avoid the funnel approach by enlarging the problematic to other issues such as mobility and pollution considerations, emphasizing that red light violation prevention cannot be reduced only to public safety consideration.

Conclusion

The economic approach provides a consistent framework for understanding RLR. It is able to account for both users and policy makers decisions and highlight possible alternatives for intervention. It can also explain why it can be rational for a driver to commit a RLR under certain circumstances, but also why red light regulation is not enforced in some cases.

Economic variables are not only at play for explaining the way drivers choose in particular situations, but the economic consequences related to road crash and traffic congestion have also to be considered for understanding the role of traffic light. Economic valuation appears as a true alternative to the engineering perspective for understanding this issue and promotes different analysis and solutions.

Regulation of RLR is achievable and requires a calibrated enforcement policy. RLC program is a possible solution for enforcing traffic safety rules, but an economic approach is needed for designing

correctly the public intervention, which could be assimilated to a particular productive process. Communication campaign, awareness program, and infrastructure modification are other available solutions. Nudging policy appears also as an interesting perspective that can be built upon a behavioral law and economic approach, providing new insights for designing traffic safety rules and enforcement policies.

References

- Avineri E (2014) Nudging safer road behaviours, technical report. Afeka Center for Infrastructure, Transportation and Logistics
- Becker G (1968) Crime and punishment: an economic approach. *J Polit Econ* 78:168–217
- Carnis L (2010) A neo-institutional economic approach to automated speed enforcement systems. *Eur Trans Res Rev* 2(1):1–12
- Carnis L, Kemel E (2012) Assessing the role of context in traffic light violations. *Econ Bull* 32(4): 3386–3393
- Carnis L, Dik R, Kemel E (2012) Should I stay or should I go? Uncovering the factors of red light runnings in a field study. In: *Proceedings of the 40th European transport conference*, Glasgow, UK
- CERTU (2010) Guide de conception des carrefours à feux, technical report. Centre d'Etudes sur les Réseaux, les Transports, l'Urbanisme et les constructions publiques
- Council FM, Persaud B, Eccles KA, Lyon C, Griffith MS (2005) Safety evaluation of red-light cameras, US Department of Transportation, Federal Highway Administration, FHWA-HRT-05-048, pp 95
- CSA (2002) Red light camera programs: Although they have contributed to a reduction in accidents, operational weaknesses exist at the local level, technical report. Bureau of State Audit
- Depken C, Sonora R (2009) Inadvertent red light violations: an economic analysis, Mimeo, p. 32. belkcollegeofbusiness.uncc.edu/cdepken/P/redlights.pdf
- Elmitiny N, Yan X, Radwan E, Russo C, Nashar D (2010) Classification analysis of driver's stop/go decision and red-light running violation. *Accid Anal Prev* 42(1): 101–111
- Erke A (2009) Red light for red-light cameras? A meta-analysis of the effects of red-light cameras on crashes. *Accid Anal Prev* 41(5):897–905
- Fehr E, Schmidt K (2006) The economics of fairness, reciprocity and altruism—experimental evidence and new theories'. In: *Handbook of the economics of giving, altruism and reciprocity*, Vol. 1, Elsevier, Amsterdam, pp 615–691

- FHA (2005) Red light camera systems, operational guidelines, technical report. Federal Highway Administration
- Garrett TA, Wagner GA (2009) Red ink in the rearview mirror: local fiscal conditions and the issuance of traffic tickets. *J Law Econ* 52(1):71–90
- Hallmark S, Orellana M, McDonald T, Fitzsimmons E, Matulac D (2010) Red light running in Iowa. *Trans Res Rec: J Trans Res Board* 2182(1):48–54
- Høye A (2013) Still red light for red light cameras? An update. *Accid Anal Prev* 55:77–89
- Langland-Orban B, Pracht EE, Large JT, Zhang N, Tepas JT (2014) Explaining differences in crash and injury crash outcomes in red light camera studies. *Eval Health Prof* 1–19 (Epub ahead of print)
- Lund AK, Kyrychenko SY, Retting RA (2009) Caution: a comment on Alena Erke's red light for red-light cameras? A meta-analysis of the effects of red-light cameras on crashes. *Accid Anal Prev* 41(4): 895–896
- McGee H, Eccles K (2003) Impact of red light camera enforcement on crash experience, a synthesis of highway practice, technical report. NCHRP synthesis
- Papaioannou P (2007) Driver behaviour, dilemma zone and safety effects at urban signalized intersections in Greece. *Accid Anal Prev* 39(1):147–158
- Porter BE, England KJ (2000) Predicting red-light running behavior: a traffic safety study in three urban settings. *J Safety Res* 31(1):1–8
- Retting RA, Williams AF, Greene MA (1998) Red-light running and sensible countermeasures: summary of research findings. *Trans Res Record: J Trans Res Board* 1640(1):23–26
- Retting RA, Ulmer RG, Williams AF (1999) Prevalence and characteristics of red light running crashes in the united states. *Accid Anal Prev* 31(6): 687–694
- Retting RA, Ferguson SA, Farmer CM (2008) Reducing red light running through longer yellow signal timing and red light camera enforcement: results of a field investigation. *Accid Anal Prev* 40(1):327–333
- Slone S (2014) Speed and red light cameras law, technical report. Capitol Research, The Council of State Governments
- Thaler RH, Sunstein CR (2008) *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven
- Travis M, Baxandall P (2011) Caution: red light camera ahead the risks of privatizing traffic law enforcement and how to protect the public, technical report. US PIRG Education Fund
- Vanlaar W, Robertson R, Marcoux K (2014) An evaluation of Winnipeg's photo enforcement safety program: results of time series analyses and an intersection camera experiment'. *Accid Anal & Prevention* 62: 238–247
- Wang X, Abdel-Aty M (2006) Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accid Anal Prev* 38(6):1137–1150

Transaction Costs

Wim Marneffe, Samantha Bielen and
Lode Vereeck

Faculty of Applied Economics, Hasselt
University, Diepenbeek, Belgium

Abstract

Transaction costs is a generic term referring to the costs of transacting through the market (e.g., search, information, contract, monitoring costs). They are applied with different meanings to organizational structures (e.g., vertical integration), market failures (e.g., externalities), institutional choices (e.g., promotion of clubs), and public choice (e.g., administrative burden).

Introduction

If markets operate as smoothly and efficiently as free market proponents and standard economic textbooks suggest, why, for instance, do firms hire workers on a permanent basis or are professions regulated? The answer is, to a large extent, transaction costs. In principle, every single activity (e.g., typing a letter, making a phone call) can be outsourced to the market. However, searching, finding, screening, contracting, and monitoring a secretary is extremely costly. These “transaction” costs of using the (labor) market can outweigh the proclaimed benefits of allocating productive activities (i.e., a letter, a phone call) through the market. Therefore, legal instruments are created (e.g., labor contract of unlimited duration) in order to avoid transaction costs.

Similarly, the regulation of professions (e.g., licensing of dentists) not only guarantees a standard quality, but also creates barriers of market entry. However, the oligopolistic costs that stem from these barriers are typically offset by the search, information, screening, contracting, and monitoring costs of using the free market of professional services, i.e., the transaction costs.

Transactions costs should always be considered in their historical context. The use, for instance, of labor contracts of unlimited duration is a way to reduce transaction costs outweighing monopolistic costs (i.e., barring other secretaries to type a letter or make a call). However, the emergence of temporary workers' agencies that offer flexible services is a market response to high transaction costs. What is an economic reason to avoid the use of the markets today may not be valid tomorrow.

The concept of transaction costs is probably among the most discussed topics in economics and has led to the emergence of an entire body of literature with numerous theoretical and empirical articles and books. Economists' understanding of transaction costs has continuously evolved since it was first introduced by Ronald Coase trying to explain the emergence of firms (Coase 1937). Currently, transaction costs are an essential part of economic analysis and frequently invoked to explain a plethora of institutional choices and behavior (Rao 2003). Unfortunately, the transaction cost literature is plagued by the use of different definitions. Therefore, this essay presents a consistent and coherent taxonomy of transaction costs.

Transaction Costs: The Prelude

Although Ronald Coase did not explicitly use the words "transaction costs" in his 1937 article on "The Nature of the Firm," he was the first to introduce the concept in order to explain the existence of firms instead of organizing economic activity through exchange of transactions across the market (Coase 1937). According to Coase, "we had a factor of production, management, whose function was to coordinate. Why was it needed if the pricing system provided all the coordination necessary?" (Coase 1993). Before Coase, creating a firm or using the market were viewed as alternative modes for coordinating production. At the time, the mainstream microeconomic assumption was that the use of the firm or the market for production was a given, not a choice (Williamson 2010).

Coase was the first to observe that firms arise because there are substantial costs involved in using the price mechanism of the market. At first, he was not very explicit about the meaning of these transaction costs. He did not provide a clear definition, but described transaction costs as the cost of "discovering what the relevant prices are" or "negotiating and contracting costs" (Coase 1937). Back in 1937, Coase did not fully grasp his own accomplishment and was merely trying to reveal the weaknesses of the dominant Pigouvian analysis of the divergence between private and social products (Coase 1993). In the late 1950s, research on market failures started, notably after the article of Samuelson on "The Pure Theory of Public Expenditures" (Samuelson 1954). At the time, few authors (e.g., Coase, Buchanan, Calabresi) considered externalities not as a market failure, but pointed out that under certain conditions "harmful effects" are not problematic for the efficiency of market mechanisms (Marciano 2011).

Transaction Costs: The Sequel

In 1960, Coase published another article on the "Problem of Social Cost" which is unmistakably one of the most cited articles in the economic and legal literature and helped launch the economic analysis of the law (Medena 2011).

In analyzing externalities (or "social cost issues" as they were called back then), Coase used the famous example of a rancher whose cattle destroys crops on the land of a neighboring farmer. First, it is assumed that the cattle rancher is fully liable for harm caused "and the pricing system works smoothly," i.e., a zero transaction cost model. Under liability, the efficient level of both cattle and crops will be produced, either through damage payments from the rancher to the farmer or through compensatory payments from the rancher to the farmer to take land out of cultivation (if that is the least-cost solution). This way, external costs are fully internalized under the liability rule. The actual level of damage payments is determined by "the shrewdness of the farmer and the cattle-raiser as bargainers" (Coase 1960).

Second, Coase discusses a situation in which the rancher cannot be held liable for the damage caused by his cattle. In this case, the resulting level of production of both cattle and crops is also efficient, since it is in the farmer's own interest to pay the rancher to cut the size of his herd up till the point where the payment is less than the benefit from reduced crop damage. Similarly, the rancher is willing to cut the size of his herd if the farmer's payment at least equals the foregone benefit resulting from the herd reduction. Again, external costs are fully internalized and both cattle and crop outputs are equal at the levels under liability.

Coase concludes that the initial assignment of legal property rights has no impact on the efficient use of resources. As discussed in the previous paragraph, both situations liability and non-liability lead to an efficient level of production of both cattle and crops. However, this conclusion only holds when the pricing system works at zero transaction costs. This insight has become known as "the Coase theorem." Interestingly, it was Stigler who coined this term in 1966 (Stigler 1966). In case of multiple farmers, substantial transaction costs may arise. Under liability, each individual farmer sues the rancher, hence transaction costs are low and the efficient level of production is attained. In case of non-liability, however, each individual farmer faces three inefficient options: (1) do nothing and bear the costs of crop damage, (2) build a fence around the individual property which on aggregate may outweigh the costs of herd reduction, or (3) try to negotiate a herd reduction which may cause substantial transaction costs, such as gathering information, contacting and discussing with other farmers, and negotiating with the rancher. Putting aside the potential free-rider problem, the latter "transaction" costs, however, can be so substantial that the farmers resort to option one or two.

Later on, Coase (1960) clarified that "to carry out a market transaction, it is necessary to discover who it is that one wishes to deal with, to inform people that one wishes to deal and on what terms, to conduct negotiations leading up to a bargain, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed and so on." The

costs that accompany these activities may hamper transactions that would have taken place if using the pricing system did not evoke such costs (Coase 1960). Furthermore, he suggests that the existence of externalities can partly be explained by the presence of transaction costs that are sufficiently large to prevent market-functioning mechanisms to internalize external costs. At the time, Coase criticized the neoclassical Pigouvian model for ignoring the existence of transaction costs. In 1993, Coase pointed out that his article had demonstrated "the emptiness of the Pigouvian analytical system" and helped to frame the discussion on externalities in a more realistic way (Coase 1993). The Coase theorem was initially met with a lot of skepticism by other scholars who condemned the underlying assumption of efficient markets as unrealistic and even "Utopian" (Blum and Kalven 1967).

Transaction Costs: Definitions

The publication of "The Problem of Social Cost" in 1960 did not lead to the immediate absorption of the idea of transaction cost reasoning in economic literature. In 1969 Kenneth Arrow defined transaction costs as "the costs of running the economic system" (Arrow 1969). But, it was not until 1985 that transaction cost reasoning became more widely known due to Oliver Williamson, who defined it as "the economic equivalent of friction in physical systems" (Williamson 1985). In an effort to put the concept of transaction costs into practice, Williamson explains: "Transaction cost economics is an effort to better understand complex economic organization by selectively joining law, economics, and organization theory. As against neoclassical economics, which is predominantly concerned with price and output, relies extensively on marginal analysis, and describes the firm as a production function (which is a technological construction), transaction cost economics (TCE) is concerned with the allocation of economic activity across alternative modes of organization (markets, firms, bureaus, etc.), employs discrete structural analysis, and describes the firm as a governance structure

(which is an organizational construction). Real differences notwithstanding, orthodoxy and TCE are in many ways complements – one being more well-suited to aggregation in the context of simple market exchange, the other being more well-suited to the microanalytics of complex contracting and nonmarket organization” (Williamson 2008).

In turn, Barzel (1997) described transaction costs as “the transfer, capture, and protection of exclusive property rights.” This is a rather narrow definition of transaction costs, yet often used by scholars from the property rights movement.

More recently, the scope of transaction costs was broadened, leading Challen (2000) to state that transaction costs include all costs associated with any allocation decision, including the costs of uncertainty. Stavins (1995) claimed that transaction costs are “ubiquitous” in market economies, since parties must find one another to transfer, communicate, and exchange information. Douglass North (1990) went even further and considered transaction costs as a part of production costs.

Nowadays, transaction costs are an essential part of mainstream economics and are being applied with different meanings to organizational structures (e.g., vertical integration), market failures (e.g., externalities), institutional choices (e.g., promotion of clubs), and public choice (e.g., administrative burden). Transaction cost is now a generic term referring to costs occurring when making a transaction in the market. Accordingly, transaction costs can be interpreted as “the costs of any activity undertaken to use the price system” (Demsetz 1997).

Towards a Classification of Transaction Costs

The discussion above concerning the definition of transaction costs clearly shows the need for a coherent and complete classification of transaction costs. Dahlman (1979) was one of the first scholars to propose a categorization of transaction costs. In accordance with Crocker (1971), he distinguished three types of costs: (1) search and

information costs, (2) bargaining and decision-making costs, and (3) monitoring and enforcement costs. However, Milgrom and Roberts (1992) used another classification of transaction costs: on the one hand, costs stemming from information asymmetries and incompleteness of contracts among parties and, on the other hand, costs following imperfect commitments or opportunistic behavior of parties. Furthermore, Foster and Hahn (1993) brought some new elements into the discussion and emphasized the distinction between direct financial costs (of engaging in trade), costs of regulatory delay, and indirect costs (associated with the uncertainty of completing a trade). A more basic classification is made by Dudek and Wiener (1996) who included search, negotiation, approval, monitoring, enforcement, and insurance costs.

One of the most interesting classifications of transaction costs is the one by Furubotn and Richter (1997). They describe transaction costs as the costs of establishing, maintaining, adapting, regulating, monitoring, and enforcing rules as well as executing transactions. Their definition of transaction costs is “the costs of resources utilized for the creation, maintenance, use, change, and so on of institutions and organizations. [...] When considered in relation to existing property rights and contract rights, transaction costs consist of the costs of defining and measuring resources or claims, plus the costs of utilizing and enforcing the rights specified. Applied to the transfer of existing property rights and the establishment or transfer of contract rights between individuals (or legal entities), transaction costs include the costs of information, negotiation, and enforcement.” The authors identify three sorts of transaction costs: the costs of using the market (market transaction costs), the costs of exercising the right to give orders within the organization (managerial transaction costs), and the costs of running and adjusting a political system (political transaction costs). Each of these three categories comprises both fixed transaction costs (setup costs for institutional arrangements) and variable transaction costs (dependent on the number of transactions). Furthermore, it should be noted that the authors make a useful distinction between *ex ante* (e.g.,

search and information costs) and ex post (e.g., monitoring and enforcement costs) transaction costs. Thus, the Furubotn and Richter classification integrates the costs of using the market as mentioned by Coase, the managerial costs identified by Williamson, and the institutional costs put forward by North.

Unfortunately, the Furubotn-Richter taxonomy has one serious drawback. In their effort to provide a comprehensive classification, they associate “transaction costs” with the use of markets as well as regulations, which undermines the very meaning of the concept. Furubotn and Richter (1997) rightly claim that regulation entails costs. However, these costs are precisely the opposite from the (Coasean) transaction costs, which refer to the use of the market and thus terminologically unsound. Moreover, the policy goal of regulations is precisely to reduce transaction costs. Adding to the confusion, the OECD in 2001 also made the distinction between non-policy-related transaction costs (in which parties incur costs of voluntary market transactions) and policy-related transaction costs (resulting from the implementation of public policy).

It is clear that the distinction between transaction costs and regulatory costs should be strictly observed. Therefore, Marneffe and Vereeck (2011) suggest that the term “regulatory costs” is exclusively used to refer to the costs of interfering in, correcting, or barring the use of markets. They recommend to short-term the policy-related costs of regulation as “regulatory costs” and non-policy-related costs of using markets as “transaction costs.”

Cross-References

- Coase, Ronald
- Coase Theorem

References

- Arrow K (1969) The organization of economic activity: issues pertinent to the choice of market versus non-market allocation. In: US Joint Economic Committee (ed) The analysis and evaluation of public expenditure: the PPB system, vol 1. Government Printing Office, Washington, DC, pp 59–73
- Barzel Y (1997) Economic analysis of property rights. Cambridge University Press, Cambridge
- Blum WJ, Kalven H (1967) The empty cabinet of Dr. Calabresi auto accidents and general deterrence. *Univ Chicago Law Rev* 34(2):239–273
- Challen R (2000) Institutions, transaction costs and environmental policy: institutional reform for water resources. Edward Elgar, Cambridge, MA
- Coase R (1937) The nature of the firm. *Economica* 4:386–405
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Coase R (1993) Law and economics at Chicago. *J Law Econ* 36:239–254
- Crocker TD (1971) Externalities, property rights and transaction costs: an empirical study. *J Law Econ* 14:445–463
- Dahlman CJ (1979) The problem of externality. *J Law Econ* 22:141–162
- Demsetz H (1997) The firm in economic theory: a quiet revolution. *Am Econ Rev* 87:426–429
- Dudek DJ, Wiener JB (1996) Joint implementation and transaction costs under the climate change convention. Restricted Discussion Document ENV/EPOC/GEEI (96)1. OECD, Paris
- Foster V, Hahn RW (1993) Emission trading in LA: looking back to the future. American Enterprise Institute, Washington, DC
- Furubotn EG, Richter R (1997) Institutions and economic theory: the contribution of the new institutional economics. University of Michigan Press, Ann Arbor
- Marciano A (2011) Ronald Coase, “The problem of social cost” and The Coase theorem: an anniversary celebration. *Eur J Law Econ* 31:1–9
- Medena SG (2011) A case of mistaken identity: George Stigler, the problem of social cost and the Coase theorem. *Eur J Law Econ* 31:11–38
- Milgrom P, Roberts J (1992) Economics, organization, and management. Prentice-Hall, New York
- North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
- OECD (2001) Domestic transferable permits for environmental management: design and implementation. OECD Proceedings, Paris
- Rao PK (2003) The economics of transaction costs: theory, methods, and applications. Palgrave MacMillan, New York
- Samuelson P (1954) The Pure Theory of Public Expenditure. *Review of Economics and Statistics* 36:387–389
- Marneffe W, Vereeck L (2011) The meaning of regulatory costs. *European Journal of Law and Economics* 32:341–356
- Stavins R (1995) Transaction costs and tradable permits. *J Environ Econ Manag* 29:133–148
- Stigler G (1966) The theory of price, 3rd edn. Macmillan, New York
- Williamson O (1985) The economic institutions of capitalism. Free Press, New York

Williamson O (2008) Handbook of new institutional economics part I. Springer, Berlin/Heidelberg

Williamson O (2010) Transaction cost economics: the natural progression. *Am Econ Rev* 100:673–690

Transferable Discharge Permits

Cornelia Ohl

HA Hessen Agentur GmbH, Transferstelle für Emissionshandel und Klimaschutz, Wiesbaden, Germany

Abstract

There are different approaches for dealing with water, air, and soil pollution, for example, the prescription of an emission or immission standard by environmental law or pollution reduction by economic instruments like *transferable discharge permits* (TDP). The idea of TDP is to control the pollution level in the environmental media by economic incentive setting (see, e.g., Tietenberg T, Lewis L (2014) *Environmental & natural resource economics*, global edition, 10th edn. Pearson Higher Education; for an introduction in economic theory). This requires the setting of a critical threshold level for pollution control – e.g., a safe minimum standard – and the breakup of this standard (“cap”) in permits that can be traded on a market.

The most prominent application of TDP in Europe is the trading of CO₂ emissions in selected sectors of the economy (EU-ETS; see, e.g., Ellerman D, Convery F, de Perthuis C (2010) *Pricing carbon: the European Union emissions trading scheme*. Cambridge University Press, Cambridge, UK/New York (also published in French: *Le Prix du Carbone: Les enseignements du marché du carbone*, London: Pearson, 2010); Endres and Ohl (Eur J Law Econ 19:17–39, 2005)). It can be seen as a flagship approach for a European-Union-wide harmonization of environmental laws and regulations for emission control by economic incentive setting.

Synonyms

Cap-and-trade approach; Emissions trading; Tradeable discharge permits; Trading of allowances

What Is It About? Proposal for an Alternative Heading: The Idea of Transferable Discharge Permits

There are different approaches for dealing with water, air, and soil pollution, for example, the prescription of an emission or immission standard by environmental law or pollution reduction by economic instruments like *transferable discharge permits* (TDP). The idea of TDP is to control the pollution level in the environmental media by economic incentive setting (see, e.g., Tietenberg and Lewis 2014 for an introduction in economic theory). This requires the setting of a critical threshold level for pollution control – e.g., a safe minimum standard – and the breakup of this standard (“cap”) in permits that can be traded on a market.

The most prominent application of TDP in Europe is the trading of CO₂ emissions in selected sectors of the economy (EU-ETS; see, e.g., Ellerman et al. 2010; Endres and Ohl 2005). It can be seen as a flagship approach for a European-Union-wide harmonization of environmental laws and regulations for emission control by economic incentive setting.

Market Design by Environmental Legislation

The standard can be fixed on different backgrounds, on social welfare considerations, aspects of human health, and nature protection, among others. In reality, we often find a mixture of different factors including political, economic, and natural science considerations. In any case, the standard is to split into permits which with regard to the regulated subject and area allow a certain amount of pollution and are valid for a certain time period. This poses questions of monitoring and measurement.

Pollution is usually a by-product of valued goods and services which make the measurement tricky, especially in cases where direct measurement is impossible and indirect calculations or assessments are required. To ensure that the assigned amount can be traded, the certified metric needs to mean the same for any regulated body irrespective of the type of good or service it provides. For this reason, clear regulations on monitoring and measurement are necessary. It is also to ensure that the number of permits certifies the critical threshold value set by the regulating authority.

In a further step, the regulating body is to issue the permits either free of charge or by selling them for a fixed price or through bidding or auctioning mechanism. Moreover, to guarantee that the polluters perform with the standard, reporting on the polluting activities is essential. It calls for the establishment of an accounting system that can be verified by independent experts. This supports proving performance with monitoring and measurement rules and individual compliance, i.e., if the polluters keep their polluting activity within the limit allowed by the number of permits they hold. If the polluting behavior is inconsistent with the assigned amounts, an enforcement mechanism is needed that makes the polluter stick to its obligations, for example, a penalty fee in combination with a belated reduction of excessively released pollutant.

Incentives for Permit Trading

Keeping within the limits of a standard frequently requires measures for emission/immission reductions. These measures raise different costs for the regulated bodies. To keep them at minimum, economists argue for the establishment of a market where the permits can be traded.

Although the market does not guarantee that the polluters with the lowest cost actually reduce the pollutant – the polluter is generally free to decide whether to buy permits or change the polluting behavior – it nevertheless encourages minimizing the cost of pollution control. If polluters maximize profits, they will take advantage of the cheapest way of pollution reduction. The polluter thus has incentive to compare the market price of a permit

with the individual reduction costs. As long as the cost for buying permits is lower than the cost of measures for pollution control, the polluter shows a demand for permits. On the other hand, polluters with low control costs are willing to reduce their pollution level in order to sell permits on the market, at least as long as the price they receive for the permit is higher than the individual reduction costs. This mechanism ensures that the standard is enforced at minimal costs.

Despite this advantage, the market approach is often criticized. One argument is that polluters with high reduction costs lose incentives for pollution control and let others do the job. The question, however, is whether society is concerned of the pollutant to stay within the prescribed limit or whether the concern is on who is responsible for taking measures. If the goal is pollution control, the market-based approach has clear advantages in terms of both social welfare and environmental protection: it first of all draws the focus on pollution reduction to the desired extent and second to incentive setting for cost minimization. The question of responsibility, nevertheless, can be addressed by selecting the group of polluters having to perform with the environmental regulation.

Applications and refinements of TDP as well as further criticism are found in the literature (e.g., see Hansjürgens et al. 2011; OECD 2004).

Cross-References

- [Anticommons, Tragedy of the](#)
- [Coase and Property Rights](#)
- [Economic Efficiency](#)
- [Emissions Trading](#)
- [Law and Economics](#)
- [Market Definition](#)

References

- Ellerman D, Convery F, de Perthuis C (2010) Pricing carbon: the European Union emissions trading scheme. Cambridge University Press, Cambridge, UK/New York (also published in French: *Le Prix du Carbone: Les enseignements du marché du carbone*, London: Pearson, 2010)

- Endres A, Ohl C (2005) Kyoto, Europe? – an economic evaluation of the European Emission Trading Directive. *Eur J Law Econ* 19:17–39
- Hansjürgens B, Antes R, Strunz M (2011) Permit trading in different applications. Routledge, New York
- OECD (2004) Tradeable permits: policy evaluation, design and reform. OECD Publishing, Paris. <https://doi.org/10.1787/9789264015036-en>
- Tietenberg T, Lewis L (2014) Environmental & natural resource economics, global edition, 10th edn. Pearson Higher Education, Universal Free E-Book store

Further Reading (documents on EU-ETS)

- Commission Decision 2006/780/EC of 16 November 2006 on avoiding double counting of greenhouse gas emission reductions under the Community emissions trading scheme for project activities under the Kyoto Protocol pursuant to Directive 2003/87/EC of the European Parliament and of the Council [Official Journal L 316 of 16 November 2006]
- Commission Decision 2007/589/EC of 18 July 2007 establishing guidelines for the monitoring and reporting of greenhouse gas emissions pursuant to Directive 2003/87/EC of the European Parliament and of the Council [Official Journal L 229 of 31.8.2007]
- Commission Regulation (EU) No 1031/2010 of 12 November 2010 on the timing, administration and other aspects of auctioning of greenhouse gas emission allowances pursuant to Directive 2003/87/EC of the European Parliament and of the Council establishing a scheme for greenhouse gas emission allowances trading within the Community [Official Journal L 302 of 18.11.2010]
- Directive 2003/87/EC of the European Parliament and of the Council of 13 October 2003 establishing a scheme for greenhouse gas emission allowance trading within the Community and amending Council Directive 96/61/EC

Transition Economies: Rule of Law and Credible Commitment

Rosolino A. Candela¹ and Ennio Piano²

¹Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

²Department of Economics, George Mason University, Fairfax, VA, USA

Abstract

Must the rule of law spring upon us solely by accident and force, or can it also emerge as a product of political reflection and choice? As

the lessons from the political and economic transition of centrally planned economies in Europe and Asia illustrate, the transitional movement requires an element of political reflection and deliberate choice, specifically the establishment a binding and credible commitment to limits on state action. In order to do so, credible commitment to the rule of law must be signaled *ex ante* by political reformers if reform efforts are to be successful, from which economic development follows *ex post*. The historical record shows, however, that efforts to establish the rule of law in transition economies have been mixed. This chapter explains why this is the case, specifically by comparing Russia and China as case studies of transitional political economy.

Introduction

Must the rule of law spring upon us solely by accident and force, or can it also emerge as a product of political reflection and choice? In a previous chapter of this encyclopedia, we discussed the role in which interjurisdictional competition between states and intrajurisdictional competition between interest groups constrained governments to obey its political commitments to enforce property rights, in which the rule of law emerged as a by-product of this process. However, as the lessons in the political and economic transition of centrally planned economies show, the transitional movement requires an element of political reflection and deliberate choice (see Hayek 1973: 45; Wagner and Runst 2011), specifically the establishment of a binding and credible commitment to limits on state action (Boettke 2009). In order to do so, credible commitment to the rule of law must be signaled *ex ante* by political reformers if reform efforts are to be successful, from which economic development follows *ex post*. The historical record shows, however, that efforts to establish the rule of law in transition economies have been mixed. This chapter explains why this is the case, specifically by comparing Russia and China as case studies of transitional political economy.

The Problem of Credible Commitment

There lies a fundamental dilemma in every attempt to establish the rule of law in those economies undergoing a transition away from national economic planning: how do we resolve the paradox of governance? In order to create the conditions conducive to exchange and capital accumulation, governments must be strong to enforce property rights and contractual arrangements. Doing so enables individuals to devote less effort and resources preventing “first-level predation,” meaning predation by private actors, and specialize in productive entrepreneurial activities. Eliminating first-level predation, however, introduces the problem of “second-level predation,” or predation by public actors (Boettke 2009: 47; see Acemoglu and Johnson 2005). Governments that are strong enough to enforce private property rights and prevent private predation are also strong enough to break any constraints to such precommitments in the future. Public predation can manifest itself in several ways, including the outright confiscation of wealth, taxation, and inflationary finance through currency debasement (Montinola et al. 1995: 54). The cost of such predation in terms of economic development is “evasive entrepreneurship,” which includes the expenditure of resources evading the legal system (Coyne and Leeson 2004: 57), via tax evasion, bribery, or withholding efforts to invest in physical and human capital. If government is given enough power to enforce the rules necessary to maintain an economy, then it must also credibly commit to honor such rules.

Why is it difficult for government officials to credibly commit to the rule of law? This can be explained by the absence of a “political Coase theorem” (Acemoglu 2003). The Coase theorem states that when private property rights are well defined and transactions are low, individuals will bargain to an efficient outcome, irrespective of the initial assignment of property rights (Coase 1960; Stigler 1966: 113). Implicitly, the Coase theorem assumes that parties to an exchange are full residual claimants and will therefore bear the full costs and benefits of the outcomes of their decision-making. Therefore, if one party to an exchange

breaks an agreement, or if both parties are disputing an agreement, reputational mechanisms or the state is available to enforce contractual agreements. Failure to abide by an agreement will result in a loss of reputation in future trading periods and/or state punishment. The Coase theorem applied to the context of political decision-making, however, is quite different. The logic of political decision-making is to concentrate benefits on well-informed and well-organized interest groups that represent a politician’s constituency and disperse costs on the masses of the ill-informed population (Boettke 1993: 7). Such a logic prevails because the governmental official who promises institutional reform to their citizenry will not bear the full costs of renegeing on such a promise. Whereas in the previous example, a private actor faces costs in terms of state punishment and loss of reputation, government officials only bear a loss in reputation. However, even these costs are not fully concentrated on the particular public actor, but are spilled over onto future generations of public actors, who face an increasingly distrustful citizenry, making it more difficult in future “trading periods” to implement transitional reforms.

In order to overcome the expectations that the citizenry hold with regard to credibly committing to the rule of law, governments must signal their willingness to tie their hands *ex ante* against using discretion to unforeseen economic circumstances *ex post*. Signaling refers to the transmission of information that is costly for a sender to emit, but is “cheap” for another party to receive, in this case government officials being the senders and the citizenry being the recipients. Without a credible commitment to the rule of law via signaling, “it is of course perfectly rational for private agents to discount announcements of future policy reforms – or assurances of the continuation of present reforms” (Rodrik 1989: 757). Rodrik argues that any successful institutional reform will have a bias towards “overshooting” in its signaling strategy. This implies that there will be inverse relationship between policy overshooting and credibility: the more severe the credibility gap, the greater the policy must overshoot in order to send the appropriate signal and overcome

the problem of credible commitment. For example, a central banker who wishes to credibly commit to sound monetary policy could adopt a rule that pegs the exchange rate, allowing citizens to convert the domestic currency into foreign currency at a preannounced rate. However, a central bank with a long record of inflationary policy will be sending a noisy signal, especially if devaluation is expected in the face of speculative attacks. In order to overcome the gap of credibly committing to currency debasement, policymakers may have to overshoot even further by adopting a more binding constraint that eliminates its discretion completely. Examples include dollarization or the private-issuance of competitive currencies (Selgin and White 2005; see also White 2010). Therefore, establishing a credible commitment to the rule of law not only requires sending a strong signal to execute institutional reform, but also establishing a binding constraint that eliminates political discretion. Once this takes place, economic and political transition can move into a wealth-creating path.

The Rule of Law and Transitional Political Economy

Fundamentally, transitional political economy entails an institutional change in the manner in which property rights are structured and governed, both in terms of formal government enforcement and informal institutional arrangements, such as customs, traditions, and norms. Broadly speaking, property rights refer to social relationships that guide expectations about the use of scarce resources (Furubotn and Pejovich 1972). Political and economic transitions are intertwined by changes in the de facto structure of property rights. The transition to a private property rights-based economy, based on freedom of exchange and freedom of contract, leads to a reallocation of wealth previously based on political connection and patronage. Respect for private property under the rule of law promotes political freedom because it separates economic power from political power and in this way enables the one to offset the other. It enables economic strength to be a check to

political power rather than its reinforcement (Friedman 1962 [2002]: 15). Such an institutional transition will imply transaction costs that are not only economic and political, but cultural and historical as well.

Implementation of the rule of law in transitional economies implies the elimination of existing political privileges, specifically de facto control over the use of resources. Therefore, economic and political transition will generate a massive redistribution in income. This in turn will imply the opposition of a large class of government officials and military personnel whose rank and privileges will be threatened by transition to market economy under the rule of law. Theoretically, this problem could be eliminated by paying those individuals the present discounted value of the income derived from holding political power, particularly through de facto control over state-owned firms and its resources. However, when the transitional costs associated with compensating the current benefactors of the existing system are greater than the associated welfare gains dispersed among the population, this will create what Gordon Tullock refers to as a “transitional gains trap” (Tullock 1975) that impedes the implementation of the rule of law.

The process of institutional transition from a socialist economy to a market economy is not only economic and political in nature, but also cultural (Pejovich 2003). Culture refers to the context where goals of individuals and the means to be employed by individuals are shaped and given meaning (Storr 2013: 54). The way in which property rights are perceived and “what constitutes an appropriate disposition of property, are all (partly) determined by culture” (Storr 2013: 32). Therefore, culture will affect the ability for governments to credibly commit to uphold the rule of law. Steve Pejovich has argued that “*the cultural differences between Central and East European countries are a major determinant of the magnitude of their respective transaction costs*” (emphasis original, Pejovich 2003: 352). Countries that have a “command culture,” which perceives the exchange of property rights to be a zero-sum game, will face higher costs of credibly committing to the rule of law than a “culture of

exchange” that regards the exchange of property rights to be mutually beneficial (see Buchanan 1997: 95–101). This is because if the “gains from trade are seen as a redistribution of income rather than as rewards to innovators for creating new wealth,” then “[s]tate authorities are more likely to impose price controls on producers and/or merchants who earn large profits than to seek ways to create incentives for others to emulate such individuals in competitive markets” (Pejovich 2003: 351). The costs, however, of institutional transition present potential profit opportunities to be monetized by institutional entrepreneurs to change the rules of game, namely, through the establishment of private property rights (Leeson and Boettke 2009; Li et al. 2006). The role of institutional entrepreneurship will be discussed in the next section.

Moreover, not only does culture matter, but ideas also matter as well, namely, through the footprint of history. For example, one of the obstacles that China encountered during the early stages of its transition to a market economy, beginning after 1978, was the “Illusion of 1957” (Cheung 1982 [1986]: 26–27). According to this illusion, the Chinese regarded capitalism to be worse than communism by equating “capitalism” with the cronyism and corruption of the rule of the Kuomintang prior to 1949, which preceded a period of economic recovery under “communism” between 1949 and 1957. “Thus within living memory the Chinese people equate capitalism with the Kuomintang *débâcle* and communism with the ‘good years’ of 1949–1957” (emphasis original, Cheung 1982 [1986]: 28). Such historical perceptions, however false and misleading they may be, may only create further institutional inertia towards a complete and credible transition to a market economy.

Given all of these difficulties inherent not only to the problem of credible commitment via signaling, but also the transaction costs inherent to institutional transition, how have some countries been able to transition to a market economy under the rule of law and achieve economic development? Peter Boettke outlines the steps that must be followed to establish a path towards successful political and economic reform: “Reform in real

time must (1) start from the existing status quo, (2) unearth the *de facto* organizing principles of that status quo, (3) design a set of reforms which address the incentive and informational problems associated with that *de facto* system, and (4) send a clear high-quality signal that the proposed reforms are credible and commit the governance structure to the new system and in doing so close the gap for the *de jure* and *de facto* organizing system in the new regime” (emphasis original, Boettke 1999: 378). In the next section, we will show how this transition has unfolded under comparative systems of federal governance, using the transitional political economy of Russia and China as comparative case studies.

Russia and China: A Comparative Transitional Analysis

One way to explain this institutional transition is to compare the transitional path that Russia and China have taken. In doing so, we make a distinction between “market-preserving federalism” (Weingast 1995; Qian and Weingast 1997) and “cartel-federalism” (Wagner and Yokoyama 2013; Wagner 2016). Under a federalist system of governance, governance is divided into two levels of authority: a national level of authority and subnational level of authority of smaller local government entities. The distinction between market-preserving federalism and cartel-federalism is based on how the national level of governance sets the conditions of competition between subnational governments.

Under market-preserving federalism, the national government credibly commits to the rule of law, namely, by allowing the entry and exit of private-sector firms to compete with state-owned firms. As a result, the wealth created through private-sector firms within competing local jurisdictions will grow, eroding the relative value of the rents derived from the political control of state-owned firms. Therefore, with the growing extension of the market, the allocation of entrepreneurial talent becomes redirected towards productive activities, rather than unproductive activities, such as rent-seeking

(Murphy et al. 1991; Tullock 1967). However, decentralization and competitive, or market-preserving, federalism are not synonymous. For example, the Soviet Union was “decentralized” in the sense that policy was administered by local governments, but local governments lacked political autonomy and local authority over the economy (Montinola et al. 1995: 57).

Under cartel-federalism, the national government creates a framework of collusion among subnational political entities. Like a cartel of firms, subnational governments collude to act as a collective monopoly, with the government acting as the *de facto* enforcer against attempted “chiseling” by local governments. In effect, cartel federalism prevents the erosion of rents derived from political control of state-owned firms and resources, namely, by obstructing the introduction of pro-growth policies among local governments that would encourage the growth of the private sector of the economy. The different institutional arrangements adopted by Russia and China since the late 1980s are reflected in their respective economic trajectories. Ten years after the transition reforms, Russia’s GDP per capita had *fallen* by almost 30% (Leeson and Trumbull 2006). The same figures had more than doubled in China over the same period of time (World Bank 2017).

Whereas Russia’s transition can be characterized by cartel federalism, China’s transition followed more closely a model of market-preserving federalism. Under a model of market-preserving federalism, China’s *de jure* changes were far less important than the *de facto* changes that emerged spontaneously among competing local jurisdictions. However, the sequence of events which we describe should not undercut “the essential element in political and economic transitions of post communism – the establishment of a binding and credible commitment to liberal limits on state action” (Boettke 2009: 43). While the emergence of private property in China was not designed by government policy, this bottom-up reform would not have flourished without a credible commitment by the state not to obstruct the *de facto* changes in property rights. After the rise of Deng Xiaoping in 1978, the “first priority under the new economic policy was

agriculture” (Coase and Wang 2012: 157). However, agricultural reform was not initiated by the *de jure* privatization of farmland. Although land is still formally owned by the state, the introduction of a “household responsibility system” has led to a *de facto* reallocation of these rights to local stakeholders.

Under this system of landownership, peasants are allowed to lease an exclusive plot from the government, leaving the responsibility contract holder to grow and sell crops of their choice. By 1982, these responsibility contracts began to be traded among peasants, and such transfers became formally permitted by the government in 1983 (Cheung 1982 [1986]: 66). The household responsibility system arose in 1978, initially out of the institutional entrepreneurship of Yan Junchang, who was a villager in Xiao Gang, a poverty-stricken village in Anhui province of China. As Li, Feng, and Jiang describe the account, “[s]truggling to escape absolute poverty, on November 24, 1978, he and 17 other farmers signed a secret agreement to divide up the land and let each household work by itself, running the risk of jail sentences. They had the implicit support of local reform-minded officials. One year later, their innovation proved to be a big success: the total grain in production was equal to the sum of production over the previous five years” (2006: 245; see also Coase and Wang 2012: 47). This *de facto* privatization of property in agriculture occurred simultaneously with the rise of commerce in special economic zones (SEZs), which was first established in the Guangdong province on August 26, 1980, to attract foreign capital and investment (Coase and Wang 2012: 62).

It is no accident that the rise of the household responsibility system and SEZs coincided with institutional reforms that limited state action in terms of taxation and regulation within Chinese provinces. Starting in 1980, China instituted a fiscal revenue-sharing system between the provincial governments and the central government. According to this fiscal arrangement, revenue income in each province is divided between a fixed shares of revenue, which is remitted to the central government, allowing the remaining share of tax revenue to remain within the local

jurisdiction (Montinola et al. 1995: 63). Moreover, the Communist Party has retained authority to appoint and dismiss governments according to their ability to foster pro-growth policies. As an added incentive, and perhaps “as an ultimate prize, the governors whose regions perform well have been brought into the national government in Beijing” (Blanchard and Shleifer 2001: 175). This choice of institutional design addressed the incentive and informational problems associated with that *de facto* system, namely, by incentivizing and selecting for those local officials whose interests would be aligned with the pursuit of pro-growth policies that encourage increased labor productivity and capital accumulation.

Although China remains undemocratic politically speaking, its ability to achieve rapid growth since 1978 could not have occurred without a credible commitment to limits on political discretion and public predation, without which property rights in land, labor, and capital would not have emerged to foster economic development. The transitional political economy of Russia, however, can be characterized by cartel federalism. Although Russia has held democratic elections and has engaged in a massive privatization scheme, such *de jure* reforms have not accompanied *de facto* reforms. This has occurred because of the failure of political officials to signal a credible commitment to the rule of law.

The Russian experience differed drastically from China's. Boettke (1993, 1995) argues that the failure of the attempts to reform the Russian economy (first under Gorbachev and then under Yeltsin) is due to that the lack of an effective rule of law. Absent the rule of law, Russian citizens had to predict whether the government would go through with reform or, instead, reverse back to centralized control of the economy and social and political life. This prediction was rooted on the historical experience with economic reform in Russia since the time of Lenin: every time the Soviet government had announced a movement towards the devolution of economic and political decision-making, this decision was then reversed in the span of years and even months. With this reversion often came the persecution of those who had taken advantage of the new economic

opportunities. It is not surprising that the Russian people did not accept the promises of the Gorbachev government at face value. Moreover, the behavior of the regime was itself sending mixed signals to the population, making the unraveling of the “reform game” even more likely. The three major pieces of legislation under *Perestroika* (the Law on Individual Enterprises, the Law on State Enterprises, and the Law on Cooperatives) aimed at increasing the competitiveness of the Soviet economy were all radically modified during the (brief) period of their implementation. “Despite the rhetoric and promise of these laws,” Boettke writes, “they contained contradictions and ambiguities that prevented them from achieving the objectives of economic reform. Furthermore, they failed to convey any binding commitment on the part of the Gorbachev regime to true market reform. From 1985 to 1991, Gorbachev introduced at least ten major policy packages for economic reform under the banner of *perestroika*, but not a single one was fully implemented” (emphasis added, 1995 [2001]: 166).

As Russia began its process of privatization during the 1990s, it did so without a complete transition of its political institutions, within which privatization takes place. Without political reform of the institutions within which political decision-making takes place, the incentives that politicians faced during the transition period remained the same. As Milton Friedman points out, “Russia privatized but in a way that created private monopolies, private centralized economic controls that replaced government's centralized controls. It turns out that the rule of law is probably more basic than privatization. Privatization is meaningless if you don't have the rule of law. What does it mean to privatize if you do not have security of property, if you can't use your property as you want to?” (Friedman 2002: viii). The inconsistency between the spirit and the letter (and application) of the reform plan only added to the skepticism of the Russian people. If the past behavior of the Soviet government suggested that its stated goals were inauthentic, its present behavior only confirmed these suspicions.

Shliefer (1997) and Blanchard and Shleifer (2001) identify another cause of Russia's reform failure, namely, its fiscal institutions, which contrast with that of China. Given the lack of the rule of law, the introduction of a decentralized political system is unlikely to lead to the desired outcome. Unlike in China, where tax revenues are raised locally, this encourages local political officials to expand their tax base by fostering economic growth. The political context in Russia, however, was best characterized by cartel federalism, since "local government revenues comes from their share in taxes collected by the central government. Moreover, while this share is in theory fixed, in practice it is negotiated. Regional governments negotiate with Moscow, and local governments negotiate with regions" (Shliefer 1997: 403), creating a collusive rather a competitive environment between regions and disincentivizing the institutionalization of pro-growth policies. Indeed, while Russia's transition has been characterized by greater *de jure* reforms than China, there have been relatively less *de facto* reforms, precisely because of a failure to credibly commit to changing the institutional incentive structure within which economic and political decision-making takes place. As Coase remarked in his Nobel Prize Address on the heels of transition of Eastern and Central Europe, "[t]hese ex-communist countries are advised to move to a market economy, and their leaders wish to do so, but without the appropriate institutions no market economy of any significance is possible" (1992: 714).

Conclusion

The problem of economic transition is fundamentally institutional in nature. For economic and political reforms to be successful in transitional economies, this entails a credible commitment to the rule of law. Absent this commitment, aligning incentives between political and economic actors becomes extremely difficult, condemning the transition process to a likely failure. The experience of postcommunist Russia and postreform China provides evidence for this hypothesis.

Cross-References

- Capitalism
- De Jure/De Facto Institutions
- Development and Property Rights
- Fiscal Federalism
- Hayek, Friedrich August von
- Institutional Economics
- Political Economy
- Privatization
- Rule of Law and Economic Performance
- State-Owned Enterprises

References

- Acemoglu D (2003) Why not a political coase theorem? Social conflict, commitment, and politics. *J Comp Econ* 31:620–652
- Acemoglu D, Johnson S (2005) Unbundling institutions. *J Polit Econ* 113(5):949–995
- Blanchard O, Shleifer A (2001) Federalism with and without political centralization: China versus Russia. IMF Staff Pap 48:171–179
- Boettke PJ (1993) Why perestroika failed: the politics and economics of socialist transformation. Routledge, New York
- Boettke PJ (1995/2001) Credibility, commitment and soviet economic reform. In: Boettke PJ (ed) Calculation and coordination: essays on socialism and transitional political economy. Routledge, New York, pp 154–175
- Boettke PJ (1999) The Russian crisis: perils and prospects for post-soviet transition. *Am J Econ Sociol* 58(3):371–384
- Boettke PJ (2009) Institutional transition and the problem of credible commitment. *Annu Proc Wealth Well Being Nation* 1:41–52
- Buchanan JM (1997) Post-socialist political economy: selected essays. Edward Elgar, Lyme
- Cheung SNS (1982/1986) Will China go 'Capitalist?': an economic analysis of property rights and institutional change, 2nd edn. Institute of Economic Affairs, London
- Coase R (1960) The problem of social cost. *J Law Econ* 3:1–44
- Coase R (1992) The institutional structure of production. *Am Econ Rev* 82(4):713–719
- Coase R, Wang N (2012) How China became capitalist. Palgrave Macmillan, New York
- Coyne CJ, Leeson PT (2004) The plight of underdeveloped countries. *Cato J* 24(3):235–249
- Friedman M (1962/2002) Capitalism and freedom, 40th anniversary edition. University of Chicago Press, Chicago
- Friedman M (2002) Preface: economic freedom behind the scenes. In: Gwartney J, Lawson R (eds) Economic

- freedom of the world 2002 annual report. The Fraser Institute, Vancouver
- Furubotn EG, Pejovich S (1972) Property rights and economic theory: a survey of recent literature. *J Econ Lit* 10(4):1137–1162
- Hayek FA (1973) *Law, legislation and liberty*, Vol. 1: Rules and order. University of Chicago Press, Chicago
- Leeson PT, Boettke PJ (2009) Two-tiered entrepreneurship and economic development. *Int Rev Law Econ* 29(3):252–259
- Leeson PT, Trumbull WN (2006) Comparing apples: Normalcy, Russia, and the remaining post-socialist world. *Post-Soviet Aff* 22(3):225–248
- Li DD, Feng J, Jiang H (2006) Institutional entrepreneurs. *Am Econ Rev* 96(2):358–362
- Montinola G, Qian Y, Weingast BR (1995) Federalism, Chinese style: the political basis for economic success in China. *World Polit* 48(1):50–81
- Murphy KM, Shleifer A, Vishny RW (1991) The allocation of talent: implications for growth. *Q J Econ* 106(2):503–530
- Pejovich S (2003) Understanding the transaction costs of transition: it's the culture, stupid. *Rev Aust Econ* 16(4):347–361
- Qian Y, Weingast BR (1997) Federalism as a commitment to perserving market incentives. *J Econ Perspect* 11(4):83–92
- Rodrik D (1989) Promises, promises: credible policy reform via signaling. *Econ J* 99(397):756–772
- Selgin G, White LH (2005) Credible currency: a constitutional perspective. *Constit Polit Econ* 16(1):71–83
- Shleifer A (1997) Schumpeter lecture: government in transition. *Eur Econ Rev* 41:385–410
- Stigler GJ (1966) *The theory of price*, 3rd Ed. New York: Macmillan
- Storr VH (2013) *Understanding the culture of markets*. Routledge, New York
- The World Bank (2017) *World development indicators*. World Bank Group, Washington, DC. <https://openknowledge.worldbank.org/bitstream/handle/10986/26447/WDI-2017-web.pdf?sequence=1&isAllowed=y>
- Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. *West Econ J* 5(3):224–232
- Tullock G (1975) The transitional gains trap. *Bell J Econ* 6(2):671–678
- Wagner RE (2016) Politics as a peculiar business: insights from a theory of entangled political economy. Edward Elgar, Northampton
- Wagner RE, Runst P (2011) Choice, emergence, and constitutional process: a framework for positive analysis. *J Inst Econ* 7(1):131–145
- Wagner RE, Yokoyama A (2013) Polycentrism, federalism, and liberty: a comparative systems perspective. *J Public Finance Public Choice* 31: 179–197
- Weingast BR (1995) The economic role of political institutions: market-preserving federalism and economic development. *J Law Econ Org* 11(1):1–31
- White LH (2010) The rule of law or the rule of central bankers? *Cato J* 30(3):451–463

Trial: Implicit Biases

Goran Dominioni¹ and Alessandro Romano²

¹Rotterdam Institute of Law and Economics, Erasmus School of Law, Rotterdam, The Netherlands

²China-EU School of Law, China University of Political Science and Law, Beijing, China

Abstract

A large body of research in implicit social cognition indicates that implicit racial biases affect human decision-making. Building on these findings, legal scholars have explored how and under which circumstances implicit racial biases affect the functioning of trial systems. This entry reviews this literature. First, it provides an overview of the main results of, and methods employed in, studies on implicit racial biases. Second, it reviews the applications of these findings for the study of criminal and employment discrimination trials and the policies that could be implemented to reduce the effect of these biases on the decision-making of the relevant actors. It concludes with some suggestions for further research.

Definition

Implicit racial biases are shifts in judgment caused by automatic and/or unconscious attitudes/stereotypes held toward a racial group. In this context, automatic means that the bias occurs without any need for attention and that it is not easily controlled. Whereas unconscious means that introspection does not reveal the attitude/stereotype.

Introduction

In the last decades, studies in behavioral economics and psychology have highlighted the existence of various behavioral patterns that are inconsistent with the rational choice theory (Kahneman 2011 and references therein). Legal scholarship has

highlighted that many of these patterns are relevant for the study of law and policy-making (Jolls et al. 1998). Part of this literature, and especially in the context of trial settings and employment law, focuses on implicit biases (e.g., Jolls and Sunstein 2006; McAdams and Ulen 2009; Teichman and Zamir 2014). Although implicit biases do not relate only to race (but, for instance, also to gender), the wide majority of studies on this topic within behavioral law and economics focus on this particular issue. For this reason, the scope of this entry is restricted to implicit racial biases.

Scientific Basis and Research Methods

Two core concepts in the study of implicit biases are attitudes and stereotypes.

An attitude is a mental association between a racial group and an evaluative disposition (Greenwald and Krieger 2006). This evaluative disposition can be either positive or negative. For example, a person may hold either a positive or a negative attitude toward Asians. Instead, racial stereotypes are an association between a racial group and a positive or a negative characteristic (e.g., being lazy, good in math, or aggressive).

Research on implicit social cognition has studied the relationship between attitudes and stereotypes both at the implicit and the explicit level. This literature indicates that a person can hold an implicit attitude and a stereotype that point in opposite directions. For instance, it is possible to have a negative attitude toward Blacks and yet hold a positive stereotype toward them (e.g., Blacks are good in sports). In addition, existing evidence indicates that self-reported attitudes and stereotypes do not necessarily mirror implicit measures. As a consequence, it is possible for an individual to hold, and maybe act upon, an implicit attitude or stereotype that conflicts with the ones she consciously endorses (Greenwald and Krieger 2006).

Implicit attitudes and stereotypes are identified using implicit measurement procedures. Thanks to these techniques it is possible to identify attitudes and stereotypes that, for a variety of reasons, self-reporting may fail to detect

(e.g., unawareness). Among the most commonly used techniques to measure implicit biases, there are affective priming, the implicit association test (IAT), and brain imaging.

In affective priming, subjects are exposed to two types of stimuli – a target and a prime – and they are asked to categorize the target as either positive or negative. To perform the task the only relevant stimulus is the target; therefore the prime should not affect the behavior of the subjects. For instance, in studies on implicit racial attitudes, subjects are first exposed to pictures of Black faces and White faces (the prime). This exposure occurs subliminally, meaning that pictures are shown for a very short time (e.g., 200 ms), so that they are processed only unconsciously. Subsequently, subjects are exposed to a series of words (the target) and are asked to categorize them (positive/negative). When priming with a racial stimulus that facilitates the categorization of the words as negative, the subject is said to hold a negative attitude toward the group (Fazio et al. 1995).

Another widely used measurement procedure in the implicit biases literature is the IAT (Greenwald et al. 1998). This procedure allows testing the relative strength of implicit attitudes and stereotypes via the measurement of response latencies in the categorization of stimuli into classes. In particular, in a racial IAT, part of the stimuli relates to racial groups (e.g., a name more commonly associated with a Black/White person), while the remaining stimuli relate to other concepts (e.g., good/bad). Subjects are repeatedly asked to categorize each stimulus (e.g., a Black name) as belonging to one of two dyads. Thus, for instance, subjects are first asked to categorize good/bad concepts and Black/White names as belonging to either the Black/good or the White/bad dyad. And, subsequently, the task is repeated with Black/bad and White/good dyads. Differences in response time between different combinations of dyads indicate that a certain racial group is more easily associated with a certain concept.

Another technique often employed in the implicit racial attitude domain is blood-oxygen level-dependent contrast imaging in functional magnetic resonance imaging (Fazio and Olson

2003). This physiological measure identifies variations in oxygen levels in the blood present in different parts of the brain, which are positively correlated with activations of these areas. Thus, for instance, exposure to Black faces has been shown to generate a greater activation of the amygdala (the amygdala is a part of the brain related to the processing of emotions). Therefore, this strand of research indicates that the exposure to different racial groups can trigger distinctive emotional states (Kubota et al. 2012).

Main Findings

The existing research indicates that implicit racial biases are pervasive in the White population of several Western countries. The largest database regarding the diffusion of implicit biases comes from the online platform Project Implicit, which allows visitors to take a racial IAT online. An analysis of the data gathered on this platform between the years 2000 and 2006 indicates that more than 65% of the people who took the racial IAT associated relatively more easily Blacks with bad. In addition, less than 15% of the participants (mainly non-White subjects) held relatively stronger associations between White and bad.

This data also suggests that Blacks are the only racial group that do not show strong pro-White associations (Nosek et al. 2007).

Research in implicit social cognition indicates that various factors mediate the formation of implicit biases. In particular, empirical evidence indicates that implicit biases can stem from past (often forgot) experiences, which may date back to early childhood. These experiences can be either direct or indirect. As indirect experiences – for instance, via the media – generally contribute to the formation of a negative perception of Blacks, this could explain the absence of a pro-Black bias among Black people (Greenwald and Krieger 2006).

A large part of the literature on implicit biases analyzes their impact on human behavior (both spontaneous and deliberate). Two meta-analyses on racial IAT studies find that measurements of implicit biases have a higher predictive validity

than self-reported measures (Greenwald et al. 2009; Oswald et al. 2013).

A meta-analysis of racial priming studies reaches a similar conclusion (Cameron et al. 2012). Overall, this literature suggests that implicit biases can affect behavior, making them especially relevant for the study of discrimination in trial settings. In this regard, it is important to notice that even small effects size for a single decision can predict large discrimination at the societal level, especially when the bias affects many individuals or when it repeatedly affects one individual (Greenwald et al. 2015).

Applications in Trial Settings

Having introduced the background of the topic, let us now analyze how implicit biases relate to trial settings from a law and economics perspective. Broadly speaking, existing research on implicit biases in trial settings unfolds in two directions. First, studies in implicit social cognition made their ways into law reviews because of the impact that implicit biases have on trial actors' behavior. Second, implicit biases have been widely discussed among legal academics in relation to the epistemic value of implicit measurements in the context of evidence law. The second line of enquiry is *sensu stricto* not behavioral and is thus less relevant for the present discussion. In the following, we will therefore focus only on the behavioral side of the debate.

Implicit Biases and Trial Parties Decisions

Earlier law and economics literature depicts trial participants (judges, prosecutors, and lawyers) as rational agents that allocate resources following a rational calculation of the costs and benefits of their actions (McAdams and Ulen 2009). For instance, according to the efficient prosecutor model, the prosecutors maximize convictions (and attach different weights to different sentences) or seek deterrence maximization (Garoupa 2012). Later developments in the field questioned this simplistic description of trial participants and emphasized the role of institutional details and behavioral aspects of human decision-making.

Research on implicit biases is a subset of these developments.

With few exceptions, research on implicit biases in the courtroom has focused on criminal trials (Kang et al. 2012). In these settings, a figure that plays a major role in steering the unfolding of a trial is the prosecutor. Research suggests that implicit racial biases among prosecutors account for at least part of the disparities in incarceration rates between racial minorities and nonracial minorities in the US criminal law system (Smith and Levinson 2011; Kang et al. 2012). And indeed, implicit biases can be particularly effective in distorting human behavior when confronted with decisions that have the following characteristics: (i) allow for some discretion, (ii) have to be taken quickly, and (iii) do not provide accountability mechanisms. The decisions taken by prosecutors often present all these characteristics. A typical example is the decision on whether to charge a suspect or what crime to charge. In these cases, implicit stereotypes may influence prosecutors' decisions to the disadvantage of racial minorities. For instance, empirical evidence indicates that many individuals tend to have a stronger implicit association between Blacks and aggressive behaviors than between Whites and aggressive behaviors. Similarly, at the implicit level, Blacks are often more easily associated with guns than Whites. Smith and Levinson (2011) argue that these stereotypes may influence the decision of a prosecutor regarding whether to charge a crime both when the Black person is the alleged victim and when he is the alleged perpetrator. For example, following a shooting, prosecutors may have to decide whether to justify the act for self-defense. When the alleged perpetrator is Black and the victim is White, the bias might make it more plausible in the mind of the prosecutor that the harm caused by the Black was the product of an unjustified aggressive behavior. Instead, when the victim is a Black person the stereotype may lead the prosecutor to believe that his aggressive behavior may have justified the actions of the White person under investigation (Smith and Levinson 2011).

Another strand of literature focuses on defense attorneys. On this regard, the concern is twofold.

On the one hand, implicit biases may induce a defense attorney to not defend a Black client. This, for instance, may occur when the attorney undervalues the probability of winning a case at trial on the basis of a biased evaluation of the available evidence. On the other hand, implicit biases may impair the performance of an attorney by undermining trust and communication with the client. Contextual factors such as work overload and degree of discretion may enhance the effect of implicit biases on attorneys' decisions. Incidentally, US public defenders often operate with insufficient resources to properly fulfill their tasks (Richardson and Goff 2013). Eisenberg and Johnson provide evidence of the pervasiveness of implicit biases among US defense attorneys (Eisenberg and Johnson 2004).

Most importantly, implicit biases may affect adjudicators' judgment and decision-making. And indeed, empirical evidence indicates that the judgment of both jury-eligible citizens and judges is affected by implicit biases (Hunt 2015; Kang et al. 2012). Regarding judges, Rachlinski and coauthors find that racial IAT scores among 120 US judges show similar patterns to those gathered with other samples. In particular, in line with the results obtained on the Project Implicit platform, they find that the large majority of White subjects held implicit attitudes that favor Whites over Blacks, while Black judges did not show strong preferences toward a particular racial group. The study further shows that when judges are not sufficiently motivated to control them, implicit biases affect their judicial decisions (Rachlinski et al. 2009).

Implicit biases can influence adjudicators' decisions from a number of perspectives, and in particular when judges have a relatively higher degree of discretion. Therefore, the influence of implicit biases might be more pronounced when the evidence presented by the parties is ambiguous (Kang et al. 2012). Hence, it is reasonable to conjecture that the effect of implicit biases might be more severe in civil cases, due to the lower burden of proof. In addition, implicit racial biases can affect memory recalls by leading adjudicators to remember facts in a more stereotypically consistent manner (Levinson 2007). In this regard, a

growing strand of literature is highlighting the existence of various implicit stereotypes that are of particular relevance in a criminal trial context. For instance, various studies suggest the existence of an implicit and bidirectional association between Blacks and crime (Hunt 2015). Similarly, Levinson and coauthors show the existence of a stronger implicit association between Blacks and guilt than Whites and guilt (Levinson et al. 2010). Importantly, IAT scores gathered with this measurement procedure predict evaluations of items evidence in a hypothetical trial. Along similar lines, in the context of death penalty judgments, implicit associations between racial groups and value of life have been shown to predict higher rates of death penalty punishments against Black defendants (Levinson et al. 2014). On a related note, implicit biases may partially account for the relatively higher rates of death penalty sentences inflicted to persons of color. For instance, implicit biases may ease finding the behavior of an offender as being heinous, atrocious, or cruel, which is one of the aggravating conditions under which death penalty can be inflicted under US law (Smith and Cohen 2012).

Outside the criminal law sphere, legal scholars have discussed implicit biases in the courtroom mainly with regard to employment discrimination lawsuits (Kang et al. 2012; Gertner and Hart 2012). In this context, the literature mostly focuses on the criteria that US judges are expected to adopt when deciding the dismissal of a case in its early stage. Here, the main concern regards the *Iqbal* standard, under which a judge should dismiss a case when the intent to discriminate the employee is not the most plausible explanation for the conduct of the employer.

In making this evaluation, the standard encourages judges to use their common sense. In this regard, legal scholars argue that, by asking judges to rely on common sense to make this decision, the law facilitates the influence of implicit biases on trial outcomes (Kang et al. 2012; Gertner and Hart 2012).

Debiasing and Insulating

Given the relevance of implicit attitudes and stereotypes, researchers are exploring how to

decrease or eliminate their influence on trial actors' decision-making. From a general perspective, these interventions can act either on trial actors' general tendencies to be influenced by these biases or on the environment in which trial decisions are made.

One way to decrease the influence of implicit biases on the general decision-making of trial participants is via training (Rachlinski et al. 2009; Kang et al. 2012; Smith and Levinson 2011; Teichman and Zamir 2014). This training may aim at increasing awareness among trial actors about the existence of implicit biases. In turn, this may have the positive effect of decreasing actors' overconfidence in the objectivity of their own judgment while increasing their motivation to control the biases. Various studies in psychology confirm that motivation to control implicit biases can be effective in reducing their influence on behavior, and hence jurors should be advised to take the perspective of the out-group person (Kang et al. 2012). Rachlinski and coauthors further suggest that judges should take a racial IAT (Rachlinski et al. 2009). Beside increasing awareness, this may have the positive result of providing judges with a rough estimation of their biases and may therefore help them taking decisions regarding the training they may need as well as avoid overcorrections (i.e., shifting the bias against Whites).

Another strategy to reduce the impact of implicit biases is exposing trial actors' to counter-stereotypical information or increasing their contacts with members of racial minorities (Rachlinski et al. 2009; Kang et al. 2012; Smith and Levinson 2011). An obvious path is increasing racial diversification in the courtroom and in prosecutors' offices. However, this can only be effective as a long-term strategy. In the short term, debiasing can be attempted by introducing images of positive figures of minority members in judges' offices.

Alternatively, interventions could focus on the context in which decisions are made. For instance, trials could be structured so that trial agents are given sufficient time to make their decisions. And indeed, this could reduce their reliance on automatic processes and thus the influence of implicit

biases on their decision-making. Similarly, research indicates that conditions of high cognitive load (i.e., situations in which the cognitive activity imposed on working memory is high) ease the influence of implicit biases on judgment. Therefore, various authors suggest trial actors to adopt strategies that may reduce cognitive load when making decisions (Kang et al. 2012). Additional strategies include hiding racial information in the file that prosecutors use to decide whether to charge for a crime (Smith and Levinson 2011) and increasing jury diversity (Kang et al. 2012). In fact, research indicates that racially diverse juries tend to endure more thorough deliberations than all-White juries. In turn, this may reduce the reliance on automatic processes in decision-making.

Accountability can also play a role in reducing the effect of biases. For instance, it has been proposed that prosecution offices could gather data regarding the racial composition of individuals at each stage of the charging phase to provide useful feedback to prosecutors (Smith and Levinson 2011). Similarly, judges could keep track of their decisions, in order to spot potential systematic biases in their decision-making (Kang et al. 2012; Rachlinski et al. 2009). Another path to increase judges' accountability is increasing the depth of the scrutiny allotted to appellate courts in situations where the racial composition of the court of first instance fuels the suspect that the decision might be biased (Rachlinski et al. 2009).

Potential interventions to reduce the effect of implicit biases at trial are however not limited to public institutions, as also attorneys can implement strategies to help them overcome their biases. Public defenders could refer to objective standards for triage and checklists to evaluate their cases. These measures may help attorneys to impose clearer limits to their own discretion (Richardson and Goff 2013).

Future Directions

As discussed in this entry, studies in implicit social cognition can provide many insights on trial dynamics and remedies to contrast racial discrimination in the courtroom. Yet, this field of

research is one of the latest developments of the already relatively young field of behavioral law and economics. Therefore, there is still a lot to learn on how these biases operate and how they influence the overall efficiency of the legal system. In the following, we highlight various possible pathways for future research.

First, it is still unclear how implicit biases affect the efficiency of legal systems. In particular, this strand of research can analyze how, to what extent and under which circumstances, implicit biases affect system costs related to criminal investigations, litigation, and adjudication. In addition, by influencing trial actors' behavior, implicit biases also affect the primary incentives provided by the law outside the courtroom (on a similar note see McAdams and Ulen (2009)). Last, as courts' accuracy is widely acknowledged as an essential element for the achievement of deterrence (Kaplow 1994), it is particularly relevant to study how implicit stereotypes affect accuracy in adjudication.

Second, the literature on implicit biases at trial mainly concentrates on Blacks (and Whites). However, the USA and other countries are becoming prismatic societies, and therefore more attention should be given to other racial and ethnic groups (e.g., Hispanics and Asians). Moreover, European scholarship and policymakers could benefit from research conducted on implicit discrimination against Arabs, Eastern Europeans, and Roma people. Along these lines, it might be warranted to conduct some studies on implicit stereotypes against Black people in Europe. In fact, as implicit biases are partially a product of culture, it might be interesting to study cross-cultural variations in implicit stereotypes.

Another pathway of research that remains largely unexplored regards implicit biases in non-criminal settings (Teichman and Zamir 2014). In particular, there is the need for further research in the fields of civil and administrative law. For instance, it has been argued that implicit biases may affect the application of the standard of proof in civil cases (Hunt 2015). Yet, these hypotheses await testing.

Future research can also contribute to a better understanding of debiasing mechanisms. As

shown above, the existing literature has started delving in this direction. Results reached so far are informative, but it is still not clear whether, and to what extent, these interventions are effective in real-life settings (Teichman and Zamir 2014). For instance, as discussed above, one of the main recommendations to decrease implicit biases among judges relates to increased diversity in judicial bodies. Yet, Rachlinski and coauthors found little differences in the pervasiveness of implicit biases between groups of judges coming from jurisdictions with great differences in racial diversity in the judiciary (Rachlinski et al. 2009). Further research could explore in more depth the conditions under which policies aimed at reducing implicit discrimination at trial are more likely to succeed.

Last, with few exceptions, most studies on implicit biases in the courtroom have been conducted with students. The field would benefit from having research conducted with professionals involved in trials. In particular, on the one hand, it would be interesting to gather data regarding the pervasiveness of implicit biases among trial agents. On the other hand, it would be important to analyze whether expertise affects the degree by which implicit biases impact behavior in trial settings.

Cross-References

- [Behavioral Law and Economics](#)
- [Bounded Rationality](#)
- [Cognitive Law and Economics](#)
- [Judicial Decision-Making](#)

References

- Cameron CD, Brown-Iannuzzi JL, Payne BK (2012) Sequential priming measures of implicit social cognition a meta-analysis of associations with behavior and explicit attitudes. *Personal Soc Psychol Rev* 16(4):330–350
- Eisenberg T, Johnson SL (2004) Implicit racial attitudes of death penalty lawyers. *DePaul Law Rev* 53(4):1539–1556
- Fazio RH, Olson MA (2003) Implicit measures in social cognition research: their meaning and use. *Annu Rev Psychol* 54:297–327
- Fazio RH, Jackson JR, Dunton BC, Williams CJ (1995) Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *J Pers Soc Psychol* 69(6):1013–1027
- Garoupa NM (2012) The economics of prosecutors. In: Harel A, Hylton K (eds) *Research handbook on the economics of criminal law*. Edward Elgar, Cheltenham, pp 231–242
- Gertner N, Hart M (2012) Employment law: implicit bias in employment discrimination litigation. In: Levinson JD, Smith RJ (eds) *Implicit racial biases across the law*. Cambridge University Press, New York, pp 80–94
- Greenwald AG, Krieger LH (2006) Implicit bias: scientific foundations. *Calif Law Rev* 94(4):945–967
- Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol* 74(6):1464–1480
- Greenwald AG, Poehlman TA, Uhlmann EL, Banaji MR (2009) Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *J Pers Soc Psychol* 97(1):17–41
- Greenwald AG, Banaji MR, Nosek BA (2015) Statistically small effects of the implicit association test can have societally large effects. *J Pers Soc Psychol* 108(4):553–561
- Hunt JS (2015) Race, ethnicity, and culture in jury decision making. *Ann Rev Law Soc Sci* 11:269–288
- Jolls C, Sunstein CR (2006) The law of implicit bias. *Calif Law Rev* 94(4):969–996
- Jolls C, Sunstein CR, Thaler R (1998) A behavioral approach to law and economics. *Stanf Law Rev* 50(5):1471–1550
- Kahneman D (2011) *Thinking, fast and slow*. Farrar, Straus and Giroux, New York
- Kang J, Bennett M, Carbado D, Casey P (2012) Implicit bias in the courtroom. *UCLA Law Rev* 59(5):1124–1187
- Kaplow L (1994) The value of accuracy in adjudication: an economic analysis. *J Leg Stud* 23(1):307–401
- Kubota JT, Banaji MR, Phelps EA (2012) The neuroscience of race. *Nat Neurosci* 15(7):940–948
- Levinson JD (2007) Forgotten racial equality: implicit bias, decisionmaking, and misremembering. *Duke Law J* 57(2):345–424
- Levinson JD, Cai H, Young D (2010) Guilty by implicit racial bias: the guilty/not guilty implicit association test. *Ohio State Univ J Crim Law* 8(1):187–208
- Levinson JD, Smith RJ, Young DM (2014) Devaluing death: an empirical study of implicit racial bias on jury-eligible citizens in six death penalty states. *NYU Law Rev* 89(2):513–581
- McAdams R, Ulen T (2009) Criminal behavioral law and economics. In: Garoupa N (ed) *Criminal law and economics*. Edward Elgar, Cheltenham, pp 403–436
- Nosek BA, Hansen JJ, Devos T, Lindner NM, Ranganath KA, Smith CT, Olson KR, Chugh D, Greenwald AG, Banaji MR (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur Rev Soc Psychol* 30(1):36–88

- Oswald FL, Mitchell G, Blanton H, Jaccard J, Tetlock PE (2013) Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *J Pers Soc Psychol* 105(2):171–192
- Rachlinski JJ, Johnson SL, Wistrich AJ, Guthrie C (2009) Does unconscious racial bias affect trial judges? *Notre Dame Law Rev* 84(3):1195–1246
- Richardson LS, Goff PA (2013) Implicit racial bias in public defender triage. *Yale Law J* 122(8):2626–2650
- Smith RJ, Cohen GB (2012) Capital punishment: choosing life or death (implicitly). In: Levinson JD, Smith RJ (eds) *Implicit racial biases across the law*. Cambridge University Press, New York, pp 229–243
- Smith RJ, Levinson JD (2011) Impact of implicit racial bias on the exercise of prosecutorial discretion. *Seattle Univ Law Rev* 35(3):795–826
- Teichman D, Zamir E (2014) Judicial Decisionmaking: a behavioral perspective. In: Zamir E, Teichman D (eds) *The Oxford handbook of behavioral economics and the law*. Oxford University Press, New York, pp 664–702

enforcement. Retrieved from understanding the WTO: the agreements: http://wto.org/english/thewto_e/whatis_e/tif_e/agrm7_e.htm).

This entry considers this balance by looking at the two poles of intellectual property policy: providing incentives to increase innovation and optimizing access to inventions both for consumptive use and for potentially innovation-increasing experimentation. This entry also surveys the notion of *calibration*, the idea that every country or region should adapt its regulatory framework to reflect its own strengths and weaknesses in optimizing what one might refer to as its innovation policy. A calibration approach suggests that providing innovation incentives and optimizing access are not mutually exclusive objectives.

TRIPS Agreement

Daniel Gervais
Vanderbilt Intellectual Property Program,
Vanderbilt University Law School, South
Nashville, TN, USA

Abstract

The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS Agreement) was negotiated between 1986 and 1994 during the Uruguay Round of the General Agreement on Tariffs and Trade (GATT), which led to the establishment of the World Trade Organization (WTO). The TRIPS Agreement sets minimum levels of several types of intellectual property (IP) protection, including copyright, trademarks, patents, industrial design, and trade secrets protection. Membership in the WTO includes an obligation to comply with the TRIPS Agreement. According to the WTO, the Agreement attempts to strike a balance between long-term social benefits to society of increased innovations and short-term costs to society from the lack of access to inventions (World Trade Organization (n.d.) Intellectual property: protection and

Part I: Incentivizing Innovation and Optimizing Access

The TRIPS Agreement has become the rope in the tug-of-war between utilitarianism and egalitarianism. Much of the Western intellectual property system seems to be based on a utilitarian theory regarding which IP rights should be given in order to incentivize the creation of new and potentially commercially valuable knowledge, referred to as “informational goods.” This expression is meant to capture goods and related services that derive all or most of their value from the information they contain or embody.

As IP rights and protection standards increase, problems regarding access to existing informational goods arise. A maximalist utilitarian argument might be that there should be no right to access new informational goods because these goods would not have existed in the first place *but for* the existence of adequate IP rights. A more nuanced utilitarian approach might allow for some unlicensed use of certain informational goods, notably to maximize the continuing production of knowledge. Of course, in an optimal policy scenario, allowing for unlicensed uses of new goods should be done in a way that does not negatively affect incentives to innovate (Chandler and Sunder 2007). Egalitarianism, by contrast,

focuses mostly on distributional issues, notably access to new knowledge and goods.

Utilitarian and egalitarian objectives are not incompatible. In fact, incorporating egalitarian or distributive concerns into the utilitarian conversation can lead to the design of a system where both production and access are considered equally valid objectives. Unfortunately, discussions regarding the TRIPS Agreement have tended to focus either on the failure to produce sufficient incentives to innovate (i.e., the failure to supply an optimal supply of new informational goods) or on the failure to provide access to potentially life-changing and life-saving inventions. Both utilitarian and egalitarian perspectives are legitimate. From the utilitarian perspective, if new inventions or creations would not have existed *but for* the IP incentive, then negatively affecting incentives to innovate has adverse welfare impacts, including on economic development. This valid utilitarian argument can coexist with an equally valid egalitarian argument: in situations where markets fail to produce the optimal supply of the demanded informational goods due to anticompetitive practices or transaction costs, access may be improved by the use of appropriate exceptions to and limitations on IP rights, including compulsory licensing. In this scenario, access to these goods may be optimized without negatively impacting incentives to innovate in a significant way.

How does one strike the right balance of creating the incentives to enable or accelerate the development of new technologies and not restricting those technologies unduly when there is no real market to protect (e.g., in the case of a public good)? An economic analysis of IP law and policy makes it possible to dispel the normative fog of utilitarianism or egalitarianism, in which many legislative and judicial decisions are steeped. An economic lens offers novel ways of envisioning public actions and allows for the possibility of moving beyond inconsistencies and sticking points (Josselin and Marciano 2001). Striking the right balance requires the determination of what affects the value of an IP right, particularly what reduces that value.

Reduction of value may follow from granting (or even the threat of granting) a compulsory

license or eliminating the right to exclude in a certain context by granting a limited legislative or court-made exception to allow for the free use of the protected creation or invention. While limiting exclusive rights may result in the reduction of the value of an IP right, it also allows third parties to understand, copy, and potentially improve on the subject of the IP right. The equilibrium between protecting the value of an IP right and optimizing the production of the protected informational good is difficult to achieve.

The Parameters of IP Policy Equilibrium

Normative debates based on the definition of property rights rarely provide a full answer to the policy question of balance outlined above. Intellectual property usually bears *aspects* of a property right, notably the right to exclude others, but that right is often much more limited than a full ownership right in a tangible object (Waldron 1988; Mackaay 1994). Consider the law of trademarks. It requires proof of a risk of confusion on the part of the potential purchaser before the right to exclude can be applied. No comparable requirement exists with respect to tangible objects. In the same vein, exceptions to the right to exclude such as fair use in copyright law, compulsory licensing of copyrights or patents, or experimental use of a patented invention without permission provide evidence of the varying nature of informational goods and of the relative nature of IP rights. Policy equations must reflect these differences.

The impossibility of determining the exact degree of overlap between property rights (in the classical sense of the term) and IP rights is not an outright impediment to striking an efficient policy balance using economic analysis. A list of objectives that is fairly persuasive can be drawn up: (1) the creation of sufficient incentives to innovate, (2) the efficient functioning of the markets for copyright works and patented inventions embedded in informational goods, and (3) the optimal (or at least reasonable) access to works and inventions. A fourth objective should be added, namely, the recognition (or attribution) of the work to the creator or inventor. An attribution right can be justified as useful for the author or inventor in generating goodwill but also as

providing users with information on the source of the good (e.g., name of the author, trademark of maker or supplier). A user can use this information to attribute trust value to the attributed source and thus possibly have decreased search costs in future transactions concerning the same good.

In order to achieve an equilibrium, IP policy equations must be solved in a way that meets the aims of all the interested parties – including creators, inventors, firms, investors, users, and reusers. The parameters of the solution allow for the perception of an adequate level of foreseeability for every group of stakeholders:

[I]ntellectual property rights, like other property rights, set the parameters allowing investors to make a guess of the expected revenues. Only once these parameters are set will investments be undertaken. (Mackaay 1994)

Economic analysis leads policy-makers to set objectives, the achievement of which can be measured by concrete results rather than fulfillment of ideology.

Compulsory Licenses

An option to maximize access to and use of the invention or creation is to issue compulsory licenses. These licenses typically involve the payment of remuneration (determined to be adequate by competent authorities) and reporting requirements (amount made, used, exported (if any), etc.). Issuing compulsory licenses constitutes the establishment of a liability regime (instead of an exclusive right) (Reichman 1992).

Compulsory licensing is allowed under the TRIPS Agreement (World Trade Organization 2014). The debate regarding compulsory licenses, which has been on top of the agenda in international IP law and policy for decades, is centered on access to inventions but also on negative impacts on incentives to innovate. In this debate, the unfairness of the patent regime for indigent users who cannot afford new medicines is often emphasized; poor countries may only be able to procure new medicines at the expense of other urgent developmental priorities (i.e., at the expense of the production of other public goods).

Compulsory licenses may improve private parties' access to informational goods and help

to achieve efficient outcomes when markets are working suboptimally in providing access, especially when access to the informational good concerned is considered necessary (e.g., pharmaceuticals or educational materials). Governing bodies issue compulsory licenses typically to allow the use of an informational good either to increase access or, second, in the hopes that third parties will make improvements or invent new products, which, in turn, will lead to competition in the market where the IP right owner is operating. The risk is the systemic weakening of incentives to innovate. If the price paid is considered too low, then financial incentives may be negatively impacted. This is also true if the instituted exception interferes with the markets operating in other territories. For example, one who purchases products at a lower cost due to a compulsory license should not export the products to another territory; such a practice may disrupt the market of the other territory. The holder of an exclusive right might also experience a reduction in the value of that IP right if a competitor were able to use the subject of the exclusive right without permission from the holder.

Compulsory licensing is subject to two additional caveats. First, knowledge that would be transferred with a *voluntary* licensing agreement may not accompany a *compulsory* licensing scenario. This knowledge may not be disclosed or otherwise “enabled” (meaning that a person with requisite knowledge can understand how to make and use the invention from the information disclosed in the patent) by the patent subject to the compulsory license. Second, a firm subjected to a compulsory license may withdraw, delay, or cancel other investments in the country. These are factors to bear in mind before issuing compulsory license. Their relative weight in the decision will vary case by case.

The Specific Case of Cultural Goods

There are economically relevant cultural aspects to the creation of incentives to innovate in the literary and artistic fields. Problems of incentives and access can impact the creation and viability of industries that produce and market cultural informational goods. Additionally, IP rules can negatively impact access and use of such goods.

Incentives to innovate are linked to the size of potential markets for new goods, which in gross economic terms would factor in the number of prospective users and their financial ability to purchase “cultural goods,” which are a subset of informational goods more directly linked to cultural and entertainment industries, such as films, music, and books in various formats. This explains the importance of English – and specifically of the North American market – in this context. Because many buyers in those regions can afford to purchase many cultural goods, they create a solid market for English-language informational goods and thus an incentive to produce such goods. The role of English as a second language in many other parts of the world reinforces this feature.

When goods are produced for markets where prices are comparatively high, they may not be made available at lower prices in other markets, in part because price discrimination may lead to exports from lower-priced markets to higher-priced ones. Such exports may be unauthorized by the copyright holder but still legal in countries that allow “international exhaustion,” namely, the ability to import in their territory goods legally produced in other countries. The TRIPS Agreement contains no mandatory rule on exhaustion.

In 1971, an appendix was added to the main international copyright instrument, the Berne Convention. This appendix allows compulsory licenses to be issued for the translation and reproduction of foreign books in developing countries in order to increase access to knowledge. Developing countries, particularly their education systems, may not have the economic ability to purchase cultural goods in sufficient quantity to validate the incentive to innovate protected by the IP system. Because some developing countries do not have sufficient buying power (compared to more industrialized markets), they would not contribute to the monetary award that the IP right entitles the creator of the work to receive. The ability to issue compulsory licenses for translation has in fact been one of the most contentious issues since the early days of the Berne Convention.

Part II: Calibration Within TRIPS Parameters

Changes to the TRIPS Agreement and the adoption and implementation of new wide-ranging multilateral treaties are unlikely and perhaps decades away. In fact, it took over a decade to pass Article 31*bis*, the only amendment to the TRIPS Agreement thus far. Therefore, at this juncture and for the predictable future, the calibration of IP policy will remain mostly a matter of domestic policy-making and regional trade agreements. Calibration will most likely involve using flexible provisions contained in TRIPS (i.e., the specific exceptions and limitations) or interpreting the Agreement to achieve IP policy goals. For example, courts may interpret the Agreement to include fair use/fair dealing exceptions to copyright rights and setting of appropriate limits on patentable subject matter.

Structurally, calibration is not a rejection of the harmonization mentioned in the opening paragraph. The calibration approach recognizes that rules will vary to a certain degree due to differences among regions, countries, and industries. This approach suggests that variations within parameters set by TRIPS are *desirable*. Calibration suggests that, by developing a comprehensive IP strategy focused on innovation and welfare improvements, a country can limit the negative impact of transitioning to more rigorous IP protection and increase its chances of reaping the benefits thereof, including technology-related foreign direct investment (FDI) and growing domestic Web-based, pharmaceutical, and other technology-intensive industries.

The Calibration Process

How does one design a calibrated implementation of TRIPS as part of an innovation optimization strategy? One can begin by eliminating what is unlikely to work. For example, there are probably no definitive answers given to questions such as whether the optimal term for a patent is 12, 18, or 22 years. All one can say is that, beyond a certain point, patent protection *decreases* innovation (Gallini 1992). Empirical studies have verified that in countries without the necessary technology-

absorptive capacity, increasing patent protection for pharmaceuticals produces little, if any, innovation outcomes (La Forgia et al. 2009). Similarly, in the copyright realm, there are no definitive answers to how long IP protection should last. Is the optimal copyright term 14 years from publication or the life of the author plus zero, 50, or 70 years? (Sprigman 2004). What combination of rights and exceptions would better achieve the policy goals than rules in place now? Is using a *single* term of protection part of the problem? One could argue that for a certain invention (protected by patent) or creation (protected by copyright), one specific term is optimal while a different term is more appropriate for another invention or creation (Stanley 2003). Finding the optimal point of IP protection is difficult because each informational good might require a different level of protection. Thus, horizontal rules (such as a single term of protection or bundle of rights) are approximations at best. These policy discussions can take the form of continuous imprecise balancing acts, making it difficult to reach a calibration target (Falvey et al. 2004).

Instead of engaging in *systemic* analyses to issue bright-line rules (e.g., patent protection lasts 20 years), policy-makers could try the other extreme, that is, a full case-by-case approach. Such an approach would factor in the exact value added of *each* informational good. Value determination would then depend on measuring the exact size of the inventive step involved in the invention concerned. This is related to the issue of measuring *patent quality*, and various metrics have been proposed, including by the Organisation for Economic Co-operation and Development (“OECD”), by considering how many times a patent is cited in later patents (Organisation for Economic Co-operation and Development 2011). Such measurements often emphasize the need to move policy goals away from purely quantitative patent metrics (number of applications or grants) and focus on the quality of the information disclosed.

One could add to this equation the degree of competition in the industrial or economic sector impacted by the invention and, relatedly, the number of dominant players by market share (Hollis 2004). Even if such a case-by-case approach did

not encompass high transaction costs, experts could only estimate the future utility of the invention. In terms of predictability, time, and transition/protection costs, a system of copyright or patent with a single set of rights and one term of protection may therefore be better than case-by-case analysis. While perhaps theoretically less attractive, such a system is simple to understand and easy to administer.

Does this mean that one is stuck at the abstract level without any hope of achieving relative analytical clarity? Here, TRIPS only offers normative guideposts and guarantees minimum levels of protection. In implementing TRIPS, each country or region should calibrate its IP policy recognizing its own unique characteristics and needs.

Calibration by Region and Industry

Both the systemic and case-by-case approaches are suboptimal for the aforementioned reasons. A more realistic approach for policy-makers trying to effectuate calibration revolves around the level of rights and exceptions that, *within the range of TRIPS-compatible implementations*, is most likely to achieve the policy goal of maximizing innovation while minimizing negative welfare impacts. Because the TRIPS Agreement only “harmonized” national laws to a degree, it left space for calibration (Reichman 2000). One should use a combination of human *and* economic development factors in order to craft a set of objectives to help determine the best level of rights and exceptions. These objectives will be amenable at least to the kind of metrics (e.g., the United Nations Human Development Index) that evidence-based policy-making calls for. Development is “a pseudonym for a complex network of benefits associated with economic growth and human social capital” (Okediji 2009). It is the sum of the changes in social patterns and norms through which production devices couple with the population: the latter acquires the capacity to utilize the former in order to achieve what is considered to be a satisfactory growth rate, and the production devices supply products that serve the population instead of being “alien” to the population. This dialectic of production devices

and population is the essence of development (Perroux 1988).

Establishing broad developmental goals is not the same as implementing them. Each country implementing TRIPS must recognize how it compares to others in the region or countries elsewhere at a similar level of development.

Differences Among Countries

Developing countries are not all identical; in fact, they are far from identical. Developing countries can be grouped in various ways. Leaving aside the least developed countries (LDCs), for which TRIPS obligations have been suspended, one could distinguish developing countries via a “net outcomes” approach. Countries in which innovation benefits outweigh additional rent extraction can be grouped together, and countries in which additional rent extraction outweigh innovation benefits can be grouped together. Professor Llew Gibbons offers a promising taxonomy of developing countries (Gibbons 2011). He grouped these countries according to three stages of development:

Stage One

- Foreign direct investment is rare and usually limited to specialized sectors – often relating to the exploitation of natural resources or developing franchise service industries like a major international brand bottling company.
- There exists unskilled, cheap labor.
- Foreign businesses create the necessary infrastructure and invest in human capital.
- The developing country is investing in the training of skilled workers and junior managers. Successfully developing a skilled workforce is a prerequisite to entering stage two.

Stage Two

- The economy is now able to absorb technology, to imitate technology at some level, and to contribute minor improvements.
- There exists a well-educated workforce adapted to absorb new technology from other countries and then incorporate the technology into the domestic economy.

- Domestic research efforts are primarily facilitative or associated with technology transfer. Focus shifts gradually to efforts on more innovative projects.

Stage Three

- The newly industrialized country produces its own intellectual property.
- The country is very selective as to which IP rights it zealously protects.

This progression from imitation to absorption to innovation tracks the path proposed by innovation theorists, i.e., progression from imitation to adaptation to true global innovation. Countries have indeed followed the progression described; along the pathway of development, the countries have gradually created more IP rights and made better use of existing IP rights. As countries build innovation-focused industries, they generally develop more sophisticated and nuanced IP policies (Gervais 2005). Simply put, they play the IP game better.

Differences Among Industries

The TRIPS Agreement put all industries on the same footing. Yet, treating all industries as equally sensitive to IP protection is incorrect and results in a number of unintended consequences. First, there is a relatively short list of industries generally considered to be highly patent sensitive. The list includes industries consisting of certain chemical companies, producers of laboratory instruments, and makers of steel-mill products (Cheng 2012; Maskus 2000; Gervais 2005). This list is partially derived from Mansfield’s studies of the field, in which he found that while 90% of pharmaceutical innovations were dependent on patent protection, only 20% of electronics and machinery innovations were patent dependent (Cheng 2012; Mansfield 1986). Beyond the general consensus on the patent dependency in these aforementioned fields, controversy concerning the role and impact of patents on innovation quickly emerges. Do the benefits from patents to computer software and online commerce outweigh the costs of obtaining, enforcing, and defending against patent infringement claims, especially those brought by

nonpracticing entities? (Mullin 2013). This question is front and center in a number of both developed and developing countries (Shrestha 2010). Proposals to recognize differences in industries have been made in the United States. For example, some scholars suggest that if the legislator is unwilling to separate industries by applying different standards, then perhaps courts should (Burk and Lemley 2010). In a developing country, implementing a *single* patent policy for *all* industries – as is facially required by TRIPS – may thus be structurally suboptimal.

While the TRIPS Agreement identifies uniform patentability criteria, the question whether it makes sense to treat all industries the same should be on any comprehensive innovation agenda. A form of “discrimination” based on the nature of the industry concerned would be justified if one could develop a proper metric to measure whether and how innovation outcomes are achieved.

Conclusion

As policy-makers around the world navigate deeper in the calibration waters of TRIPS implementation and seek to optimize national innovation strategy within the boundaries ascribed by TRIPS, they must attempt to avoid the negative consequences of increasing IP protection, from patent trolls, and the prevention of access to essential medicines by patents to the chilling of free expression and prevention of fair uses by copyrights and trademarks. While avoiding the Scylla of excessive IP, one must also not run to the Charybdis of insufficient IP. This is the inevitable and desirable process of calibration that an economic analysis of IP law can both inform and guide.

Cross-References

- Copyright
- Intellectual Property: Economic Justification
- Trade Secrets Law

- Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond
- WTO: Procedural Rules

References

- Burk D, Lemley M (2010) The patent crisis and how the courts can solve it. *Syracuse Sci Tech Law Rep* 23:1–22
- Chandler A, Sunder M (2007) Is Nozick kicking Rawls ass?: intellectual property and social justice. *UC Davis Law Rev* 40:563
- Cheng T (2012) A developmental approach to the patent-antitrust interface. *Northwest J Int Law Bus* 33:1–79
- Falvey R, Foster N, Greenaway D (2004) Intellectual property rights and economic growth. Internationalisation of economic policy research paper no 2004/12, 2. Retrieved from <http://ssrn.com/abstract=715982>
- Gallini N (1992) Patent policy and costly imitation. *Rand J Econ* 23(1):52–63
- Gervais D (2005) Intellectual property and development: the state of play. *Fordham Law Rev* 74(74):505–535
- Gibbons L (2011) Do as I say (not as I did): putative intellectual property lessons for emerging economies from the not so long past of the developed nations. *SMU Law Rev* 64(3):923–973
- Hollis A (2004) An efficient reward system for pharmaceutical innovation. Retrieved from <http://www.who.int/intellectualproperty/news/Submission-Hollis6-Oct.pdf>
- Josselin J-M, Marciano A (2001) L’analyse économique du droit et le renouvellement de l’économie politique des choix publics. *Economie Publique/Pub Econ* 7:6
- Mackaay E (1994) Legal hybrids: beyond property and monopoly? *Columbia Law Rev* 94:2630
- Mansfield E (1986) Patents and innovation: an empirical study. *Manag Sci* 32(2):173–181
- Maskus K (2000) Lessons from studying the international economics of intellectual property rights. *Vanderbilt Law Rev* 53(6):2219–2239
- Mullin J (2013) In historic vote, New Zealand bans software patents: patent claims can’t cover computer programs ‘as such’. Retrieved from *Ars Technica*: <http://arstechnica.com/tech-policy/2013/08/in-historic-vote-new-zealand-bans-software-patents>
- Okediji R (2009) History lessons for the WIPO development agenda. In: Netanel W (ed) *Intellectual property and developing countries*. Oxford University Press, Oxford, pp 137–162
- Organisation for Economic Co-operation and Development (2011) *Competing in the global economy, technology: quality in OECD science, technology and industry scoreboard*. Retrieved from oecd.org
- Perroux F (1988) The pole of development’s new place in a general theory of economic activity. In: Higgins B, Savoie D (eds) *Regional economic development: essays in honour of Francois Perroux*. Unwin Hyman, Boston, pp 48–76

- Reichman J (1992) Legal hybrids between the patent and copyright paradigms. In: Altes WK (ed) *Information law towards the 21st century*. Kluwer Law, Deventer, pp 325–341
- Reichman J (2000) The trips agreement comes of age: conflicts or cooperation with the developing countries. *Case West Reserve J Int Law* 32:441
- Shrestha S (2010) Trolls or market-maker? An empirical analysis of nonpracticing entities. *Columbia Law Rev* 110:114–160
- Sprigman C (2004) Reform(aliz)ing copyright. *Stanf Law Rev* 57:552
- Stanley C (2003) A dangerous step toward the over protection of intellectual property: rethinking *Edler v. Ashcroft*. *Hamline Law Rev* 679:694–695
- Waldron J (1988) *The right to private property*. Clarendon, Oxford
- World Trade Organization (2014) Members accepting amendment of the TRIPS Agreement. Retrieved from intellectual property: trips and public health: http://www.wto.org/english/tratop_e/trips_e/amendment_e.htm
- World Trade Organization (n.d.) Intellectual property: protection and enforcement. Retrieved from understanding the WTO: the agreements: http://wto.org/english/thewto_e/whatis_e/tif_e/agrm7_e.htm

Further Reading

- La Forgia F, Osenigo L, Montobbio F (2009) IPRs and technological development in pharmaceuticals: who is patenting what in Brazil after TRIPS. In: Netanel W (ed) *The development agenda: global intellectual property and developing countries*. Oxford University Press, Oxford, pp 293–319
- Nachbar T (2004) Intellectual property and constitutional norms. *Columbia Law Rev* 104:338–339
- Prud'homme D (2012) Dulling the cutting-edge: how patent-related policies and practices hamper innovation in China. Beijing, European Union Chamber of Commerce in China
- Ragavan S (2012) *Patent and trade disparities in developing countries*. Oxford University Press, Oxford

Type-I and Type-II Errors

Matteo Rizzolli

LUMSA University, Rome, Italy

Abstract

Adjudicative procedures meant at establishing truth about facts on defendants' behavior are naturally prone to errors: defendants can be found guilty/liable when they truly were not (type-I errors) or they can be acquitted when

they should have been convicted (type-II errors). These errors alter the incentives of defendants to comply with norms. We review the literature with a particular focus on type-I errors.

Introduction

The word *adjudication* has its Latin roots into the words *jus* (right, justice) and *dicere* (to say). All organizations set goals and have adjudicative procedures to “establish the truth” about members' compliance. Performance appraisal committees decide whether employees earn rewards within business organizations; teachers must assess students' progress in learning; courts adjudicate whether citizens have committed crimes; disciplinary committees within sport leagues and professional organizations as well as religious tribunals assess whether members' conduct has been conforming. Adjudicative procedures must evaluate and reward something they cannot directly observe – being it effort, intention, or act – and this makes them obviously prone to errors. Although our framing will be mostly applied to criminal (See also ► “[Criminal Sanctions and Deterrence](#)” and ► “[Crime, Incentive to](#)”), administrative, and civil courts, most of the results here presented apply to any generic adjudicative procedure. We thus consider the general case of an adjudicative authority who (i) must assess whether the observed behavior of an individual *conforms* or *deviates* from the prescribed behavior and (ii) must incentivize or sanction such behavior accordingly. In judging behavior, errors inevitably arise, and they generally undermine individuals' incentives. These errors take mainly two forms: (i) the adjudicative authority may assess non-compliance when in fact the subject is duly complying and (ii) the adjudicative authority may assess compliance when in fact the subject is deviating. Individual's compliance with the prescribed behavior can be interpreted as the null hypothesis, so that the adjudicative authority can both incorrectly reject the null and sanction a complying subject (a type-I error) and incorrectly accept the null and exculpate an undeserving subject (type-II error). In the context of crime

deterrence, type-I errors amount to wrongful convictions of innocents. We model the relation between type-I and type-II errors below within a standard optimal deterrence framework. Finally, we discuss the empirical relevance of type-I errors.

Basic Setup

Let y_0 be the initial endowment equal for all agents and b the benefits from deviating from the prescribed behavior (e.g., committing crime). b is distributed among the agents with a generic distribution $z(b)$ and a cumulative $Z(b)$ with support $[0, \bar{b}]$. Let also h be the harm/externality generated by each individual's deviation (each individual takes this decision only once). For the sake of simplicity, all individuals are audited and brought in front of an adjudicative authority. The authority observes the amount of inculpatory evidence e that is produced against a defendant, and if this overcomes a certain threshold, $\tilde{e}$ then the authority imposes a monetary sanction s . For the sake of simplicity, we also assume that there is no welfare-improving deviation as in Becker seminal paper on crime (See also ▶ “Crime and Punishment (Becker 1968)” by Becker 1968 and also to “Becker, Gary S.”) (this would be a crime for which $b > h$) and that monetary sanctions are transferred from the defendant to society. Furthermore, in the function of social costs, we do not consider the private benefits from crime but only its social costs.

Therefore let e have a frequency distribution of $i(e)$ for the conforming defendant (innocent) and of $g(e)$ for the deviating defendant (guilty). Let $I(e)$ and $G(e)$ be the cumulative distributions of $i(e)$ and $g(e)$, respectively, and note that $I(\tilde{e})$ and $G(\tilde{e})$ are the probabilities of being acquitted for the complying and for the deviating defendant, respectively, given the evidence threshold $\tilde{e}$. To keep notation compact, we will often use I and G for $I(\tilde{e})$ and $G(\tilde{e})$, respectively.

The evidence is stochastically distributed, albeit in general more incriminating evidence is available against deviating defendants than against complying ones. First-order stochastic dominance is assumed $I(e) > G(e) \forall e \in]0, e_{\max}[$. Without

FOSD evidence would be produced randomly for the complying and the deviating alike, and therefore the whole criminal procedure would be pointless. Note also that G is the probability of type-II error and $1 - I$ is the probability of type-I error. Let us also define $\Delta(\tilde{e}) = I - G$ as the *accuracy* of the adjudicative procedure; Δ represents the ability of the procedure to distinguish complying from deviating defendants.

For our purpose, we assume the social planner optimizes deterrence only by affecting the threshold $\tilde{e}$ which in turn determines the error's trade-offs: for instance, an increase in $\tilde{e}$ generates both an increase in the number of wrongful acquittals G and a decrease in the number of wrongful convictions $1 - I$.

The risk-neutral individual does not deviate as long as the returns from deviating behavior are smaller than the expected returns of conforming. Since b varies across individuals, there exists a level of $\tilde{b}$ for which the individual is indifferent between conforming and not, and this determines the proportion of the population $Z(\tilde{b})$ who conforms.

Social welfare is thus $(1 - Z(\tilde{b}))h$: the social costs of harm caused by those defendants who deviate. On the other hand, the social planner only acts on the threshold $\tilde{e}$ which implicitly defines the trade-off between type-I and type-II errors. The link between the social planner's choice of the evidentiary standard $\tilde{e}$ which in turn determines the error's trade-off and the defendant's choice of conformity determined in $\tilde{b}$ are the ingredients to understand the role of adjudication in deterrence.

Let us begin by assuming agents to be risk-neutral utility maximizers. The returns from conforming are $E\pi_I = y_0 - (1 - I)s$, while the returns from deviating are $E\pi_G = y_0 + b - (1 - G)s$. All defendants for which $E\pi_I \geq E\pi_G$ will conform and therefore the threshold level of b which implicitly defines the conforming population is

$$\tilde{b}_m = (1 - (1 - I) - G)s \quad (1)$$

By looking at Eq. 1, we can single out the typical “deterrence effect” as $\tilde{b}$ increases both with the magnitude of the sanction ($\uparrow s \Rightarrow \uparrow \tilde{b}$)

and via an increase in the detection probability for the deviating defendants which in this model corresponds to a decrease in the probability of type-II errors ($\downarrow G \Rightarrow \uparrow \tilde{b}$). Furthermore, a “compliance effect” of type-I errors can be seen: $s \tilde{b}$ increases when the probability of being punished decreases for conforming defendants ($\uparrow I \Rightarrow \uparrow \tilde{b}$). Also a “screening effect” can be established: the higher is the accuracy Δ , the better the procedure can discriminate between conforming and non-conforming behaviors and the greater the advantages of staying conforming ($\uparrow \Delta \Rightarrow \uparrow \tilde{b}$). Finally, by simple inspection of Eq. 1, it is evident that marginal change in either $1 - I$ or G determines an equal decrease of $\tilde{b}$ as $\frac{\partial \tilde{b}}{\partial (1-I)} = \frac{\partial \tilde{b}}{\partial G} = s$. Under risk neutrality, type-I errors ($1 - I$) and type-II errors (G) have the same negative impact on the defendant’s incentive to comply. This is because on one hand type-II errors undermine compliance inasmuch as they decrease the probability of nonconforming defendants being finally sanctioned. On the other hand, type-I errors increase the opportunity costs of conforming relative to deviating.

Now that the threshold level $\tilde{b}$ is defined, the social welfare can be computed and derived with respect to the evidence threshold:

$$\begin{aligned} \frac{\partial SW}{\partial e} &= \partial(1 - Z(\tilde{b}_m))h \\ &= -z(\tilde{b}_m)(i(\tilde{e}) - g(\tilde{e}))sh \end{aligned} \quad (2)$$

Let $\tilde{e}_{\text{neutral}}$ be implicitly defined by $i(\tilde{e}) = g(\tilde{e})$. By inspection of Eq. 2, the optimal evidence threshold $\tilde{e}$ that minimizes social costs is $\tilde{e}_{\text{neutral}}$. In fact accuracy reaches its maximum level when the social planner chooses $\tilde{e}_{\text{neutral}}$ so that the distance between the two cumulative functions is maximized. If the social planner chooses a higher evidence threshold $\tilde{e}_{\text{pro-defendant}} > \tilde{e}_{\text{neutral}}$, then the error trade-off tilts in favor of the defendant as the probabilities of both correct and wrongful acquittals $-I$ and G , respectively, increase. $\tilde{e} > \tilde{e}_{\text{neutral}}$ necessarily also implies $g(\tilde{e}) > i(\tilde{e})$ by definition of the frequency distribution of $i(e)$ and $g(e)$. Notice that for levels of $\tilde{e} > \tilde{e}_{\text{neutral}}$, G grows faster than I and therefore accuracy cannot be maximal.

Evidence Thresholds, Standard of Evidence, and Error Ratios

While our analysis focuses on the evidentiary threshold $\tilde{e}$ that determines the probabilities of both type-I and type-II errors, there are other two common concepts that concern adjudication and that must be put in relation with our analysis.

The first one is the **standard of evidence**: it is generally understood as the level of certainty the adjudicative authority must reach in order to establish guilt in a criminal proceeding (or liability in civil one). Among the most common standards of proof used in different adjudicative procedures, there are the *preponderance of evidence* (*poe*) standard, the *clear and convincing evidence* (*cace*) standard, and the *beyond any reasonable doubt* (*bard*) standard. Although giving probabilistic interpretations of these standards of proof is very controversial (see Kaplow 2012, footnote 76 for a discussion), they are commonly understood to roughly coincide with the 50%, 75%, and 95% thresholds, respectively. Under *poe* (or *cace* or *bard*), the adjudicative authority must answer to the question of whether, given the evidence available, the likelihood that the defendant has deviated is larger than 50% (or 75% or 95% depending on the standard applied). These probabilities must be understood as Bayesian posterior probabilities of having deviated, and these are functions – following the Bayes’ rule – of the likelihood of the signal given by the densities i and g of the evidentiary threshold $\tilde{e}$ and on the prior probability of being brought in front of the adjudicative authority. The probability that a defendant has deviated or not also depends on the base rates of the two actions, $(1 - Z(\tilde{b}))$ and $Z(\tilde{b})$, respectively. The $\tilde{b}$ are determined endogenously by defendants’ decisions and ultimately depend on the evidence threshold $\tilde{e}$. Therefore in order to identify the proper threshold $\tilde{e}$ – in case the *poe* standard applies – one should ask what value of $\tilde{e}$ implicitly solves the equation $g(\tilde{e}) \cdot (1 - Z(\tilde{b})) = i(\tilde{e}) \cdot Z(\tilde{b})$. If the adjudicative authority needs to apply the *cace* or *bard* standard, one could simply multiply by either 3 or 19 the right side of the previous equation. As Lando

(2002) and Kaplow (2012) point out, the two notions – the one based on the optimal *evidence threshold* and the one based on the *standard of evidence* – are strikingly different. To begin with, the optimal evidence threshold is derived from welfare analysis and seeks to find the level of $\tilde{e}$ that maximizes social welfare. By contrast, within the *standard of evidence* framework, $\tilde{e}$ is derived by asking under what circumstances would the probability that the defendant before the adjudicative authority has actually deviated be 50% (or 75% or 95% or other conventional probabilities). In fact the optimal *evidence threshold* could be associated with any probabilistic *standard of evidence* whatsoever.

Another approach focuses on the **ratio of errors** and expresses the pro-defendant bias of adjudicative procedures in terms of error ratios. There seems to be something specific about type-I errors in the context of crime: scholars and rule makers across time and societies advocated a specific attention to the avoidance of type-I errors even at the expense of many type-II errors; arguably the most famous statement in this respect is the one of William Blackstone (1769) recommending that *it is better that ten guilty persons escape than that one innocent suffer*. Dekay (1996) systematizes the relation between the standards of evidence and the error ratios. We can interpret these as ratios of errors' frequencies where the frequency of erroneous acquittals is the conditional probability that a truly deviating defendant is acquitted (type-II error) multiplied by the base rate of the action $(1 - Z(\tilde{b}))$, while the frequency of erroneous conviction is the conditional probability that a truly complying defendant is convicted (type-I) multiplied by the base rate $Z(\tilde{b})$. So the type-I error ratio (sometimes also called the Blackstone's error ratio) is defined as $\frac{G \cdot (1 - Z(\tilde{b}))}{(1 - I) \cdot Z(\tilde{b})}$.

All else being equal, higher standards of evidence that affect the trade-off between G and I do imply higher Blackstone-like error ratios. However, it should be noticed that the optimal *evidence threshold* could be associated with many different error ratios depending on the base rates.

Risk and Loss Aversion

Subjects are known to be generally risk-averse in their utility of income. We thus introduce risk aversion in the measure of the monetary gains from crime b following Rizzolli and Stanca (2012). If b are monetary gains for which utility $U(\cdot)$ can be derived, then the expected utility of complying is $IU(y_0) + (1 - I) \cdot U(y_0 - s)$, while the expected utility of deviating is $GU(y_0 + b) + (1 - G) \cdot U(y_0 + b - s)$. The threshold level of $\tilde{b}_{eu}$ that triggers a defendant to deviate is implicitly defined by

$$\begin{aligned} & I[U(y_0) - U(y_0 - s)] \\ & - G[U(y_0 + b) - U(y_0 + b - s)] \quad (3) \\ & \geq U(y_0 + b - s) - U(y_0 - s) \end{aligned}$$

Equation 3 shows that when there is an increase in either of the errors (increase in G or decrease in I) on the left-hand side of the equation, defendants find deviation convenient for lower levels of b (on the right-hand side). However, given the concavity of the utility function, the negative impact of type-I errors $(1 - I)$ on the threshold level $\tilde{b}_{eu}$ and thus on social welfare is stronger than that of type-II errors (G). To see why, note that $U(y_0) - U(y_0 - s) > U(y_0 + b) - U(y_0 + b - s)$. In order to maintain the same level of deterrence, a given percentage increase of $1 - I$ must be compensated by a smaller percentage decrease of G . Therefore, assuming risk aversion, type-I errors $(1 - I)$ create more disutility and thus induce more deviation than comparable type-II errors (G); therefore, social costs are minimized for a $\tilde{e}^* > \tilde{e}_{neutral}$. The opposite result holds if we instead assume risk-seeking behavior.

Another interesting extension concerns the introduction of loss aversion: a departure from the expected utility framework that has been incorporated in models such as the cumulative prospect theory (Dhami and al Nowaihi 2013). These models build on the empirical observation that people tend to think of possible outcomes of a choice under uncertainty relative to a certain reference point and tend to prefer the avoidance of losses (outcomes below the reference point) than the acquisition of comparable gains (outcomes above the reference point). Incorporating

reference-dependent preferences and loss aversion in the model is not trivial (see Nicita and Rizzolli 2014); however, the intuition and the results are quite simple: type-I errors always imply a potential loss relative to the status quo, while this is not necessarily true for type-II errors. To conclude, in presence of loss aversion, type-I errors ($1 - I$) represent a net loss and impact the defendant value function more than comparable type-II errors (G); therefore, social costs are minimized for a $\tilde{e}^* > \tilde{e}_{\text{neutral}}$.

Cost of Sanctions

So far we have assumed that the sanction s is monetary and that it implies – once imposed – a costless transfer from the defendant to the society. However, the imposition of sanctions implies both private costs of punishment to defendants and to society as well. Nonmonetary sanctions are a social cost (Shavell 1987) as their imposition implies a disutility for the defendant that is not transferred to society. Furthermore, all sanctions – including monetary fines – must be administered and therefore imply a social cost (Polinsky and Shavell 1992).

Define c as the total cost (both to the defendant and to the society) of imposing a sanction. The social welfare function (assuming risk neutrality) should be rewritten as the following:

$$SW = [1 - Z(\tilde{b})]h + [1 - Z(\tilde{b})](1 - G)c + Z(\tilde{b})(1 - I)c \quad (4)$$

The first term of Eq. 4 represents the harm/externality of deviating, as before. The second term represents the expected total costs of imposing sanctions on deviating defendants, and the third term represents the expected total costs of punishing complying defendant (type-I errors). The problem lies in defining the optimal $\tilde{e}$ that minimizes the expected total costs from crime, including the costs of punishment. As before, the first term is minimized for $\tilde{e} = \tilde{e}_{\text{neutral}}$. However, the second and third terms are minimized for $\tilde{e} \rightarrow \infty$. In fact for an evidence threshold $\tilde{e}$

arbitrarily high, the probability of correctly imposing a sanction on a deviating defendant ($1 - G$) or erroneously imposing a sanction on a complying defendant ($1 - I$) decreases to zero and – since nobody is punished – there are no costs of punishment for society. When social costs are considered, the costs of harm implied by the first term must thus be balanced against the costs of punishment of the second and third term. Therefore, in the presence of costs of punishment, the social costs of harm must be weighted against the social costs of punishment, and therefore social costs are minimized for $\tilde{e}^* > \tilde{e}_{\text{neutral}}$. This result is based on Rizzolli and Saraceno (2013).

Identity Errors

Lando (2006) introduced a distinction between *mistakes of act* and *mistakes of identity*. *Mistakes of act* happen when a defendant is judged deviating when in fact he was complying. These are adjudicative errors we have been focusing on so far, for which the main concern of the adjudicative authority is whether there actually was any deviation at all. Note that, in case of mistakes of act, the two errors are independent: an increase in wrongful convictions does not imply any change in the number of wrongful acquittals. Then there are *mistakes of identity*, by which in the presence of deviations that can be easily observed, such as a murder or a robbery in the context of crime, the wrong person can be incriminated for an act that actually did happen. These are the cases where the occurrence of the deviation cannot be denied and the authority is concerned with who committed the crime. Note that in this case the two errors for a given crime are linked, as the conviction of an innocent person implies the acquittal of the person actually responsible for it.

Suppose that at time t_1 there exists an exogenous probability $\beta_{i,g}$ that a defendant is sanctioned for a deviation that has already happened at t_0 and which the subject is not responsible for (a mistake of identity). This exogenous probability can vary depending on the decision of the defendant at t_1 : it seems reasonable to assume that abstaining from a crime at t_1 reduces the

probability of a mistake of identity, so that $\beta_i \leq \beta_g$. Thus the returns from conforming at t_1 are $E\pi_I = y_0 - (1 - I)s - \beta_i s$, while the returns from deviating are $E\pi_G = y_0 + b - (1 - G)s - \beta_g s$. The threshold level of b which implicitly defines the conforming population is

$$\tilde{b}_{\text{identity}} = (1 - (1 - I) - G)s - (\beta_i - \beta_g)s \quad (5)$$

Inspection of Eq. 5 and comparison with Eq. 1 highlight the role of mistakes of identity vis-à-vis deterrence implicitly defined by $\tilde{b}_{\text{identity}}$. The first part is equal to Eq. 1, while in the second part, if $\beta_i = \beta_g$ as Lando (2006) hypothesized, then identity errors occurred at t_0 have no impact on deterrence at t_1 . However, if $\beta_i < \beta_g$, then identity errors actually have a positive impact on deterrence. The reason is intuitive: the decision to deviate in t_1 triggers a net increase in the probability of being wrongfully convicted because of a mistake of identity. Of course this result is based on the assumption that the probability of identity errors in t_1 is determined exogenously, and it is not a function of $\tilde{e}$. Furthermore, identity errors impose a necessarily constraint between the input of wrongful acquittals and the output of wrongful identity convictions; Garoupa and Rizzolli (2013) show that once this constraint is considered, mistakes of identity have a net negative impact on deterrence.

Errors and the Precaution of Harm

Another main area where the role of adjudicative errors has been explored is tort law (see Png 1986; Lando and Mungan 2017, among others). The standard model of tort law substitutes the dichotomous choice between complying and deviating with a continuous choice about the level of activity/care. We will discuss the main implications below. However, some novel conclusions can be drawn also from applying the dichotomous choice model. In this framework, the defendant chooses between *conforming* to the prescribed standard of care or *deviating* and not taking any precaution. Since taking precautions is costly, we can interpret b as the opportunity cost of conforming (by deviating, the defendant saves

b). Furthermore, the sanction is equal to the harm inflicted since the goal of the tort system is compensation, and the decision to conform only reduces the expected harm: when conforming, harm h_i is produced with probability α_i , while when deviating, harm h_g is produced with probability α_g , where $h_g > h_i$ and $\alpha_g > \alpha_i$. Adjudicative errors can occur in the usual way, and therefore, a risk-neutral defendant's returns from conforming are $E\pi_I = y_0 - (1 - I)\alpha_i h_i$, while the returns from deviating are $E\pi_G = y_0 + b - (1 - G)\alpha_g h_g$. All defendants for which $E\pi_i \geq E\pi_g$ will conform and therefore the threshold level of b which implicitly defines the conforming population is

$$\tilde{b}_{\text{care}} = (1 - G) \cdot \alpha_g h_g - (1 - I) \cdot \alpha_i h_i \quad (6)$$

By comparing Eq. 6 with Eq. 1, one immediately realizes that type-I errors have a smaller impact on the incentive to comply than type-II errors as $\frac{\partial b}{\partial(1-I)} = \alpha_i h_i < \frac{\partial b}{\partial G} = \alpha_g h_g$; this is because complying causes a smaller expected harm. Also social welfare changes as now also complying defendant causes harm. To find out the optimal $\tilde{e}$, we compute $\frac{\partial SW}{\partial e} = 0$ and thus

$$\partial Z(\tilde{b})\alpha_i h_i + \partial(1 - Z(\tilde{b}))\alpha_g h_g = (\alpha_i h_i i(\tilde{e}) - \alpha_g h_g g(\tilde{e}))(\alpha_i h_i - \alpha_g h_g) = 0 \quad (7)$$

Rearranging Eq. 7, we have that $i(\tilde{e}) = \frac{\alpha_g h_g}{\alpha_i h_i} g(\tilde{e})$, and since $\frac{\alpha_g h_g}{\alpha_i h_i} > 1$, the equality can be satisfied only for $\tilde{e}^* < \tilde{e}_{\text{neutral}}$. We can thus conclude that when defendants face a dichotomous choice between complying and causing a smaller expected harm and deviating and causing a larger expected harm, type-I errors impact deterrence less than type-II errors, and therefore welfare is maximized for a level of evidentiary standard smaller than the neutral one.

Precautionary Activities and Chilling of Desirable Behavior

In both the crime and the tort contexts, the choice of deviating causes social harm at least in

expected terms. Compliance causes no harm in the crime context, while it produces a smaller social harm in the tort context. In many situations, however, defendant compliance can have both harmful consequences and benign ones. One may think of the case of competition policy, where the threat of antitrust sanctions may discourage efficient, pro-competitive behavior; another case may be medical malpractice, where worries about false positives may prevent cost-effective care. Kaplow's (2011) model envisages a population that can engage in a harmful act that produces a private benefit as well as a negative externality and another population that can only engage in a benign act that produces no externality. The two types of act are initially indistinguishable to the authority, but the adjudicative procedure gives rise to an evidence signal e that is higher for harmful acts than for benign ones. As before, the authority sanctions subjects whose acts produce an evidence signal higher than a certain cutoff value $\tilde{e}$. However, now the expected sanction raises both the costs of the harmful act and that of the benign one, thus chilling desirable behavior. Kaplow (2011) shows that the optimal $\tilde{e}$ that equates the (falling) marginal benefits of deterring harmful acts with the (rising) marginal costs of chilling benign acts is such that $\tilde{e}^* > \tilde{e}_{\text{neutral}}$. Intuitively it is generally optimal to raise the sanction and simultaneously raise $\tilde{e}$, holding deterrence constant. In fact the only consequence of this policy is a reduction in chilling costs.

A similar model is proposed by Mungan (2011) where subjects can choose between *inaction* (precautionary activity) and *action*, and this second choice can produce *no externality* (desirable behavior) or a *negative externality* (harmful activity). The authority cannot distinguish with certainty whether the activity is harmful or benign but must rely on an evidence signal e and balance the usual errors' trade-off. The expectation of sanctions wrongfully imposed on desirable behavior induces subjects at the margin to switch over to precautionary activities. Again, Mungan (2011) shows that the optimal evidence threshold is such that $\tilde{e}^* > \tilde{e}_{\text{neutral}}$.

Judicial Errors When the Choice of Care Is Continuous

In the model presented so far, the defendant's choice between complying and deviating is dichotomous. However, other situations like torts are best described by a continuous choice of care level x . In the prevailing model of tort, a legal standard $\bar{x}$ is set in order to determine liability by a potential injurer: the defendant avoids liability if his level of care is equal or above the standard one which is usually equated to the optimal level x^* . Craswell and Calfee (1986) introduce legal errors in this context by proposing a model where such legal standard is uncertain in the sense that defendants who choose a level of care x only know that there is a probability $F(x)$ (decreasing in x) that they will be sanctioned so that choosing higher levels of x decreases the probability to be punished. So if they choose $\underline{x} < x^*$, there is a $1 - F(\underline{x})$ probability of type-II error (the defendant is not made liable even if he took less than the efficient level of care), while if they choose $\bar{x} > x^*$, there is a $F(\bar{x})$ probability of type-I error (the defendant is made liable even if he took enough care). Assuming that also both the sanction s and the opportunity cost of care b are increasing functions of x , Craswell and Calfee (1986) show that with respect to the socially optimal level of x , the defendant's choice of x can be either undercomplying (the defendant chooses $\underline{x} < x^*$) or overcomplying ($\bar{x} > x^*$). This is because, on one hand, there is always a positive chance $1 - F(x)$ of acquittal, and this increases the returns of taking lower levels of care. But on the other hand, the expected sanction depends on the probability $F(x)$, and this can be driven down by increasing the level of care. The relative impact of these two countervailing effects on the final level of x can be of either sign as it depends on various features of the legal environment and in particular on the amount of uncertainty. Craswell and Calfee (1986) show that under some plausible assumptions concerning the distribution of errors, defendants will usually take an excessive level of care. In other words, while the possibility of escaping liability when the defendant has not taken enough precaution (type-II error) has the usual adverse

effect on the incentives to take precaution, the possibility of being wrongfully held liable even when one has taken enough precautions (type-I error) induces the defendant to increase the level of precautions (under some plausible conditions).

Further Effects on Evidentiary Standards and on Type-I Errors

In addition to the literature survey above, many authors have also explained the high evidence threshold usually observed in legal trials using deterrence-based arguments that point at (i) biased evidence selection (Schrag and Scotchmer 1994), (ii) parties' evidence production expenditure (Yilankaya 2002), (iii) optimal exercise of care by parties (Demougin and Fluet 2006), (iv) marginal deterrence (Ognedal 2005), (v) repeated offenders (Chu et al. 2000), and (vi) emotional costs of indignation (Nicita and Rizzolli 2014).

Furthermore, without making reference to a specific utilitarian approach based on the deterrence rationale, a consistent number of papers simply postulate that wrongful convictions of innocents are morally repugnant and thus inherently worse than type-II errors. Arguments that justify this position are reviewed in Epps (2015). These arguments are mainly deontological and transcend the utilitarian framework (See also ► [“Retributivism”](#)) although they can still be considered in our model by overweighting the impact of type-I errors on the social welfare function along the lines of Miceli (2009).

Empirical Relevance

The role of type-II errors has been greatly analyzed empirically and experimentally; there exists a vast literature testing Becker's deterrence hypothesis with real data on incarceration and on the death penalty (Chalfin and McCrary 2017). There is also a small stream of literature testing the deterrence hypothesis in the lab (see literature cited in Khadjavi 2015); (See also ► [“Experimental Law and Economics”](#)).

Most of this literature, however, ignores type-I errors and their impact on deterrence and behavior. This asymmetry is easy to understand once one considers that type-II errors (crimes that go unpunished) are far more easy to be observed and measured than type-I errors (wrongful convictions can in fact be mistaken for correct convictions). Empirical research on type-I errors has only recently taken off either comparing agreement rates of judges and juries (Gould et al. 2014) or by using DNA testing introduced in the 1990s. Many innocent defendants used DNA testing to clear themselves after conviction whenever biological evidence from the crime scene had been retained. By adopting this strategy, Risinger (2007) estimated the type-I error rate in capital rape-murder cases to be between 3.3% and 5% in 1982–1989. Gross and O'Brien (2008), using post-1973 US data on death sentences (See also ► [“Death Penalty”](#)), estimated a type-I error frequency of wrongful death sentences to be at least 2.3%. Most of this literature is concerned with measuring the magnitude of type-I errors and less with the assessment of the impact of type-I errors on general deterrence (Gould et al. 2014). A small number of controlled lab experiments try to assess the impact of type-I errors on deterrence: Grechenig et al. (2010) first showed that both errors greatly undermine deterrence in a voluntary contribution mechanism (VCM) type of game. Rizzolli and Stanca (2012) disentangled the effects of each type of error and found that type-I errors are more detrimental to deterrence than type-II errors. Marchegiani et al. (2016) found the same effect within a principal-agent setting, while Markussen et al. (2016) using a VCM design found instead the opposite effect: that type-I errors are less detrimental than type-II errors.

References

- Blackstone W (1769) Commentaries on the laws of England, vol 4. Clarendon Press, Oxford
- Chalfin A, McCrary J (2017) Criminal deterrence: a review of the literature. *J Econ Lit* 55:5–48
- Chu CC, Hu S-C, Huang T-Y (2000) Punishing repeat offenders more severely. *Int Rev Law Econ* 20:127–140

- Craswell R, Calfee JE (1986) Deterrence and uncertain legal standards. *J Law Econ Org* 2:279–303
- Dekay ML (1996) The difference between Blackstone-like error ratios and probabilistic standards of proof. *Law Soc Inq* 21:95–132
- Demougin D, Fluet C (2006) Preponderance of evidence. *Eur Econ Rev* 50:963–976
- Dhami S, al Nowaihi A (2013) An extension of the Becker proposition to non-expected utility theory. *Math Soc Sci* 65:10–20
- Epps D (2015) The consequences of error in criminal justice. *Harv Law Rev* 128:1065
- Garoupa N, Rizzolli M (2013) Wrongful convictions do lower deterrence. *J Inst Theor Econ* 168:224–231
- Gould JB, Carrano J, Leo RA, Hail-Jares K (2014) Predicting erroneous convictions. *Iowa Law Rev* 99:471–2299
- Grechenig K, Nicklisch A, Thöni C (2010) Punishment despite reasonable doubt – a public goods experiment with sanctions under uncertainty. *J Empir Leg Stud* 7:847–867
- Gross SR, O'Brien B (2008) Frequency and predictors of false conviction: why we know so little, and new data on capital cases. *J Empir Leg Stud* 5:927–962
- Kaplow L (2011) On the optimal burden of proof. *J Polit Econ* 119:1104–1140
- Kaplow L (2012) Burden of proof. *Yale Law J* 121:738–859
- Khadjavi M (2015) On the interaction of deterrence and emotions. *J Law Econ Organ* 31:287–319
- Lando H (2002) When is the preponderance of the evidence standard optimal? *Geneva Pap Risk Insur Issue Pract* 27:602–608
- Lando H (2006) Does wrongful conviction lower deterrence? *J Leg Stud* 35:327–338
- Lando H, Mungan MC (2017) The effect of type-I error on deterrence. *Int Rev Law Econ* 53:1
- Marchegiani L, Reggiani T, Rizzolli M (2016) Loss averse agents and lenient supervisors in performance appraisal. *J Econ Behav Organ* 131:183–197
- Markussen T, Putterman L, Tyran J-R (2016) Judicial error and cooperation. *Eur Econ Rev* 89:372–388
- Miceli TJ (2009) Criminal procedure. Edward Elgar Publishers, vol 3 of Criminal law and economics – encyclopedia of law & economics, Edward Elgar (ed), Cheltenham
- Mungan M (2011) A utilitarian justification for heightened standards of proof in criminal trials. *J Inst Theor Econ* 167:352
- Nicita A, Rizzolli M (2014) In Dubio Pro Reo. Behavioral explanations of pro-defendant bias in procedures. *CESifo Econ Stud* 60:554. ift016
- Ognedal T (2005) Should the standard of proof be lowered to reduce crime? *Int Rev Law Econ* 25:45–61
- Png IPL (1986) Optimal subsidies and damages in the presence of judicial error. *Int Rev Law Econ* 6:101–105
- Polinsky A, Shavell S (1992) Enforcement costs and the optimal magnitude and probability of fines. *J Law Econ* 35:133–148
- Risinger DM (2007) Innocents convicted: an empirically justified factual wrongful conviction rate. *J Crim Law Criminol* 97:761–806
- Rizzolli M, Saraceno M (2013) Better that ten guilty persons escape: punishment costs explain the standard of evidence. *Public Choice* 155:395–411. <https://doi.org/10.1007/s11127-011-9867-y>
- Rizzolli M, Stanca L (2012) Judicial errors and crime deterrence: theory and experimental evidence. *J Law Econ* 55:311–338
- Schrag J, Scotchmer S (1994) Crime and prejudice: the use of character evidence in criminal trials. *J Law Econ Org* 10:319–342
- Shavell S (1987) The optimal use of nonmonetary sanctions as a deterrent. *Am Econ Rev* 77:584–592
- Yilankaya O (2002) A model of evidence production and optimal standard of proof and penalty in criminal trials. *Can J Econ* 35:385–409

U

Ultra-broadband Networks

- [Next-Generation Access Networks](#)

Ultrafast Broadband Networks

- [Next-Generation Access Networks](#)

Underground Economy

- [Informal Sector](#)

Unofficial Economy

- [Shadow Economy](#)

Using Updated Beliefs in Modeling Good Faith and Its Impact on Pareto Efficiency

- [Good Faith and Game Theory](#)

Value Incommensurability

Tomáš Sobek¹ and Josef Montag^{2,3}

¹Faculty of Law, Masaryk University,
Brno, Czech Republic

²International School of Economics, Kazakh-
British Technical University, Almaty, Kazakhstan

³Faculty of Law, Charles University, Prague,
Czech Republic

Definition

Incommensurability is a failure, or impossibility, to make value comparisons of different choices. It represents an important criticism of the proportionality test in judicial decision-making and of the cost-benefit analysis as a tool for assessing public policy. This entry summarizes the basic theories of value incommensurability.

Motivation and Delineation

Incommensurability is a failure, or impossibility, to make value comparisons. Its consequence is that choices among incommensurable options cannot be led by reason, i.e., cannot be rational. A few examples may help to illustrate such choices: upon arrival at Auschwitz, Sophie is forced to choose which one of her two children would be gassed and which would proceed to the labor camp; an individual needs to choose

whether to pursue a career as farmer or as a violinist; one wonders, whether Mozart is a better composer than Michelangelo was sculptor; a city needs to find a trade-off between air quality and heating during winter time.

Conceptually, it is useful to distinguish two types of incommensurability, positive and normative. Positive incommensurability obtains when a value comparison is objectively (technically) impossible, as is arguably the case for Sophie. It would be difficult to argue that her choice to save the son and sacrifice the daughter was arrived at by comparing the two and finding that her son's life had more value to her than the life of her daughter. In fact, the choice itself is a tragedy here; quite apart of the tragic outcome leading to the death of one of her children. This is why Sophie never recovers from that experience and eventually commits suicide. Normative incommensurability is the claim that some things ought not to be compared, since doing so would lead to undesirable consequences. One may say, for instance, that trying to express air quality in terms of money would lead to its "commodification" and demeaning.

At the individual level, people constantly make choices between alternative courses of action whose values are difficult, or even impossible, to ascertain objectively. Should "I get the green shirt or blue pants? Or, should I invite friends over for dinner instead?" "Should I buy more children's books or a child seat?" "Should I cross the road right here, saving myself 15 s of walk, or go to the

zebra, lowering the chance of being run over?" And so on.

Economists have long been aware of issues pertaining to objective valuation and particularly cautious when it comes to the possibility of interpersonal comparisons. They usually assume that individuals' choices follow from innate preferences (see Stigler and Becker 1977). Preferences enable individuals to determine the relative value of alternative courses of action to *them*, guiding their choices. And, as long as those choices do not affect others, or do so to a negligible extent, they are best left to those very individuals to be made. In this line of thinking, incommensurability is indistinguishable from indifference, except for extreme dilemmas, such as Sophie's, or curiosities, such as choosing between Mozart and Michelangelo. In fact, economists avoid the problem of incommensurability altogether by assuming it away, that is by assuming that preferences are rational, i.e., complete and transitive (see Mas-Colell et al. 1995, chap. 1). Individual rationality, i.e., the ability to make reasonable choices in accordance with one's individual preferences, is also a default assumption underlying much of private law (Lucy 2007, chap. 3; Voyiakis 2017, chap. III.C).

However, issues arise when individuals' choices have significant external effects onto others, as is commonly the case. In such cases, the society must somehow determine what are the appropriate courses of action (Coase 1960). In other words, it must be determined what *we* should do. Are people allowed to wear green shirts that are out of fashion? Are people allowed to wear hijabs in public schools? In such cases, interpersonal comparisons, even if implicit, are inevitable.

At the societal level, such *we*-type decisions are frequently made through two types institutions: political (parliaments, executive) and judicial. While incommensurability issues pertain various types of decision-makers and contexts, judges are in a specific position because (a) they inevitably make decisions that favor one party (broadly defined) at the expense of another, (b) they do so on behalf of the society, and (c) they do so as an unelected bureaucratic entity,

i.e., without democratic legitimacy. In effect, judges are tasked to make choices for the society, the *we*-type choices, similarly as individuals make choices for themselves, the *I*-type choices. This is when the incommensurability issues loom large.

When the standard, legalistic, approaches fail, that is when the law does not state clear "social preferences," courts opt for the proportionality test (see the entry "[► Proportionality Test](#)" in this encyclopedia for more details). This legal method allows them to decide hard cases, i.e., cases where two or more legitimate rights collide and a decision necessarily leads to one right prevailing at the expense of another. In order to decide such cases correctly, the court balance the satisfaction of some rights and the damage to other rights resulting from a judgement. The proportionality test can be viewed as a method through which the courts uncover the social preference and that forces them to do so transparently.

This enterprise seems to be, however, dubious if the colliding rights are incommensurable. The intuition was put vividly by Antonin Scalia who characterized the balancing approach as something "more like judging whether a particular line is longer than a particular rock is heavy" (Bendix Autolite Corp. v. Midwesco Enterprises 1988). Incommensurability is also presented as an objection to the cost-benefit analysis, which assumes that any available options can be valued along a common scale whose units are typically money (Frank 2008; Sunstein 1994).

Positive Incommensurability

Failure of Transitivity

Two options are commensurable if one of them is better than the other or if their value is the same. Two options are incommensurable if none of the three value relations hold. For instance, if it is neither true that an option A is better than an option B nor that B is better than A, and yet it is not the case that they are of the same value. Take the example of Mozart and Michelangelo. Arguably, neither was more creative than the other nor were their creativities the same. One can

then conclude that their creative qualities are incommensurable.

Incommensurability is the failure of transitivity, which can be shown by the Small-improvement Argument (Raz 1986, chap. 13, p. 325). Suppose we are comparing Mozart and Michelangelo and agree that neither is better than the other. Can we say that they are of the same quality? Think of a hypothetical Mozart that is in some dimension slightly better than the actual one and is the same in all other respects. Clearly, this “Mozart⁺” is better than the original Mozart. However, if it is still the case that Mozart⁺ isn’t better than Michelangelo, then Mozart and Michelangelo are incommensurable.

Implications for Rational Choice

At least at first glance, incommensurability seems to provide disturbing prospects, because it “can leave us paralysed, not knowing what to choose” (Waldron 1994, p. 815). True dilemma means that there is no objective value basis to say about one decision rather than the other that it is the right answer.

From the point of view of practical rationality, we can distinguish three normative statuses of action: rationally demanded, rationally prohibited, and rationally permitted. If incomparability embodies the third status, then it renders both alternatives for choice rationally permissible, that is, there is sufficient reason to choose either alternative. Therefore, from the point of view of the rationality requirement, the choice of any of the two alternatives is faultless and blameless (Raz 1999, chap. 2).

Parity

Chang (2002) argues that failure of transitivity does not always imply incommensurability. Building up on the Small-improvement Argument, she develops the Unidimensional Chaining Argument to show this. Returning to our example with Michelangelo and Mozart, think of a very poor sculptor, she names him Talentlessi. Clearly, Talentlessi will be a worse sculptor than Michelangelo and arguably less creative than Mozart. Now think of adding units of talent to him. Talentlessi⁺ will be better than Talentlessi but still worse than both Mozart and

Michelangelo. The same will probably be the case for Talentlessi⁺⁺, and so on. However, as Talentlessi becomes more and more talented, his qualities will be approaching those of Michelangelo. At some point, we will have Talentlessi^{*} who is exactly like Michelangelo.

Because Talentlessi^{*} as well as Michelangelo are comparable with all the lesser Talentlessis and because those are comparable with Mozart, Michelangelo and Mozart are comparable too.

This argument relies on the Small Unidimensional Difference Principle stating that a small unidimensional difference cannot trigger incomparability where before there was comparability. Thus, if Talentlessi is comparable to both Michelangelo and Mozart, so must be Talentlessi⁺, Talentlessi⁺⁺, and eventually Talentlessi^{*}.

Chang (2002) argues that in situations of transitivity failure, i.e., none of the three value relations hold, but when the objects are comparable, a fourth relation obtains: *parity*. Because Mozart and Michelangelo are comparable and yet neither is superior nor are they the same, one may say that they are on par.

Normative Incommensurability and Commodification

Certain transactions give rise to moral opposition and some are even illegal. Some contemporary authors criticize the exchange of money for health care, human organs, personal privacy, human dignity, voting rights, pollution rights, public education, surrogacy and sexual services, works of art, and so on (see Sunstein 1994). These value-bearers, due to their special moral, sacred, cultural, or generally symbolic significance, are supposed to be inherently incompatible with market exchanges and qualified as incommensurable (or incomparable) to money.

Commodification is usually understood as an assignment of price to something not previously considered in economic terms. This may then transforms goods, relationships, services, ideas, and also people’s lives into objects of trade. Here, incommensurability is not the failure of

value comparison. Rather, the argument here is that even if those values are in principle comparable, doing so has negative consequences, particularly in the sense of demeaning their true value.

In order to understand the phenomenon of normative incommensurability, one needs to think about it in the broader social and cultural context. Interpersonal transactions have no universal nature; their moral and symbolical significance is dependent on social conventions of relevant community (Cohen 2003). When people strictly adhere to certain principles or stubbornly reject some trade-offs, it does not have to mean that they are irrational. Perhaps they signal that they are loyal to certain values that their community recognizes as significant. At the same time, they build, intentionally or not, a reputation as good people who can be reliable partners in different collaborative projects (Posner 1998, 2000, p. 193).

One example of criticized commodification concerns the tort law. Radin (1993) criticizes the conception of harm compensation because it conceives the harm to the injured victim as commensurable with money. The market approach takes the bodily harm in the same way as any economic cost and she rejects it as a form of inappropriate reductionism. The market idea that each person's attributes have a monetary equivalent is incompatible with an ideal of individual uniqueness (Radin 1987). If bodily injury and money aren't commensurable to each other, then no payment of money can restore persons to the status quo ante. This means that the corrective justice, as a fundamental ethical principle of the tort law, must demand something other than rectification. Radin (1993) proposes a noncommodified conception of compensation, which requires a payment as a way both to bring the wrongdoer to recognize that he has done wrong and to show the victim that her rights are taken seriously.

Another example is the social institution of friendship. Perhaps we can hire someone to behave like our friend but it will not be a real friendship. It appears that commodification is detrimental to the very nature of some values by imposing a substitution regime on them (Sandel 2012, 2013). Friendship can't be for sale without

losing its own nature. If we were to offer someone a financial compensation for giving up his friend, he would probably have rejected such an offer as an insult. This is different than when someone is offended by getting an inappropriately low financial offer for something that is normally for sale. At the same time, it does not mean that friendship is more valuable than any amount of money. The main explanatory reason of the different approach is rather the fact that we understand friendship and business as two mutually independent social institutions that fulfill very different purposes (Regan 1989; Sunstein 1994, see also Khalil and Marciano 2018).

A third example is blood donation. Selling blood is something different from a free donation, where blood is symbolically valued as the gift of life itself (Anderson 1990). Of course, sold blood can save human lives as well. But we may need to take account of nonmarket norms that people recognize as morally significant. Titmuss (1971) argued that the commodification of blood erodes the people's sense of moral duty and therefore their willingness for free donation is significantly weakened. As a result, the number of voluntary donors is declining and they are replaced by poor people who sell their blood as a way of making money. This fact can also have a harmful effect on the quantity and quality of available blood (Titmuss 1971).

Summary

Positive value incommensurability presents a challenge for rational decision-making, particularly by public authorities and courts. A way out is to say that any action (inaction) is rationally permissible as neither decision is better or worse than the other. Normative incommensurability, on the other hand, typically seems to serve as an argument *for* a particular action or inaction. A full-fledged discussion of various nuances of the incommensurability literature and the criticisms is beyond the scope of this entry. In order to learn more about the phenomenon, its criticisms, and implications, we refer the reader to the readings below.

Cross-References

- [Conflict of Laws](#)
- [Cost–Benefit Analysis](#)
- [Judicial Decision-Making](#)
- [Proportionality Test](#)

References

- Anderson E (1990) The ethical limitations of the market. *Econ Philos* 6:179–205
- Bendix Autolite Corp. v. Midwesco Enterprises (1988) <https://www.law.cornell.edu/supremecourt/text/486/888>
- Chang R (2002) The possibility of parity. *Ethics* 112:659–688
- Coase RH (1960) The problem of social cost. *J Law Econ* 3:1–44
- Cohen G (2003) The price of everything, the value of nothing: reframing the commodification debate. *Harv Law Rev* 117:689–710
- Frank RH (2008) Why is cost-benefit analysis so controversial? In: Hausman DM (ed) *The philosophy of economics: an anthology*. Cambridge University Press, New York, pp 251–269
- Khalil EL, Marciano A (2018) A theory of tasteful and distasteful transactions. *Kyklos* 91:110–131
- Lucy W (2007) *Philosophy of private law*. Oxford University Press, Oxford
- Mas-Colell A, Whinston MD, Green JR (1995) - *Microeconomic theory*. Oxford University Press, Oxford
- Posner E (1998) The strategic basis of principled behavior: a critique of the incommensurability thesis. *Univ Pa Law Rev* 146:1185–1214
- Posner E (2000) *Law and social norms*. Harvard University Press, Cambridge, MA
- Radin M (1987) Market inalienability. *Harv Law Rev* 100:1849–1937
- Radin M (1993) Compensation and commensurability. *Duke Law J* 43:57–86
- Raz J (1986) *Morality of freedom*. Oxford University Press, Oxford
- Raz J (1999) *Engaging reason*. Oxford University Press, Oxford
- Regan D (1989) Authority and value: reflections on Raz's morality of freedom. *South Calif Law Rev* 62:995–1095
- Sandel M (2012) *What money can't buy: the moral limits of markets*. Farrar, Straus and Giroux, New York
- Sandel M (2013) Market reasoning as moral reasoning: why economists should re-engage with political philosophy. *J Econ Perspect* 27:121–140
- Stigler GJ, Becker GS (1977) De gustibus non est disputandum. *Am Econ Rev* 67:76–90
- Sunstein C (1994) Incommensurability and valuation in law. *Mich Law Rev* 92:789–861
- Titmuss R (1971) *The gift relationship: from human blood to social policy*. Pantheon, New York
- Voyiakis E (2017) *Private law and the value of choice*. Hart Publishing, Oxford
- Waldron J (1994) Fake incommensurability: a response to professor Schauer. *Hastings Law J* 45:813–824

Further Reading

- Adler M (1998) Law and incommensurability: introduction. *Univ Pa Law Rev* 146:1169–1184
- Alder J (2001) Incommensurable values and judicial review: the case of local government. *Public Law* 4:717–735
- Anderson E (1993) *Value in ethics and economics*. Harvard University Press, Cambridge
- Anderson E (2000) Why commercial surrogate motherhood unethically commodifies women and children: reply to McLachlan and Swales. *Health Care Anal* 8:19–26
- Binmore K (2009) Interpersonal comparison of utility. In: Ross D, Kincaid H (eds) *Oxford handbook of philosophy of economics*. Oxford University Press, Oxford, pp 540–559
- Brennan J, Jaworski PM (2015) Markets without symbolic limits. *Ethics* 125:1053–1077
- Chang R (1997) Introduction. In: Chang R (ed) *Incommensurability, incomparability, and practical reason*. Harvard University Press, Cambridge, MA
- Chang R (2012) Are hard choices cases of incomparability? *Philos Issues* 22:106–126
- Chang R (2013) Incommensurability (and incomparability). In: LaFollette H (ed) *International encyclopedia of ethics*. Blackwell Publishing, Malden, pp 2591–2604
- Cook PJ, Krawiec KD (2018) If we allow football players and boxers to be paid for entertaining the public, why don't we allow kidney donors to be paid for saving lives? *Law and Contemporary Problems* 81:9–34
- Craswell R (1998) Incommensurability, welfare economics, and the law. *Univ Pa Law Rev* 146:1419–1464
- Ellis S (2008) The main argument for value incommensurability (and why it fails). *South J Philos* 46:27–43
- Epstein R (1995) Are values incommensurable, or is utility the ruler of the world? *Utah Law Rev* 3:683–715
- Ertman M, Williams JC (2005) *Rethinking commodification: cases and readings in law and culture*. New York University Press, New York
- Eyal N, Tieffebach E (2016) Incommensurability and trade. *Monist* 99:387–405
- Friedman D (1982) What is 'fair compensation' for death or injury? *Int Rev Law Econ* 2:81–93
- Griffin J (1977) Are there incommensurable values? *Philos Public Aff* 7:39–59
- Hsieh N (2016) Incommensurable values. *Stanford encyclopedia of philosophy*. Online at <https://plato.stanford.edu/entries/value-incommensurable>. Last accessed on 30 May 2018
- Kelly C (2008) The impossibility of incommensurable values. *Philos Stud* 137:369–382

- Larmore CE (1987) Patterns of moral complexity. Cambridge University Press, Cambridge
- Lavetti K (forthcoming) The estimation of compensating wage differentials: lessons from the deadliest catch. *J Bus Econ Stat*. <https://doi.org/10.1080/07350015.2018.1470000>
- Pildes RH, Anderson ES (1990) Slings arrows at democracy: social choice theory, value pluralism, and democratic politics. *Columbia Law Rev* 90:2121–2214
- Regan D (1989) Authority and value: reflections on Raz's morality of freedom. *South Calif Law Rev* 62:995–1095
- Sunstein C (1996) Health-health tradeoffs. *Univ Chic Law Rev* 63:1533–1571

Vertical Integration

Giuseppina Gianfreda
University of Tuscia, Viterbo, Italy

Abstract

Economists have long been inquiring into the determinants of vertical integration. Theories which explain the rationale for firms' decision to expand vertically also predict different implications in terms of welfare. A body of literature traces vertical integration to the problem of the origin and the boundary of a firm and explains integration on the bases of the economization of costs related to market transactions and contract incompleteness; this literature dates back to Coase's (*Economica* 4:386–405, 1937) seminal article on the nature of the firm and has been developed within the Transaction Costs Economics framework and subsequently under the Property Right approach. Another set of contributions studies vertical integration with respect to firms' competitive environment. In this context, two main views emerge: on the one hand vertical integration has been considered as a way to exploit market power and implement (otherwise banned) price or exclusionary strategies and may thus raise antitrust concerns; on the other hand, integrating firms can cope with negative "vertical" externalities and increase efficiency. Another set of contributions, often referred to

as vertical equilibrium or dynamic models, explains vertical integration with respect to the degree of market maturity, as the Stigler model, or to demand fluctuations. Other models explain the incentive for vertical integration by factors related to uncertainty, i.e., over the supply of the input good or in the demand of the final good.

Introduction

Vertical integration refers to the organization of successive stages of production or distribution – i.e., a supplier and a retailer – within a single firm. Two aspects are relevant in the definition of vertical integration: (i) the ownership or control by the same firm over the successive stages of production or distribution process and (ii) the substitution of external with internal exchanges.

Vertical integration can be *full* or *partial*. Under the first aspect, i.e., ownership or control, integration is full (partial) if the firm acquires all (part of) the shares in a vertically related firm. Under the second aspect, i.e., the internalization of exchanges, full vertical integration means that the entire output of an upstream unit is employed in a downstream unit or the entire quantity of an intermediate input in the downstream unit is obtained from the upstream process, while partial integration means that most of the output of the upstream unit is employed as most of the input in the downstream unit, that is to say that the stages of production are not internally self-sufficient.

Vertical integration can result from a process of *internal* growth or may be due to *external* expansion, i.e., the merging of a firm with another firm operating in a successive stage of production or distribution. Integration can take place in two direction, i.e., *upward* and *forward*; in the first case, a firm expands to get control over an upstream firm, for example a manufacturer who acquires an input supplier, while in the second case the control is on a downstream firm, for example a manufacturer who acquires a retailer.

Transaction Costs and Contract Incompleteness

Transaction Costs Economics

Under Transaction Costs Economics (TCE), different governance structures, i.e., internal organization versus market transactions, are evaluated with respect to the comparative costs of ensuring task completion.

TCE dates back to the pioneering article of Ronald Coase “The Nature of the Firm” in 1937 and has been enriched with further important contributions since the 1970s as the concept of transaction costs has been widened.

Coase’s starting point is that the distinguished mark of the firm is the suppression of the price mechanism; within a firm market transactions are eliminated and the allocation of resources is dependent on the entrepreneur-coordinator who directs production. The reason is that using the price mechanism has costs, which would be reduced or avoided by internalizing production. Coase focuses one type of costs in particular, i.e., long-term contracts, such as labor. Whenever a long-term contract for the supply of a service is preferred to several short period contracts – because in such a way some costs which are incurred in making each contract are avoided or because of the parties’ risk attitudes – the purchaser would not want to specify all details that the supplier is expected to do owing to uncertainty and difficulty of forecasting; in other words, he will not know which course of action he will want the supplier to take. The contract will then state only the limits of the supplier action and the direction of resources will depend on the buyer. Coase defines the firm as a system of relationship which emerges when the direction of resources rests on the entrepreneur.

As firms emerge because they allow to save market transaction costs, they will expand until the cost of organizing an extra transaction within the firm will equal the costs of carrying out that transaction in the market or the cost of organizing it in another firm. Firm size is then limited by the decreasing returns to the entrepreneur function, by the difficulty in allocating inputs in the uses they value most as the internalized transactions

increase, and by the increase in the input price due to the advantage of smaller firms.

If Coase explains the nature of the firm on the basis of the comparative transaction costs differences, Williamson (1971, 1975, 1985) develops further the concept of transaction specifying where those differences reside. His approach rests on two behavioral assumptions which break with the orthodox view of the maximizing man, i.e., the concept of bounded rationality and that of opportunistic behavior. The concept of bounded rationality was introduced by Herbert Simon (1955) who questioned the assumption of the “economic man” as being capable of making absolutely rational decisions; the capacity to acquire the relevant information for decision taking is limited, and so is the skill to compute and calculate among the different alternatives the course of action which leads to optimization.

In the presence of bounded rationality, economic agents can disclose misleading and distorting information in an opportunistic way; as a result, the assumption of self-interest seeking individual broadens as to encompass deceit. Under these circumstances, the neoclassical frictionless transaction paradigm may not always apply, in particular when bilateral interdependence between traders intrudes conflicts may arise which impair task completion. Alternative governance structures have then to be evaluated with respect to the comparative costs of ensuing – i.e., planning, adapting, and monitoring – task completion.

Williamson pinpoints three dimensions of transactions which are critical for TC evaluation: (i) asset specificity, (ii) frequency of transactions, and (iii) uncertainty. Asset specificity is the crucial condition for TC; it refers to the degree to which an asset can be reused in alternative ways and by alternative users without any loss of productive value. When investments are asset specific, their cost opportunity is much lower than they would be in the best alternative use or for an alternative user should the transaction not be completed; hence, the identity of the parties involved in the transaction becomes relevant as well as the continuity of the relationship. Asset specificity gives rise to bilateral interdependence.

Moreover, as the supply relation has to be adapted through time in response to disturbances, in evaluating the comparative advantage of internal organization versus market transactions, the ease of effecting intertemporal adaptations must be taken into account. On the one hand, market transactions allow production cost control more efficiently; on the other hand, in presence of asset specificity, they impede ease of adaptation. Vertical integration has the advantage of allowing the harmonization of interest, thus facilitating the sequential adaptive decision-making process.

Klein et al. (1978) focus on the possibility to appropriate quasi-rents from specialized asset – defined as the excess of asset's value over the value it would have in its next best use – by an opportunistic contractor as a reason for vertical integration. They refer in particular to the post contractual opportunistic behavior which will emerge once a specific investment is made no matter the enforceability of the contract. Even if a contract could unambiguously specify all relevant dimensions the threat of production delays should a litigation initiate may be an effective bargaining device.

Property Right Approach

Hart and Grossman (1986) observe that if TC approach helps understand integration when the costs of contracting between independent firms are higher, it fails to explain the limits to integration – if the right conditions are set and the owner of one of the firms becomes the employee of the other firm, any two independent owners can be better off integrating. Hart and Grossman thus sharpen Williamson TC-based arguments and define integration in terms of ownership of assets. Here again the problem rests in the difficulty in writing a contract which provides for all the payments and actions under every observable state of nature. Whenever it is costly for a party to write a contract which specifies the list of rights it desires over a list of another party's assets, it may be optimal for the first party to purchase all rights except those specifically mentioned in the contract; vertical integration is then the acquisition of the residual right of control over a supplier or a purchaser.

In this framework the costs of integration stem from the symmetry of control, i.e., when residual rights are purchased by one party, they are lost by the other. Symmetry of control creates distortions because of contractual incompleteness. One problem arises because the allocation of ownership rights changes the returns of the investment and by this way the level of investment. A controlling firm can use its residual right of control to increase the share of ex post surplus, and this induces the controlled firm to underinvest; as a result integration is optimal only if one firm decision is particularly relevant as compared to another.

Hart and Moore (1990) further specify the model by interpreting ownership right over an asset as the ability to exclude others from the use of this asset. Assuming asset-specific or person-specific productivity acquisition and incomplete contracts, the model predicts how changes in ownership affect employees and manager incentives; integration is then efficient – holdup is reduced – whenever the acquiring firm investment is important or the acquiring firm is an important trading partner or in the presence of asset complementarity.

Vertical Integration and the Competitive Environment

The rationale for vertical integration has been widely debated in the literature with respect to markets' competitive structure. On the one hand, integration can be seen as a way to achieve foreclosure or replicate otherwise banned price strategies such as price squeeze; from this perspective vertical integration may be a matter of concern for antitrust authorities. On the other hand, vertical integration can be explained as a response to “vertical externalities” (Tirole 1988); integrating firms may cope with efficiency losses due to variable proportions, double marginalization, and free riding problems, or, in the monopsony case, input rents extraction.

At the heart of those theories is the neoclassical firm assumption. In this setting integration is seen as an alternative to vertical agreements such as exclusive dealings, franchising, royalties, resale

price agreements, or other price strategies. Vertical integration is then at the extreme end of the spectrum; theories on the pros and cons of vertical integration from the standpoint of competition usually apply to vertical restraints.

Integration, Foreclosure, and Price Discrimination

Hart and Tirole (1990) trace the foreclosure effect of vertical integration to a commitment problem. Suppose an upstream firm supplying two downstream firms. The upstream firm could serve the downstream firms at a price equal to half the monopoly profit; however, with the absence of the capacity to commit – for example, through exclusive dealings – the upstream producer would have the incentive to renege and increase the quantity sold to one of the downward firms. In those circumstances, the downward firms would anticipate renegotiation and refuse to enter the contract unless the price charged by the upstream producer is lower; by vertically integrating with one of the downstream firms, the upstream producer would lose the incentive to supply the downstream rival and restore market power. Vertical integration may then entail market foreclosure.

Vertical integration has also been analyzed as an exclusionary strategy relying on the raising of rivals' costs. In this setting foreclosure occurs as a result of integration if nonintegrated downward competitors are not supplied with the inputs produced by the upward integrated firms or if the nonintegrated upstream competitor is preempted from selling to the downward integrated firm. For example (Salop and Scheffman 1983, 1987), a dominant firm facing a competitive fringe may integrate upward and – as long as the upstream firm has market power – raise the input price to its rivals. The final price to consumers will increase. The upstream profits will be sacrificed; however, price squeeze allows the dominant firm to extend profits disproportionately. Several objections have been raised as to the possibility of preemption by raising rival costs; once that integration takes place, the reduction in the input demand by the downward integrated firm may limit the capacity of the upward independent suppliers to

raise input prices; the integrated upward firms may find it unprofitable to refuse to deal with independent downstream firms; downward firms may find it profitable to integrate in turn. Some of these objections have been addressed in the literature; according to Ordover et al. (1990), foreclosure emerges in equilibrium whenever the downstream firm's revenue is increasing in the input price, as the profit squeeze for the downward independent firms is less than the increase in the upward unintegrated supplier profits.

Vertical integration can be also a way to practice implicit price discrimination when explicit price discrimination is banned. Suppose a monopolist – or a dominant firm – supplies two downstream firms with an input and the elasticity of the derived demand for that input is different for each firm; the monopolist would then maximize his profits by charging a higher price to the firm having a more rigid demand. If price discrimination was prohibited, the same effect could be achieved by integrating the elastic demand firm and increasing the input price to the independent firm. The result would be a price squeeze effect, since the integrated subsidiary would be able to lower price to final consumers. This argument was first set out by Stigler (1951) and then developed in various directions. Carlton and Perloff (1981) analyze vertical integration as a means to price discriminate in natural resource industries. Katz (1987) focuses on the threat to upward integration by a retail chain competing with local stores as a determinant of price discrimination; vertical integration emerges only if price discrimination is banned.

Integration and Vertical Externalities

One of the traditional defenses of vertical integration is the theory of variable proportions. A firm employing a monopolistically supplied input in variable proportions with other inputs produced by a competitive industry would substitute the monopolistically produced input by the latter, thus causing efficiency losses. The monopolistic supplier would then be able to reap profits by integrating downward and correcting the input mix. The theory was first set out by McKenzie (1951) and Vernon and Graham (1971). The

literature has debated the welfare implication of this analysis; on the one hand, vertical integration would solve the problem of inefficiency losses, thus increasing welfare; on the other hand, the higher monopolistic price for the product would cause countervailing effects. A major critique to this analysis is that integration is not a necessary condition for this result, which could be achieved by the monopolist through an appropriate pricing strategy.

Vertical integration may be welfare increasing when it avoids double marginalization. The model dates back to the pioneering article by Spengler (1950). When a firm with market power supplies an input to a downstream retailer also having market power, both the price charged by the upstream firm to the downstream retailer and the price charged by the downstream firm to final consumers will include a markup. As a result consumers will pay too high a price and final output will be “too much” restricted. If one markup could be eliminated, i.e., if retailer could price at marginal cost, final output would be increase so as joint profits. Integrating both firms can coordinate price strategies and increase both producers’ and consumers’ welfare. The avoidance of intermediate distortions presupposes market power both upstream and downstream; however, the same result is obtained if vertical integration occurs in a monopolistically competitive industry (Dixit 1983).

Vertical integration may also be a way to solve free riding problems when a manufacturer serves competing retailers providing services which are not appropriable by a single seller. The setting has been first analyzed by Telser (1960) with respect to resale price maintenance (RPM). Suppose that for the selling of a product some investment in selling activity is needed, for example, in order to inform about the technical characteristics of a product. If a seller makes this effort, the buyers can benefit from his activity but may purchase the product elsewhere where it is sold at a lower price; the possibility of free riding entails under provision of the selling activity from the retailers. In this setting the manufacturer who has an interest in that consumers be informed about the features

of the produce may prevent undercutting by the competing retailers by imposing a higher price to retailers (RPM) or integrating downward.

Incentives for upward integration may arise when a monopsonist faces an upstream competitive industry which produces an input at rising marginal cost (Perry 1978). In this model the incentive to integrate relies in the extraction of the suppliers’ rents; by acquiring suppliers one at a time, a monopsonist is able to expand the employment of the input produced by the integrated suppliers and to reduce his dependence from the unintegrated suppliers. In this process the price of the input decreases and so the independent suppliers’ rents, allowing the monopsonist to acquire upstream firms at a lower price. The monopsonist gains from this strategy are of two types: the internalization of the efficiency losses due to the underemployment of the input (efficiency effect) and the reduction of the rent component of the input cost (rent effect).

Stigler Market Equilibrium Model

Stigler’s (1951) theory of vertical integration focuses on the firm as a set of functions; firms’ decision to internalize certain function rather than acquire the corresponding services from other firms depends on the maturity of the industry, and the relation between maturity and integration is U shaped. Firms in young industries usually have to execute more tasks on their own, as they require new materials, overcome technical problems, design specialized equipment, and so on; at the early stage of industry maturity, vertical integration is then a likely outcome. However, as markets grow, some functions become sufficiently important to be provided by specialized firms, so that integration tends to decrease. Internalization takes place again in declining industries as surviving firms have to assure functions which are no longer carried out at a sufficient rate to support independent firms.

More recent contributions (Perry 1984; Green 1986) analyze vertical integration as a response to demand fluctuations in the presence of

competitive markets. When the intermediate markets are subjects to external fluctuations caused by the exogenous net supply of intermediate goods, on the one hand, firms can have incentives to integrate to avoid uncertainty and, on the other hand, they could enhance profits whenever they can respond to market fluctuations. The decision to integrate amplifies fluctuations in the intermediate markets which in turn affect the gains from integrating. Different equilibria may arise as a result of price flexibility.

Vertical Integration and Uncertainty

Other models concentrate on vertical integration in the presence of uncertainty. Arrow (1975) focuses on the uncertainty in the supply of the upstream good and on the consequent need for information by downstream firms; acquiring one or more upstream firms improves forecasts of the spot prices of the input and so the firm's ability to choose the level of capital.

According to Carlton (1979), vertical integration is a means of transferring risk from one sector to another. If the downstream firms' (the retailers) demand is random, a firm must make production decision before the demand can be observed; the derived demand which the upstream firm faces is also random. Because of the risk that the produced good goes unsold, the price of the input must exceed the marginal production cost and in this difference lies firm's incentive for forward integration; in addition integrating the firm copes with the risk of being rationed by the wholesaler, which would prevent sales. However, according to the model, vertical integration implies a lower level of expected utility because downstream firms are less efficient risk absorbers with respect to upstream firms, which causes higher total input costs in market structures with vertical integration. The lower efficiency in risk absorption in turn is due to the higher number of downstream than upstream firms; for this reason the probability that a unit of factor input will be used if it is held in the upstream firms is higher.

Cross-References

- [Property Rights: Limits and Enhancements](#)
- [Transaction Costs](#)

References

- Arrow KJ (1975) Vertical integration and communications. *Bell J Econ* 6(1):173–183
- Carlton DW (1979) Vertical integration in competitive market under uncertainty. *J Ind Econ* 27(3):189–209
- Carlton DW, Perloff JM (1981) Price discrimination, vertical integration and divestiture in natural resource markets. *Resour Energy* 3:1–11
- Coase R (1937) The nature of the firm. *Economica* 4:386–405
- Dixit A (1983) Vertical integration in a monopolistically competitive industry. *Int J Ind Organ* 1:63–87
- Green J (1986) Vertical integration and assurance of markets. In: Mathewson, Stiglitz (eds) *New developments in the analysis of market structure*. MIT Press, Cambridge, MA
- Grossman SJ, Hart O (1986) The costs and benefits of ownership: a theory of vertical and lateral integration. *J Polit Econ* 94:691–719
- Hart O, Moore J (1990) Property rights and the nature of the firm. *J Polit Econ* 98:1119–1158
- Hart O, Tirole J (1990) Vertical integration and market foreclosure. *Brook Pap Microecon*: 205–276
- Katz ML (1987) The welfare effects if third degree price discrimination in intermediate good markets. *Am Econ Rev* 77:154–167
- Klein B, Crawford RG, Alchian AA (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J Law Econ* 21:297–326
- McKenzie LW (1951) Ideal output and the interdependence of firms. *Econ J* 61:785–803
- Ordover JA, Saloner G, Salop SC (1990) Equilibrium vertical foreclosure. *Am Econ Rev* 80:127–142
- Perry MK (1978) Vertical integration: the monopsony case. *Am Econ Rev* 68:561–570
- Perry MK (1984) Vertical equilibrium in a competitive input market. *Int J Ind Organ* 2:159–170
- Salop SC, Scheffman DT (1983) Raising rivals' costs. *Am Econ Rev* 73:267–271
- Salop SC, Scheffman DT (1987) Cost-raising strategies. *J Ind Econ* 36:19–34.
- Simon HA (1955) A behavioral model of rational choice. *Q J Econ* 69(1):99–118
- Spengler JJ (1950) Vertical integration and antitrust policy. *J Polit Econ* 53:347–352
- Stigler GJ (1951) The division of labor is limited by the extent of the market. *J Polit Econ* 59:185–193
- Telser LG (1960) Why should manufacturers want fair trade? *J Law Econ* 3:86–105

- Tirole J (1988) *The theory of industrial organization*. MIT Press, Cambridge, MA
- Vernon J, Graham D (1971) Profitability of monopolization by vertical integration. *J Polit Econ* 79:924–925
- Williamson OE (1971) The vertical integration of production: market failure considerations. *Am Econ Rev* 61:112–123
- Williamson OE (1975) *Markets and hierarchies: analysis and antitrust implications*. Free Press, New York
- Williamson (OE) (1985) *The economic institutions of capitalism*. Free Press, New York

Violence, Conflict-Related

Alessandra Cassar¹, Pauline Grosjean² and Sam Whitt³

¹University of San Francisco, San Francisco, CA, USA

²School of Economics, University of New South Wales, Sydney, NSW, Australia

³High Point University, High Point, NC, USA

Definition

Actions and events occurring in the context of warfare and armed conflict that result in economic losses, physical, and/or psychological trauma. These actions and events may include, and are not limiting to, injury, loss of life, property theft and destruction, displacement, sexual assault, forced displacement or detention, torture, and participating in or witnessing traumatic actions committed against others. The perpetrators and/or recipients of such actions need not be formal institutional actors. Violence may be caused by an actor's willful intent and influence on others, or through negligence and failure to act to prevent trauma from taking place. Perpetrators of violence may also be adversely affected by actions they undertake or witness against others.

Literature

War and other forms of conflict-related violence have taken a profound toll on many people throughout the world. Between 1945 and 1999,

Fearon and Laitin (2003) estimate that over 3 million people died in 25 interstate wars and more than 16 million died in 122 civil wars in 73 countries. Although war and conflict-related violence have been examined extensively through various macro-level data collection efforts (Singer and Small 1994; Small and Singer 1982; Gleditsch et al. 2002; Collier and Hoeffler 2004; Lacina and Gleditsch 2005; Raleigh and Hegre 2005), evidence of conflict effects at the microlevel is currently limited (Blattman and Miguel 2010).

Our primary goal here is to survey recent contributions to the literature on possibly positive and negative legacies of violence. Given the importance of institutions and social norms for growth and development, the long-term consequences of war and conflict-related violence on a society's prospects for development are likely to depend on the social and institutional legacy of conflict. We will review here papers that rely on different sources of evidence, microlevel surveys, behavioral experiments, or macroeconomic datasets in order to study the social and political legacy of conflict. Table 1 lists recent studies on the relationship between violence and political and social behavior. We will review some of these studies in more detail. We find that, the findings from this literature are mixed. While a number of recent studies show that exposure to conflict intensifies certain positive prosocial elements within individuals and communities, others point to negative consequences of conflict on trust. We will attempt to shed light on these disparities by considering how individuals and communities are impacted by conflict. We find that, the effects of conflict may vary significantly depending on microlevel heterogeneity of conflict experiences.

We will not review studies of the effect of conflict on human capital outcomes, such as health or education, nor studies of the effect of conflict on physical capital and aggregate growth rates. Excellent reviews of this literature can be found elsewhere, namely, in Blattman and Miguel (2010). Instead, we focus on microlevel responses to violence. In particular, we consider the effects of conflict on prosocial preferences such as trust, because they have been found vital to solving cooperation and coordination problems and

Violence, Conflict-Related, Table 1 Heterogeneous effects of violence in political and social behavior reported in recent studies

Study authors	Year	Case/cases	Source of violence or trauma	Outcome and/or main effects	Methods	Results
Bauer et al.	2014	Republic of Georgia Sierra Leone	War	+	Experiment	Higher egalitarianism and parochialism among victimized children
Gilligan et al.	2013	Nepal	War	+	Experiment	Communities with greater exposure to violence exhibit greater levels of social capital
Voors et al.	2012	Burundi	War	+	Experiment	Individuals exposed to violence display more altruistic behavior toward their neighbors
Gneezy and Fessler	2012	Israel	War	+	Experiment	Violence increases cooperation within the group
Blattman and Annan	2010	Uganda	War	+	Survey	Formerly abducted are much more likely to engage in political life
Bellows and Miguel	2009	Sierra Leone	War	+	Survey	Violence leads to more engagement in the community, politics, and public good contribution
Blattman	2009	Uganda	War	+	Survey	Violence leads to more political participation
Bateson	2012	Large N	Crime	+	Survey	Violence leads to more participation in politics
Whitt and Wilson	2007, 2014	Bosnia	War	+	Experiment	Surprisingly high levels of out-group altruism after the war
Cassar et al.	2012	Tajikistan	War	—	Experiment	Victimization by violence undermines local trust
Jha and Wilkinson	2012	India	War	—	Macro	Communities with combat experience exhibit greater ethnic cleansing
Humphreys and Weinstein	2007	Sierra Leone	War	—	Survey	Ethnically mixed combatant groups more likely to abuse civilians
Becchetti et. al	2013	Kenya	Riots	—	Experiment	Violence leads to less trustworthiness
Beber et al.	2014	Sudan	Riots	—	Survey	Violence exposure leads to support for partition and opposing citizenship for out-groups
Nunn and Wantchekon	2011	Africa	Slave trade	—	Macro	Exposure to slave trade leads to lower trust in relatives and neighbors
Rohner et al.	2013	Uganda	War	—/+	Survey	Fighting has a negative effect on the economic situation in highly fractionalized counties, but has no effect in less fractionalized counties
Carter and Castillo	2005	Honduras	Natural disaster	+/—	Experiment	While negative shocks might promote cooperation, too large negative a shock appears to diminish it
Kesternich et al. 2014.	2014	Europe 13 countries	World War II	—	Survey	Negative effect of conflict on health and economic outcomes

(continued)

Violence, Conflict-Related, Table 1 (continued)

Study authors	Year	Case/cases	Source of violence or trauma	Outcome and/or main effects	Methods	Results
Grosjean	2014	Europe 35 countries	World War II	—	Survey	Negative legacy of conflict on political trust and perceived effectiveness of national institutions
Kim and Lee	2012	Korea	War	—	Survey	Children exposed to conflict become more risk averse as adults and more conservative politically

therefore crucial for economic and social development. Societal trust has been positively associated with growth and market development (e.g., Knack and Keefer 1997; Zak and Knack 2001; Henrich et al. 2010). Recent studies have shown how other-regarding preferences are critical for human cooperation in large groups (Bowles 2006; Boyd and Richerson 2005) and collective action (Bowles and Gintis 2006). The role of impersonal social trust in sustaining economic exchange is the object of an ever-growing literature. A prerequisite for the successful development of market economies is to depart from closed group interactions and to enlarge exchange to anonymous others (Fafchamps 2006; Algan and Cahuc 2010). In this regard, generalized trust appears as a keystone for successful market development. Generosity, a sense of fairness, and trustworthiness may also help sustain trade and cooperation in countries where institutional contract enforcement is weak by preventing contract violations. Even in countries with well-functioning institutions, trust may play an important role, given the incomplete nature of contracts. We now explore seemingly paradoxical positive and negative effects of conflict on prosocial preferences from recent studies in more detail.

On the Positive Effects of Conflict on Prosocial Behavior

Recent literature on the behavioral legacies of conflict offers surprisingly consistent evidence

of increased prosocial behavior after violence, with disconcertingly positive implications for the effects of war on social capital building. In particular, Bellows and Miguel (2009) find a significant increase in collective action among the individuals more affected by the war in Sierra Leone. Blattman (2009) reports higher voting and political action among former child soldiers in Uganda. Using lab-in-the-field experiments, Voors et al. (2012) find that individuals with exposure to greater levels of violence during the war display more altruistic behavior toward their neighbors, are more risk seeking, and have higher discount rates. Bauer et al. (2014) provide evidence of higher egalitarianism and parochialism among victimized children in the Republic of Georgia after the war with Russia, as well as among subjects victimized as children during the civil war in Sierra Leone. At the community level, Gilligan et al. (2011) find that communities with greater exposure to violence during the Maoist rebellion in Nepal exhibit more trust and higher levels of collective action. Whitt and Wilson (2007) and Whitt (2014) find surprisingly high levels of altruism toward ethnically defined out-groups among Bosniaks, Serbs, and Croats in postwar Bosnia, suggesting how prosocial norms can recalibrate quickly among former rivals and adversaries. Collectively, these studies imply that conflict may not have long-term damaging effects on prosocial norms. People can quickly recover or revert to prosocial norms once the fighting stops.

On the Negative Effects of Conflict on Prosocial Behavior

While the above research points to growing evidence of prosocial behavior after violence, some recent studies also provide cautionary evidence to the contrary. Becchetti et al. (2011) report lower trustworthiness among individuals most affected by violence in Kenya. Rohner et al. (2013) find detrimental effects of conflict on interethnic trust and trade in Uganda. The effect is particularly strong among ethnically divided communities. Nunn and Wantchekon (2011) also show that violence, and a history of violence going as far back as the slave trade in Africa, can still impact contemporary trust negatively and strongly. They find a remarkably robust negative legacy of the slave trade on general trust which they attribute to the destruction of social ties by interethnic slave raiding that took place hundreds of years ago. In 35 countries of Europe and Central Asia, Grosjean (2014) finds that victimization during World War II is, to this day, associated with lower perceived legitimacy and effectiveness of national institutions.

Using the case of Tajikistan, Cassar et al. (2013a) report the results of a series of behavioral experiments and individual surveys designed to investigate whether, more than 10 years after the end of the 1992–1997 Tajik civil war, there is any imprint on preferences and social norms that are thought to sustain the development of impersonal exchange. They find negative and persistent effects of violence on the norms that support impersonal exchange, in particular on trust within local communities. The magnitude of the effect is substantial. They find that victimization during the civil war is associated with a 40% average decrease in trust, as measured in a standard behavioral trust game, when respondents are matched to another individual from the same village. Survey-based evidence on actual behavior and stated preferences corroborates the experimental findings and points to natural consequence of lost trust: a reduced willingness to trade. Former victims are both less likely and less willing to participate in local

markets, especially when they do not have a personal connection with the trader they are dealing with. The results also indicate that experiences of victimization are associated with reinforced kinship-based norms of morality and behavior, at the expense of the rule of formal law.

However, although the net effect of conflict is negative, Cassar et al. (2013a) also find that the negative effects of violence depend on the extent to which infighting took place within local communities. Negative effects are more pronounced in regions where opposing groups intermix and where local allegiances were, and still are, split. This indicates that effects of conflict on local norms are mediated both by the specificity and the salience of wartime divides. This, in turn, may explain why the emerging literature on conflict and prosociality has found both positive and negative effects, as we now explore in more detail.

Making Sense of Conflicting Findings

Civil wars can come in many different forms. Some wars pitch socially and politically well-defined ingroups against equally cohesive out-groups. In other conflicts, even ostensibly ethnic conflicts, it is hard to identify who is ingroup and who is out-group, i.e., who is a friend and who is a foe. And in other cases, where cleavages are more based on ideology and beliefs than ascriptive characteristics, it may be nearly impossible to identify friend from foe using rudimentary visual or auditory cues. The Tajik civil conflict, in its vast majority, fits into this last category. It was characterized by localized intragroup conflict where people were unable to apply basic heuristics to identify friend from foe within their communities. As perpetrators of violence and authors of denunciations may still be close by, victims might be especially wary and guarded about whom they can trust – even years after the conflict has ended.

Cassar et al. (2013a) find indeed that the negative legacy of conflict in intra-community trust is most stringent in areas where the population was,

and still is, of intermixed political allegiance. They find that the association between conflict exposure and undermining of local trust, willingness to trade, and strength of kinship is only present in villages that are above the median in terms of political polarization but absent in villages that are below the median. They also contrast the effect in the two regional polar-opposite cases: the Pamir region where no community infighting occurred and the environs of the capital city Dushanbe, one of the most intermixed regions of the country. Consistently with the above interpretation, they find that the negative results are strongest in the Dushanbe region but absent in Pamir.

These results suggest that civil war may have particularly deleterious effects in regions where opposing groups intermixed and where allegiances are difficult to identify from basic observable characteristics. This last implication is consistent with recent findings by Rohner et al. (2013) that ethnic conflict in Uganda has hindered cohesion in ethnically divided districts but has had little effect in ethnically homogenous districts.

Another observation is that many studies of positive effects of conflict on social capital rest upon the Putnam (1995) interpretation of social capital based on incidence and intensity of group membership and participation in associations. Putnam argues that group membership (bonding social capital) fosters trust toward other groups (bridging social capital). Indeed, group membership and civic participation have been widely used in the literature as measures of social capital and positive development outcomes (for a recent review, see Guiso et al. 2010). However, social capital as group membership and association may also have negative consequences if identifying strongly with a group leads to the exclusion of outsiders (Bourdieu 1985; Portes 1998).

For example, on a positive note, Cassar et al. (2013a) find in postwar Tajikistan and Grosjean (2014) again in 35 European countries after World War II that conflict stimulates collective action and membership in social organizations (Bellows and Miguel 2009; Voors et al. 2012). However, they also shed light on the complex and dark nature of conflict-induced

social capital formation. While Grosjean (2014) reveals that victimization leads to collective action, it is of a type that appears to erode political trust. Cassar et al. (2013a) similarly find that civic engagement among victims of violence decreases trust, especially when victims of violence associate with religious groups.

In the case of Tajikistan, participation in religious groups may be perceived as a form of opposition to the secular government. In a companion paper, Cassar et al. (2013b) find that both war veterans and those who participated in fighting even after the 1997 peace agreement are significantly more active in mosque attendance and other measure of religiosity. They interpret this finding as evidence of the effect of conflict experience on fostering bonding rather than bridging social capital. These conclusions resonate with recent findings by Satyanath et al. (2013), which illustrate that a dense network of civic associations was instrumental in the downfall of democracy in interwar Germany.

Conclusion

Field research on the effects of violence on institutions and social norms is ongoing. Directions for future research include important questions about time horizons for institutional development and norm recovery after violence, the extent to which norms and institutions change when conflict is still active and ongoing, and estimating conflict effects through natural experiments and quasi-experimental designs. There is also an increasing need for replication of results from prior studies. To our knowledge, new field research projects are active or recently completed in conflict-ridden societies, for example, in Afghanistan, Bosnia, Congo, Kosovo, Pakistan, South Sudan, Somalia, Syria, Ukraine, and Yemen. With each year, the body of evidence from microlevel research into patterns of conflict effects is expanding and improving in terms of scope and sophistication in research design and linking empirical results to existing theories of violence and its heterogeneous consequences for growth and development.

At the same time, the conflict literature faces major identification problems. It is virtually impossible to find perfect randomization of treatments (such as exposure to victimization) in naturally occurring field settings. In the absence of such ideal conditions, a multitude of less-than-perfect observational studies pointing to similar conclusions are still better than nothing at all.

In conclusion, we believe that more clarification of ingroup/out-group identification problems could help explain why some conflict studies find dramatic rebound and recovery of social norms, while others observe lasting detrimental legacies of violence.

Cross-References

► Violence, Interpersonal

References

- Algan Y, Cahuc P (2010) Inherited trust and growth. *Am Econ Rev* 100(5):2060–2092
- Bauer M, Cassar A, Chytilová J, Henrich J (2014) War's enduring effects on the development of egalitarian motivations and in-group biases. *Psychol Sci* 25(1):47–57
- Becchetti L, Conzo P, Romeo A (2011) Violence and social capital: evidence of a microeconomic vicious circle. Working papers 197, ECINEQ, Society for the Study of Economic Inequality
- Blattman C (2009) From violence to voting: war and political participation in Uganda. *Am Polit Sci Rev* 103:231
- Bourdieu P (1985) The social space and the genesis of groups. *Soc Sci Inf* 24(2):195–220
- Bowles S (2006) Group competition, reproductive leveling, and the evolution of human altruism. *Science* 314:1569
- Bowles S, Herbert G (2006) Social preferences, homo economicus and zoon politikon. In: Robert E, Goodin RE, Charles T (eds) *The Oxford handbook of work of contextual political analysis*. Oxford University Press, New York
- Boyd R, Richerson PJ (2005) Solving the puzzle of human cooperation. In: Levinson S (ed) *Evolution and culture*. MIT Press, Cambridge, MA, pp 105–132
- Cassar A, Grosjean P, Whitt S (2013a) Legacies of violence: trust and market development. *J Econ Growth* 18(3):285–318
- Cassar A, Grosjean P, Whitt S (2013b) Social preferences of ex-combatants: survey and experimental evidence from postwar Tajikistan. In: Wärneryd K (ed) *The economics of conflict: theory and empirical evidence*. MIT Press (Forthcoming)
- Collier P, Hoeffler A (2004) Greed and grievance in civil war. *Oxford Econ Pap* 56(4):563–595
- Fafchamps M (2006) Development and social capital. *J Dev Stud* 42(7):1180–1198
- Fearon JD, Laitin DD (2003) Ethnicity, insurgency, and civil war. *Am Polit Sci Rev* 97(1):75–90
- Gilligan MJ, Pasquale BJ, Samii C (2011) Civil war and social capital: game-game evidence from Nepal. SSRN working paper number 1911969
- Gleditsch NP, Wallensteen P, Eriksson M, Sollenberg M, Strand H (2002) Armed conflict 1946–2001: a new dataset. *J Peace Res* 39(5):615–637
- Grosjean P (2014) Conflict and social and political preferences: evidence from World War II and civil conflict in 35 European countries, comparative economic studies, Mar 2014
- Guiso L, Sapienza P, Zingales L (2010) Civic capital as the missing link. NBER working papers 15845, National Bureau of Economic Research, Inc.
- Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, Bolyanatz A et al (2010) Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327(5972):1480–1484
- John B, Miguel E (2009) War and collective action in Sierra Leone. *J Public Econ* 93(11–12):1144–1157
- Kesternich I, Siflinger B, Smith JP, Winter J (2014) The effects of World War II on economic and health outcomes across Europe. *Rev Econ Stat* 96(1):103–118
- Kim Y Il, Lee J (2012) Long run impact of traumatic experience on attitudes toward risk: study of Korean war and its impact on risk aversion. IZA working paper
- Knack S, Keefer P (1997) Does social capital have an economic payoff? A cross-country investigation. *Q J Econ* 112(4):1251–1288
- Lacina B, Gleditsch NP (2005) Monitoring trends in global combat: a new dataset of battle deaths. *Eur J Popul* 21(2–3):145–166
- Nathan N, Wantchekon L (2011) The slave trade and the origins of mistrust in Africa. *Am Econ Rev* 101(7):3221–3252
- Portes A (1998) Social capital: its origins and applications in modern sociology. *Annu Rev Sociol* 24:1–24
- Putnam R (1995) Bowling alone: America's declining social capital. *J Democr*:65–78
- Raleigh C, Hegre H (2005) Introducing ACLED: an armed conflict location and event dataset. Paper presented to the conference on 'disaggregating the study of civil war and transnational violence', University of California Institute of Global Conflict and Cooperation, San Diego, 7–8 Mar
- Rohner D, Thoenig M, Zilibotti F (2013) Seeds of distrust: conflict in Uganda. *J Econ Growth* 18(3):217–252
- Satyanath S, Voigtlaender N, Voth H-J (2013) Bowling for fascism: social capital and the rise of the Nazi party in Weimar Germany, 1919–33. NBER working paper no. 19201
- Singer DJ, Small M (1994) Correlates of war project: international and civil war data, 1816–1992. Inter-

- University Consortium for Political and Social Research, Ann Arbor
- Small M, Singer JD (1982) *Resort to arms: international and civil war, 1816–1980*. Sage, Beverly Hills
- Voors M, Nillesen E, Verwimp P, Bulte E, Lensink R, van Soest D (2012) Does conflict affect preferences? Results from field experiments in Burundi. *Am Econ Rev* 102(2):941–964
- Whitt S (2014) Social norms in the aftermath of ethnic violence: ethnicity and fairness in non-costly decision making. *J Confl Resolut* 58(1):93–119
- Whitt S, Wilson RK (2007) The dictator game, fairness and ethnicity in postwar Bosnia. *Am J Polit Sci* 51(3):655–668
- Zak P, Knack S (2001) Trust and growth. *Econ J* 111:295–321

Further Reading

- Blattman C, Miguel E (2010) Civil war. *J Econ Lit* 48(1):3–57

Violence, Interpersonal

Pauline Grosjean

School of Economics, University of New South Wales, Sydney, NSW, Australia

Definition

Interpersonal violence is the intentional use of physical force or power, threatened or actual, against another person that results in a high likelihood of injury and/or death. In what follows, we will abstract from psychological forms of abuse and from sexual violence and focus on assault and homicide.

Interpersonal Violence: Roots, Remedies, and Consequences

In the *Leviathan*, published in 1651, Hobbes famously described the pervasiveness of interpersonal violence in the absence of a strong state. He wrote that “during the time men live without a common power to keep them all in awe, they are in that condition which is called war, and such a war as is of every man against every man.” He

also evoked some of its consequences and in particular the fact that “in such condition there is no place for industry.”

In what follows, I will discuss the following: (i) why agents resort to interpersonal violence, with a focus on the economic origins of interpersonal violence; (ii) how violence is tamed and, in particular, why interpersonal violence has declined so steadily since the emergence of strong states during the Middle Ages in the Western World; and (iii) some of the consequence of violence on economic activity. We will see, as was already discussed by Hobbes, that the origins, prevalence, and consequences of violence are intricately linked to the nature and quality of the state.

The Roots of Violence

Until recently, economists have essentially focused on the defense of property rights and its converse, expropriation, as a motive for violence. As described by Hobbes, if the state is absent or too weak to enforce property rights, agents have to take matters into their own hands. Violence by individuals creates a “semblance of property rights” to use the words of Skaperdas (1992) and becomes a substitute for formal law enforcement. At equilibrium – an equilibrium of fear – property rights are sustained by respective investments in power and by the threat to exercise this power.

A common prediction of these models focused on expropriation is that actual force is used in equilibrium only in the presence of bargaining inefficiency, as in Gonzalez (2010) or in the presence of incomplete information, as in Donohue and Levitt (1998) or in Aney (2012). Otherwise, if every agent knows who the strong party is, it is better to concede rather than fight, since fighting is costly.

Castillo et al. (2013) provide an empirical validation of the theoretical intuitions concerning the link between expropriation motives and violence. They show, in the context of the Mexican drug trade, that higher incentives to exert violence due to a sharp rise in the price of cocaine due to large seizures of cocaine in Colombia (Mexico’s main cocaine supplier) between 2006 and 2010 account for 21.2% and 46% of homicides and drug-related

homicides, respectively, in Mexico. However, in the context of more advanced economies with better formal law enforcement such as the United States, Levitt (1999) finds that violent crime is not responsive to increases in predation motives generated by widening inequality between rich and poor.

Other papers have recently introduced other motives for violence in addition to direct expropriation. Chassang and Padro-i-Miquel (2010) show how the presence of uncertainty introduces a new motive for violence: preemption. Players, while second-guessing each other's moves, may decide to attack in order to avoid suffering a debilitating surprise attack from an opponent who is expected to be aggressive. Reputation is another major motive for violence. This is formally illustrated by Silverman (2004) and Ghosh et al. (2013), who develop reputation-based theories of crime and violence in dynamic games of incomplete information. Silverman (2004) shows that in an environment with incomplete information, agents may want to behave violently in order to build a reputation for being tough and deter future attacks. This "culture of violence" equilibrium prevails if the cost of participating in crime is low enough, for example, if the state is weak or absent.

The reputation motive for violence is also stressed in sociopsychological explanations of violence. In a famous and debated hypothesis, Nisbett and Cohen (1996) argue that the Southern culture of honor is rooted in the original settlement of the backcountry of the United States by settlers from the fringes of Britain, the Scottish Highlanders, and the Ulster Scots-Irish, who were traditionally practicing pastoralist agriculture. Cultures of honor, which rely heavily on aggression and male honor, are common adaptations among populations that depend upon easily stolen herds. Grosjean (2014) stresses another element, in addition to pastoralism, which explains why eighteenth century Scottish Highlanders and Scots-Irish were more prone to interpersonal violence than other Western European settlers. They originated from areas where formal states were weak and where the concept of state-administered punishment for crime was foreign (O'Donnell

2005). Lawlessness, lack of political centralization, and violence had characterized the Anglo-Scot borderlands for much of the 250 years during which Scotland and England were in open conflict with one another (from 1296 to 1551). Ulster, the last Irish province to come under English domination, had been particularly ravaged by the Nine Years' War and left in a power vacuum by the Flight of the Earls in 1607. In the absence of third-party law enforcement, aggression and a willingness to kill can be essential to build a reputation for toughness and deter animal theft.

Sociopsychological and economic theories based on reputation differ from previous expropriation models in that they move away from the immediate costs and benefits of exerting violent acts into the concept of a "culture" of violence. Evolutionary anthropologists argue that rules of dispute resolution belong to the set of culturally transmitted norms of behavior, that is to say preset behaviors that save on the cost of developing new responses to changing environments (Paciotti and Richerson 2002; Richerson and Boyd 2005). The absence of formal law enforcement entails the prevalence of self-help justice, which is sustained by specific cultural traits – the culture of honor being a particular example.

Grosjean (2014) formally tests Nisbett and Cohen's (1996) culture of honor hypothesis. Combining contemporary homicide data with historical census data, she finds that Scot or Scots-Irish historical presence contributes to higher homicide and aggravated assaults in the United States, even to this day. Consistently with a culture of honor, the effect is specific to white offenders and to a type of homicide that aims at the defense of one's reputation: between acquaintances and perpetrated with a handgun, pistol, or, in an even more demonstrative way, a blunt object, such as hammer or club. The persistence of this violence over more than 200 years is taken as evidence of the existence of a culture of violence. The author finds direct evidence that such violent cultural traits are inheritable and were transmitted within families to subsequent generations of people of Scots-Irish ancestry.

However, Grosjean (2014) finds that the relationship between historical Scots-Irish presence and contemporary interpersonal violence is observed in the South only. She hypothesizes, and confirms, that the historical presence of Scot or Scots-Irish herders is associated with higher homicide only in areas where the quality of formal institutions was low. The interpretation is that a culture of violence exists but, more importantly, that it does not exist in a vacuum. In particular, persistence has vanished where formal institutions were strong. The following subsection describes why that might be the case.

Taming Violence

The premise of the argument made so far lies in the relationship between lawlessness, economic vulnerability, and interpersonal violence. Violence plays an essential role for the defense and enforcement of property rights in the absence of third-party enforcement. It follows that better third-party enforcement should unravel violence.

The development of formal, state-based institutions guarantees the security property rights and provides alternative means of dispute adjudication. It therefore lowers the returns to interpersonal violence. The development of the state also increases the costs of interpersonal violence, since the monopolization of violence by the state goes hand in hand with the penalization of interpersonal violence (Weber 1958). This provides the foundation for a negative relationship between state development and the prevalence of violence.

The fact that state development unravels not only violence but also cultures of violence is at the heart of Elias's "civilizing process" (Elias 1994). According to this theory, cultural norms reflect the social structure. As the state develops and monopolizes violence, violent instincts and the inclination to solve disputes with fists, sticks, blades, or guns are gradually placed under an increasingly strong social control.

Pinker (2011) provides abundant historical evidence consistent with the view that interpersonal violence has been steadily decreasing since the

emergence of strong states in Europe in the Middle Ages. Detailed empirical studies are provided by Grosjean (2014), Couttenier et al. (2014), and Restrepo (2014). We have already reviewed Grosjean's (2014) result that the Scots-Irish culture of honor persisted only in areas of the United States where formal institutions were weak. Restrepo (2014) shows that areas of Canada with a weaker monopoly of force by the state at the time of European settlement were historically more violent and still are more violent – despite the later consolidation of the Canadian state. Couttenier et al. (2014) exploit the fact that some mineral discoveries occurred sometime before and sometime after governmental territorial expansion in the West of the United States and compare these two distinct institutional contexts. They find that when discoveries predated the establishment of the state, they were associated with higher levels of interpersonal violence. These differences have persisted over time. Today, still, places with "early" historical mineral discoveries experience homicide rates 40% higher than the sample average, even though, obviously, the state is now fully consolidated in these places. In contrast, no effect is observed if discoveries occurred after the state was established. These differences hold even when comparing discoveries that occurred within short time intervals from one another, either just before or just after incorporation. These findings speak for the importance of the quality of formal institutions for determining whether violence takes root: even short-term variations in the institutional environment can generate large differences in the prevailing levels of interpersonal violence, and these differences persist over time.

The question of how the monopolization of violence by the state comes about is beyond the reach of this article, but interested readers can be referred to a recent empirical study by Sanchez de la Sierra (2014) on the determinants of state formation in the Democratic Republic of Congo.

When the state's monopoly violence is incomplete, individuals can take matters into their own hands (man against man as expressed by Hobbes),

but specialized groups can also emerge as a substitute for the state to provide property rights security. This is essentially Gambetta's (2003) theory of the emergence of the Sicilian Mafia. He argues that the Mafia arose to substitute for formal, state-based property rights enforcement after the collapse of the Bourbon Kingdom in the nineteenth century. This idea is investigated empirically in a recent paper by Buonanno et al. (2012). At the time the state collapsed, the price of sulfur – Sicily's most valuable export – was soaring due to rising demand on international markets. The Mafia provided alternative means of property rights security. Buonanno et al. (2012) show that the Mafia took root in areas that were richest in sulfur, around the time the Bourbon Kingdom collapse created a high demand for property rights protection by Sicilian landlords.

The Mafia is not the only type of organization that can provide alternative means of property rights security when the state is absent or weak. Drug cartels and militias are others. In that regard, Dell (2011) and Castillo et al. (2013) provide fascinating studies of the interplay between state policies and cartel behavior. Both study the consequences of the United States and Mexican Drug War on cartel behavior. As in previously cited studies of the Mafia, cartel activity is proxied by the number of homicides. Dell (2011) provides a careful analysis of the political economy game between local authorities and drug cartels in Mexico to explain the sharp rise in drug trade-related violence in Mexico since 2007, which has claimed 50,000 lives. She shows that the crackdown led by Mexico's conservative National Action Party (PAN) has had a direct and positive effect on the level of prevailing violence. Using a regression discontinuity framework between closely contested local elections, in which it can be assumed that who wins office – and hence the intensity of repression – is randomly assigned, she shows that PAN's victories and the associated repression are systematically associated with a subsequent rise in homicide. She explains this phenomenon by the fact that the state's repressive policies weaken the position of the dominant drug

cartel and open the way for a war for dominance between cartels.

Angrist and Kugler (2008) document that exogenous increases in coca prices increase violence in rural districts in Colombia because combatant groups fight over the additional rents. Similarly, Castillo et al. (2013) show that violence exerted by Mexican cartels increases when the price of cocaine rises. Interestingly they find that violence is highest in locations contested by several cartels, that is to say areas where each cartel is furthest away from enjoying a monopoly of violence, a result that resonates with Dell's (2011) interpretation.

Violence and organized crime are commonly perceived as a severe obstacle to the economic development of several regions around the world, for instance, Latin American countries such as Mexico and Colombia or former republics of the Soviet Union. The next section reviews some of the negative economic consequences of violence and organized crime.

The Consequences of Violence

Besides the immediate costs of violence in terms of loss of lives, physical injuries, and psychological trauma, violence, or the threat of violence, also disrupts economic activity. We will review several studies that try to estimate the cost of violence. As we will see, this is an arduous task because it is difficult to separate the effect of violence itself from that of what gives rise to violence in the first place, that is to say insecure property rights and a weak rule of law. Indeed, several studies document the direct negative effects of institutional quality on economic growth and development (see the seminal study by Acemoglu et al. 2001). Making matters even more complicated, as we will see, is that violence itself can cause institutional quality to deteriorate even further.

The most immediate costs of violence consist obviously in loss of life and other bodily or psychological damages. Besides a few studies that try to estimate the value of statistical life – with sometimes widely varying estimates (see Leon and Miguel 2013) – economists have mostly focused

on the costs in terms of economic growth and development.

The first underlying mechanism through which violence negatively affects economic growth and development is expressed by Hobbes in the opening quote: violence or its threat leaves little place for “industry” because it makes the fruit of labor uncertain and hence reduces incentives to invest.

BenYishay and Pearlman (2013) estimate the impact of homicide-related deaths during the drug war in Mexico on the labor supply of the general population. They find that the increase in homicides has reduced average hours worked by 1–2%. They conclude that the fear of violence can lead to behavioral changes that lower economic activity.

Even in the context of more advanced economies, the direct economic costs of violence or their indirect costs, through several behavioral changes in the labor market, are substantial. In the United States, Hamermesh (1999) finds that higher rates of homicide explain declines in evening and early morning work and that these shifts resulted in losses on the order of four to ten billion dollars. Pinotti (2011a) estimates the impact of the presence of organized crime in Southern Italy on GDP. Applying a synthetic control approach, he compares Mafia-stricken regions to a weighted average of other regions that are otherwise similar in terms of industrial structure and outcomes prior to the rise of organized crime in the 1970s. He finds that the presence of the Mafia has lowered GDP per capita by 16% in Mafia-stricken regions.

A related question deals with the distribution of these costs across different segments of society. The few studies dealing with this issue find that it is the poor who bear the bulk of the costs of violence. In Mexico, BenYishay and Pearlman (2013) find that the negative impacts of violence on hours worked are larger for the self-employed who work from home, generally small-scale artisans and shopkeepers. Still in Mexico, Ajzenman et al. (2014) find that the impact of violence on housing prices is borne entirely by the poorer sectors of the population. They find that an increase in homicides equivalent to one standard

deviation leads to a 3% decrease in the price of low-income housing.

The decrease in GDP in Mafia-stricken regions of Southern Italy documented in Pinotti (2011a) is primarily driven by a contraction in private investment. While in principle, public investment could substitute to foregone private investment, Pinotti (2011) shows that this process is incomplete, what results in a net aggregate deadweight loss. First, he finds that an increase in public investment has not fully compensated the decrease in private investment. Second, this public investment was relatively inefficient and in fact further distorted by the presence of the Mafia.

This second result is explored in more details in a follow-up paper by Pinotti (2011b), which shows that Mafia presence has a direct negative effect on the quality of institutions and more precisely on the quality of politicians. The theoretical intuition is found in Dal Bo et al. (2006) who argue that the presence of criminal organizations deters high-quality politicians from running for office, as they anticipate that their political action might be constrained or distorted by criminal organizations and that their life could be endangered if they were not ready to compromise with criminal organizations. As a result, only dishonest politicians who are ready to bend down to the wills of organized criminal groups are left to run for office.

Pinotti (2011b) finds empirical support for this prediction. Politicians in Southern Italy who are appointed during more violent elections, in which the Mafia is presumably more active, exhibit on average a higher probability of being subsequently involved in scandals. They are also much less likely to hold a college degree.

Couttenier et al. (2014) document a related “curse” of homicide-related violence on the quality of judicial institutions in the historical context of the United States. They find not only that mineral discoveries that occurred before the consolidation of the state are associated with more violence but also with a lower quality of (subsequent) judicial institutions. Using a historical panel dataset of judicial quality indicators at the state level, they find that mineral discoveries are associated, yearly, with a reduction by

between 2% and 5% in the quality of judicial institutions.

The effect of institutional quality on growth and development has attracted a large literature, since the seminal studies by Engerman and Sokoloff (1997) and Acemoglu et al. (2001). More recently, scholars have also focused on the positive role of social norms of cooperation and trust, and social capital in general, for economic prosperity (see Knack and Keefer 1997). In this dimension also, the consequence of violence is dire. For example, Nunn and Wantchekon (2011) show that ethnic groups that were most severely affected by the violence associated with the slave trade in Africa still display lower levels of trust and have lower social capital, to this day.

These destructive effects of violence on formal and informal institutions provide additional channels through which violence negatively affects economic growth and development.

Conclusion

The literature reviewed here describes interpersonal violence as the outcome of a cultural-institutional arrangement, in which violence, or its threat, is a response to the need to protect and enforce property rights when formal institutions are weak or absent. Even though improvements in third-party enforcement could unravel this unfortunate equilibrium, in practice this is very difficult to achieve. This is because improvements in formal institutions are hard to come but also because the violence itself causes the quality of political and judicial institutions to worsen even further, thereby feeding the initial vicious circle.

On a more positive side, a couple of papers find that it is possible to break out of this vicious circle. Grosjean (2014) finds that a culture of violence by eighteenth century Scots-Irish settlers was not transmitted to further generations in areas where settlers faced strong formal institutions. Voigtländer and Voth (2012) also find that anti-Semitic attitudes disappeared in the most active trading regions of Germany where the cost of discriminating against outsiders was high.

Cross-References

► Violence, Conflict-Related

References

- Acemoglu D, Samuel J, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. *Am Econ Rev* 91:1369–1401
- Ajzenman N, Galiani S, Seira E (2014) On the distributive costs of drug-related homicides. Center for Global Development working paper 364
- Aney MS (2012) Conflict with quitting rights: a mechanism design approach. Working papers 18–2012, Singapore Management University, School of Economics
- Angrist JD, Kugler AD (2008) Rural windfall or a new resource curse? Coca, income, and civil conflict in Colombia. *Rev Econ Stat* 90:191–215
- BenYishay A, Pearlman S (2013) Homicide and work: the impact of Mexico's Drug War on labor market participation. Mimeo, University of New South Wales
- Buonanno P, Durante R, Prarolo G, Vanin P (2012) Poor institutions, rich mines: resource curse and the origins of the Sicilian Mafia. Working papers wp844, Dipartimento Scienze Economiche, Università di Bologna
- Castillo JC, Mejia D, Restrepo P (2013) Scarcity without Leviathan: the violent effects of cocaine supply shortages in the Mexican drug war. Center for Global Development working paper 356
- Chassang S, Padro-i-Miquel G (2010) Conflict and deterrence under strategic risk. *Q J Econ* 125(4): 1821–1858
- Couttenier M, Grosjean P, Sangnier M (2014) The wild west is wild: the homicide resource curse, no 2014–2012, discussion papers, School of Economics, The University of New South Wales
- Dal Bo E, Dal Bo P, Di Tella R (2006) Plata o Plomo?: Bribe and punishment in a theory of political influence. *Am Polit Sci Rev* 100(1):41–53
- Dell M (2011) Trafficking networks and the Mexican Drug War. Mimeo, Harvard
- Donohue JJ, Levitt SD (1998) Guns, violence, and the efficiency of illegal markets. *Am Econ Rev* 88(2): 463–467
- Elias N (1994) The civilizing process. Blackwell, Oxford/Cambridge
- Engerman SL, Sokoloff KL (1997) Factor endowments, institutions, and differential paths of growth among new world economies: a view from economic historians of the United States. In: Harber S (ed) *How Latin America fell behind*. Stanford University Press, Stanford, pp 260–304
- Gambetta D (1993) *The Sicilian Mafia: the business of private protection*. Harvard University Press, Cambridge, MA

- Ghosh S, Gratton G, Shen C (2013) Terrorism: a model of intimidation. Mimeo, University of New South Wales
- Gonzalez FM (2010) The use of coercion in society: insecure property rights, conflict and economic backwardness. In: Garfinkel MR, Skaperdas S (eds) Oxford handbook of the economics of peace and conflict. Oxford University Press, New York
- Grosjean P (2014) A history of violence: the culture of honor and homicide in the United States South. *J Eur Econ Assoc*, 12(5):1285–1316
- Hamermesh D (1999) Crime and the timing of work. *J Urban Econ* 45:311–330
- Hobbes T (1651) Of man, being the first part of Leviathan. The Harvard Classics. 1909–14. Chapter XIII
- Knack S, Keefer P (1997) Does social capital have an economic payoff? A cross-country investigation. *Q J Econ* 112(4):1251–1288
- Levitt S (1999) The changing relationship between income and crime victimization. *Fed Reserve Bank N Y Econ Policy Rev* 5(3):87–98
- Nisbett R, Cohen D (1996) Culture of honor: the psychology of violence in the south. Westview Press, New York
- Nunn N, Wantchekon L (2011) The slave trade and the origins of Mistrust in Africa. *Am Econ Rev* 101(7):3221–3252
- O'Donnell, Ian (2005) Lethal violence in Ireland 1841–2003. *British journal of criminology*
- Paciotti BM, Richerson PJ (2002) Incorporating a theory of cultural evolution into explanations of male dispute-related violence. Mimeo, UC Davis
- Pinker S (2011) The better angels of our nature. Viking, New York
- Pinotti P (2011a) The economic consequences of organized crime: evidence from Southern Italy. Mimeo, Università Bocconi
- Pinotti P (2011b) The economics of organized crime: lessons from the Italian Mafia. Mimeo, Università Bocconi
- Restrepo P (2014) The mounties and the origins of peace in the Canadian prairies. Mimeo, MIT
- Richerson PJ, Boyd R (2005) Not by genes alone. How culture transformed human evolution. The University of Chicago Press, Chicago
- Sanchez de la Sierra R (2014) On the origin of states: stationary bandits and taxation in Eastern Congo. Mimeo, Columbia
- Silverman D (2004) Street crime and street culture. *Int Econ Rev* 45(3):761–786
- Skaperdas S (1992) Cooperation, conflict, and power in the absence of property rights. *Am Econ Rev* 82(4):720–739
- Voigtländer N, Voth J (2012) Persecution perpetuated: medieval origins of anti-semitic violence in Nazi Germany. *Q J Econ*, 127(3):1339–1392
- Weber M (1958) The three types of legitimate rule, vol IV. Publications in Society and Institutions, Berkeley, p 1

Further Reading

- Sanchez de la Sierra R (2014) On the origin of states: stationary bandits and taxation in Eastern Congo. Mimeo, Columbia. https://raulsanchezdelasierra.files.wordpress.com/2014/06/paper_1_20140616.pdf

Voting Power Indices

Manfred J. Holler

Center of Conflict Resolution (CCR), University of Hamburg, Institute of SocioEconomics (ISE), Hamburg, Germany

Abstract

This entry introduces Shapley-Shubik index, Banzhaf index, Deegan-Packel index and Public Good Index. It discusses the properties of these measures of a priori voting power focusing on monotonicity. Applications are listed.

The Indices

On the occasion of the 1787 Constitutional Convention held in Philadelphia, Luther Martin, a Maryland delegate, complained “that the voting power of the large states in the US House of Representatives would be disproportionately too large compared to that of the small states, assuming that the representatives of each of the 13 states would always vote as a bloc” (Felsenthal and Machover 2005, p. 486). He observed that the seat shares are a pure proxy of the a priori voting power, i.e., the voting power that abstracts from political ideologies, personal preferences, geographical and cultural closeness, etc. and considers seats shares and the decision rule (e.g., a majority quota q), only. For an illustration, let us assume $q = 51$ and the seat distribution is $w = (48, 46, 6)$. Can we conclude that the first bloc of voters, controlling 48 seats, is 8 times as powerful as the third bloc controlling 6? In fact, we may ask whether it is more powerful at all. Voting power indices try to answer to such questions, discussing voting situations in the form of a weighted voting game (or, more general, in the

form of “simple games”) that assigns to a winning coalition a value of 1 and to a losing coalition a value of 0. The blocs of voters labeled players and their vote shares are identical to the weights.

Luther Martin measured the a priori voting power of state i by the number of possible minimum winning coalitions (MWC) that have i as a member. A coalition S is a MWC if S is losing if one of its members is leaving this coalition. As a consequence, all members of a MWC are critical to its winning, i.e., are “swing players.” This measure has been reintroduced by Holler (1982) under the label of Public Good Index (PGI) and axiomatized in Holler and Packel (1983) without being aware of Luther Martin’s proposition. Already before, Deegan and Packel (1979) have introduced a hybrid form based on MWCs: the Deegan-Packel index (DP). Their measure divides the benefits of winning by the number of players in the particular MWC. This implies that it treats the value of a coalition as a private good.

There is another case of reintroduction. In 1946, Lionel Penrose proposed a power measure that, to determine the a priori voting power of i , counted all coalitions S which had i as a swing player. Therefore, to contribute to the power value of S , i had to be critical for the winning of S , but S can have surplus voters, i.e., voters that have no impact on the winning or losing of S . This measure became popular as Banzhaf index (Bz) only when Banzhaf (1965) reintroduced it, and the measure was applied to the design of real world juries. For the latter, see Grofman and Scarrow (1979).

The favorite power measure for many game theorists, especially if they have some mathematical inclination, is the Shapley-Shubik index (SS) which applies the Shapley value (Shapley 1953), a solution concept for cooperative games, to situations of weighted voting. Shapley and Shubik (1954) is the corresponding paper. A straightforward calculation is to list all permutations (i.e., orderings) of voting blocs and, starting from the left side, give the bloc i that turns a losing coalition into a winning coalition, called pivot, the value 1. The values get normalized such that the values of all players add up to 1. Note that there are $n!$ permutations if there are n

players. However, there is a rather sophisticated formula that can be used as a shortcut. The formula is based on answering the question: What is the probability that a specific pivot i is in the ordering positioned exactly where it is?

One of the beauty of SS is that power weights can directly be interpreted as probabilities to affect the outcome – this relates to Max Weber’s notion of power “as the chance of a man or of a number of men to realize their own will in a communal action even against the resistance of others who are participating in the action” (Weber, 1948[1924], p.180).

The Right Index

There are at least another dozen of voting power indices discussed in the literature. This raises the question which is the right index. One way to answer this question is to look into the underlying stories and accept that alternative measures are meant for alternative issues. For instance, Bz and PGI are meant to measure the influence of voters, while SS and DP imply a sharing. However, the formal analysis is not conclusive. Alternatively, the measures are discussed with respect to their properties. Of course, the underlying axioms give some insights. Yet, more popular arguments refer to paradoxes and, more specifically, failures in satisfying monotonicity. PGI and DP do not satisfy local monotonicity (LM), i.e., it can happen that a player i with a larger vote share than player j gets assigned a smaller power value than j . For instance, given a vote distribution $w^\circ = (35, 20, 15, 15, 15)$ and a majority quota 51, the PGI values are $h(v^\circ) = (4/15, 2/15, 3/15, 3/15, 3/15)$. The player with weight 20 shows a lower value than the players with weights 15. Such a reversal, indicating a violation of LM, cannot happen if we apply Bz or SS as long as there are no a priori unions considered and coalitions and permutations, respectively, enter the calculation with equal weights (See Alonso-Mejide et al. (2009)). In a recent publication, Freixas and Kurz (2016) suggest a class of new indices that result out of convex combinations of PGI and Bz index such that LM is satisfied.

Both SS and Bz (as well as PGI and DP) suffer of the paradox of redistribution: If we redistribute the vote shares such that the vote share of player i increases, it can happen that i 's power value decreases. A classical discussion of paradoxes we find in Brams (1975). However, one might wonder whether it makes sense to compare power indices to vote distributions, in order to evaluate the properties of the indices, when Luther Martin observed already that seat shares are a pure proxy of (a priori) voting power. Under the assumption that the election outcome is represented by the seat distribution, which is not a trivial to achieve (see De Meur and Gassner 1987; Gambarelli and Palestini 2013), the seat distribution could serve as norm that should be equalized by voting power. Equalization can be achieved (or at least approximated) by a reshuffling of seats or by the choice of decision rule. Berg and Holler (1986) and Holler (1985) proposed a randomization over alternative decision rules to equalize voting power and seat distribution (See Turnovec 2011). Of course, with the strict correspondence of power and seat shares, the violation of LM and the paradox of redistribution is avoided. However, this does not answer which index to choose.

Applications

Earlier applications of voting power indices focused on both the US legislation – characterized by the interrelationship of Senate, Congress, and President – and the UN Security Council (see, e.g., Shapley and Shubik 1954). Over the last thirty years, however, numerous articles have been published on the power distribution in EU political institutions, i.e., Council of Ministers and the European Parliament, and their interplay (although with the Commission), e.g., in the co-decision procedure. Each of the two collection of papers Holler and Owen (2001) and Holler and Nurmi (2013) contains a section on “Voting Power in the European Union” “presenting five, respectively, six highly relevant articles. (See list of Further Reading below.) Moreover, there are the two volumes of *Homo Oeconomicus*, in honor of Hannu Nurmi

(Holler and Widgrén 2009; Holler et al. 2011), which contain a number of contributions on this subject. In addition, I would like to recommend Napel and Widgrén (2011) which analyzes the voting power in the EU Council of Ministers for the consultation procedure and compares the results with the outcome of a noncooperative (strategic) game designed accordingly. For the latter, (political) preferences have to be assumed, and thus the a priori of the power indices does no longer apply. However, in the given study, power indices delivered a good approximation of the “relative strategic power” of the noncooperative analysis.

There is also voting in bureaucracy, at university, and in business. Gambarelli and Pesce (2004) observe that the application of power indices allows the investor to define his power of control on the basis of the number of shares owned (See also Gambarelli and Owen 2002).

Cross-References

► [Power Indices](#)

References

- Alonso-Mejide JM, Bowles C, Holler MJ, Napel S (2009) Monotonicity of power in games with a priori unions. *Theory and Decision* 66:17–37
- Banzhaf JF (1965) Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review* 19: 317–343
- Berg S, Holler MJ (1986) Randomized decision rules in voting games: a model of strict proportional power. *Quality and Quantity* 20:419–429
- Brams SJ (1975) *Game theory and politics*. Free Press, New York
- Deegan J Jr, Packel EW (1979) A new index of power for simple n-person games. *International Journal of Game Theory* 7:113–123
- De Meur G, Gassner M (1987) Problems of equity in multiparty representational systems: What we want and what we can't. In: Holler MJ (ed) *The Logic of multiparty systems*. Kluwer, Dordrecht/Boston/Lancaster, pp 391–404
- Felsenthal DS, Machover M (2005) Voting power measurement: a story of misreinvention. *Social Choice and Welfare* 25:485–506
- Freixas J, Kurz S (2016) The cost of getting local monotonicity. *Eur J Oper Res* 251:600–612

- Gambarelli G, Owen G (2002) Power in political and business structures. In: Holler MJ, Kliemt H, Schmidtchen D, Streit ME (eds) Power and fairness (Jahrbuch für Neue Politische Ökonomie 20). Mohr Siebeck, Tübingen, pp 57–68
- Gambarelli G, Pesce S (2004) Takeover prices and portfolio theory. In: Gambarelli G (ed) Essays on cooperative games: in honor of Guillermo Owen. Special issue of Theory and Decision. Kluwer Academic Publishers, Dordrecht, vol 36. pp 193–203
- Gambarelli G, Palestini A (2013) Minimax multi-district apportionments. In: Holler MJ, Nurmi H (eds) Power, voting and voting power: 30 years after. Springer, Heidelberg, pp 169–186
- Grofman B, Scarrow H (1979) Iannucci and its aftermath: the application of the Banzhaf index to weighted voting in the state of New York. In: Brams SJ, Schotter A, Schwödiauer G (eds) Applied game theory. Physica, Würzburg-Wien
- Holler MJ (1982) Forming coalitions and measuring voting power. Political Studies 30:262–271
- Holler MJ (1985) Strict proportional power in voting bodies. Theory and Decision 19:249–258
- Holler MJ, Widgrén M (eds) (2009) Essays in honor of Hannu Nurmi: Volume I (Homo Oeconomicus 26). Accedo, Munich
- Holler MJ, Nohn A, Vartiainen H (eds) (2011) Essays in honor of Hannu Nurmi: Volume II (Homo Oeconomicus 28). Accedo, Munich
- Holler MJ, Packel EW (1983) Power, luck and the right index. Zeitschrift für Nationalökonomie (Journal of Economics) 43:21–29
- Napel S, Widgrén M (2011) Strategic versus non-strategic voting power in the EU Council of Ministers: the consultation procedure. Social Choice and Welfare 37:511–541
- Shapley LS (1953) A value for n-person cooperative games. In: Kuhn H, Tucker AW (eds) Contributions to the theory of games, vol 28, Annals of mathematical studies. Princeton University Press, Princeton, pp 307–17
- Shapley LS, Shubik M (1954) A method for evaluating the distribution of power in a committee system. American political Science Review 48:787–792
- Turnovec F (2011) Strict proportional power and optimal quota. Homo Oeconomicus 27:463–479
- Weber M (1948[1924]) Class, status and party. In: Gerth HH, C. Wright Mills CW (eds), Essays from Max Weber, Routledge/Kegan Paul, London

Further Reading

- Felsenthal D, Machover M (1998) The measurement of voting power. Theory and practice, problems and paradoxes. Edward Elgar, Cheltenham
- Holler MJ, Owen G (eds) (2001) Power indices and coalition formation. Kluwer, Boston/Dordrecht/London
- Holler MJ, Nurmi H (eds) (2013) Power, voting and voting power: 30 years after. Springer, Heidelberg

Way of Trade, The

► [Lex Mercatoria](#)

Wealth Maximization: Efficiency and Equality Considerations

Yew-Kwang Ng
Department of Economics, Nanyang
Technological University, Singapore, Singapore
School of Economics, Fudan University,
Shanghai, China

Abstract

Posner's proposal for wealth maximization in judiciary decisions has not been widely accepted, but the influence of the economic analysis of law it propelled has increased tremendously. In the face of criticism, Posner himself has retreated into a pragmatism with wealth maximization not the only principle used. This leads to a lack of both moral justification and consistency. This entry shows that, if not used as the ultimate objective, but as an instrument for welfare maximization, wealth maximization is much more acceptable, if it is also supplemented by appropriate redistribution in the general equality policy. This is especially so as efficiency supremacy in specific

issues including the judiciary (close to wealth maximization) would then make every income group better off. This is so despite the presence of disincentive effects in general redistribution. The separation of the judiciary and the legislative may then also be justified.

Introduction

Around four decades ago, eminent judge Posner (1973, 1979) pushed the interest in law and economics to a new height by proposing the use of wealth maximization in the judiciary. Posner proposed this principle both for descriptive/scientific analysis and proscriptive/normative guidance. In his own words, it “provides a good explanation of many common law rules and principles . . . [and also is] an ethically attractive guide to social policy, especially in the common law context” (Posner 1980a, pp. 243–244). The proposal led to intense discussion both on descriptive/scientific accuracy and the proscriptive/normative acceptability in the 1980s and after (e.g., Dworkin 1980; Kronman 1980; Veljanovski 1981; Mathis 2009; Campbell 2012). Recent interest and controversies in the issue continue (e.g., Driesen and Malloy 2015; Fernandes 2015; Hatzis and Mercuro 2015; Pino-Emhart 2016; Posner 2015). Despite the widespread nonacceptance of Posner's proposal especially at the normative level, the influence of the economic analysis of law has grown spectacularly (e.g., see Hatzis and

Mercuro 2015, Preface). Under severe criticisms, Posner himself has, since as early as the mid-1980s, retreated from wealth maximization as the single principle, into a pragmatism of admitting different considerations (“wealth maximization as an incomplete guide,” Posner 1985, p. 101; see also Mathis 2009, ch. 8), if not into an “anti-theoretical populist” position (Fernandes 2015, p. 210). This pragmatism is conceptually unsatisfactory as it does not provide some appropriate tradeoffs between the different considerations. In my view, some ultimate guide such as utilitarianism or its generalization (welfarism, which is utilitarianism without necessarily the maximization of the unweighted sum of net welfare) is desirable, at least at the more ultimate level of the long-term choice of what rules/principle/rights to accept and impose.

The present entry shows that, if “wealth maximization” is used not as the ultimate objective, but as a qualified instrument for some intrinsic objective (e.g., welfare maximization), it will be much more acceptable. Though Posner himself has also moved towards the instrumentality of wealth maximization (e.g., see Mathis 2009, Sub-section 8.4.4.3), he has provided neither the ultimate objective the instrument is clearly for, nor an adequate justification. Posner appears to be both rejecting and citing the support of utilitarianism. His criticism of utilitarianism is valid only against a very naïve version that ignores indirect effects. Taking account of these effects, utilitarianism is consistent with most widely accepted individual rights and constitutional provisions.

However, wealth maximization itself is not sufficient for welfare maximization. It has at least to be supplemented by the promotion of equality. Nevertheless, from the principle of treating a dollar as a dollar (efficiency supremacy) **in specific issues** (Ng 1984), the role separation between the legislative and the judiciary, as mentioned by Posner (1990, p. 101f.), may be justified. Posner justifies the separation on the ground of the lack of “flexible instruments for redistribution” by courts; “Courts can do very little to affect the distribution of wealth in a society” (Posner 1984, p. 132). Here we go further, arguing for refraining from redistribution even if the judiciary

is capable in doing so, at least on specific issues including judiciary decisions, leaving redistribution to be done in the general equality promotion policy. The focus of the judiciary on efficiency then comes close if not exactly coinciding with Posner’s wealth maximization. Despite the prospering of the economic analysis of law in the academia, “Law and Economics seems to be dying in tort law, at least as a normative principle” (Pino-Emhart 2016, concluding paragraph). The acceptance of our argument here may help to revive law and economics in this respect. Put it differently, the critics of Posner may have won the battle in this front, this entry tries to make Posner winning the war, at least in terms of wealth maximization for welfare maximization.

As pointed out by Mathis (2009, Section 8.6), Posner himself had retreated into a pragmatism of “rapprochement towards utilitarianism,” with “division of labour between the judiciary and the legislature,” and effectively using wealth maximization as a nonexclusive instrument. Thus, in a sense, the present entry is much in line with Posner’s later position. However, it provides a more consistent defence from the level of moral philosophy (intrinsic values) to that of practical policy, with deeper analysis, especially regarding the issue of equality or distribution.

To focus on our main points within an entry, we largely ignore some complications like the difference between the willingness to pay and the willingness to accept (or between offer and asking prices), loss aversion, endowment effects, etc. For simplification, we will also largely use an a-temporal analysis except when intertemporal issues are directly discussed, thus largely ignoring such issues like the difference between stocks and flows. Posner’s failure to adequately distinguish between ex ante compensation and hypothetical compensation (Posner 1980b), the “naturalistic fallacy” (Mathis 2009, p. 164) of violating Hume’s law, and other methodological ambiguities that have been discussed by others (e.g., Mathis 2009, ch. 8; Campbell 2012) will also largely not be covered here. On the issue of compensation, the actual loss of a lottery ticket buyer may be said to be ex-ante consented to (hence compensated for in a sense) at the time of buying

the ticket, but many losses of those made worse-off by using the Kaldor-Hicks criterion or wealth maximization with hypothetical, but not actual compensation cannot be said to be similarly consented to.

Though the issues of distribution/equality and that wealth cannot be an ultimate value that are focused here have also been discussed before (e.g., Dworkin 1980; Mathis 2009, ch. 8), this entry pushes the discussion to a deeper and more comprehensive level to provide a more adequate basis for making wealth maximization acceptable as an instrument for welfare maximization. In particular, instead of equating wealth maximization with the acceptance of the Kaldor-Hicks criterion (Posner 1980a, p.244), we argue for wealth maximization in specific issues, showing that, when supplemented with appropriate redistribution, all income groups will be made better off. This provides a much stronger basis than the Kaldor-Hicks criterion.

Wealth Maximization for Welfare Maximization

Posner (1979, 1980a) rejected utilitarianism and proposed wealth maximization to replace utilitarianism. Obviously, wealth maximization differs from utilitarianism or welfare maximization. A very large amount of wealth, if very unequally distributed, with the majority starving, may clearly be a situation of very low welfare. Thus, wealth maximization cannot be an acceptable ultimate objective. I have defended utilitarianism and welfarism elsewhere (Ng 1975, 1990, 2016; the last specifically against Posner's arguments). This section shows how wealth maximization may be used (with supplementation) for welfare maximization. Since wealth is not intrinsically valuable while welfare is, this is a much reasonable position to take. In fact, perhaps in the back of his mind, Posner may also have something similar. In the concluding paragraph of his 1979 paper (p. 136), he proposed a system that "provides foundation and accommodation both of individual rights and of the material prosperity upon which, in the modern world, the happiness of most people

depend." This sounds like welfarism. A welfarist also wants material prosperity and individual rights for the pursuit of happiness.

The obvious deficiency ("monstrousness") of wealth maximization without appropriate supplementation is that a large amount of wealth, if very unequally distributed, may still be disastrous. Thus, an obvious way to supplement wealth maximization is to achieve an appropriate, usually more equal, distribution. Whenever distribution is too unequal, especially when some individuals are too poor, redistribution is made to avoid disastrous outcomes. More appropriately, redistribution policies are used routinely to avoid extreme inequality and poverty. Many such policies are in fact maintained in most high-income countries in the form of progressive income taxation and welfare spending including unemployment benefits. In the presence of such supplementation, wealth maximization may be pursued in specific areas, including the judiciary as advocated by Posner, with all income groups, the rich and the poor alike, being made better off, as shown below.

If redistribution may be carried out costlessly, it is clear that wealth maximization makes sense, at least when only economic issues are involved. The highest feasible amount of wealth should be achieved, and then redistribution may be carried out to achieve the highest level of social welfare. Redistribution is of course not costless; substantial costs include administrative, compliance, and policing. However, economists emphasize the efficiency costs of distortion or disincentive (which is a distortion in the work-leisure choice). Especially for high-income countries, this emphasis makes sense, as the essential costs of raising government revenues and some basic public spending are largely given. Choices involve only the extra costs of raising more or less revenues. Then for these relevant *marginal* costs, the disincentive effects are the more important costs. Focusing on the efficiency costs of disincentive effects of tax/transfer, it is argued below that using the rule of efficiency supremacy or wealth maximization in specific issues (including in the judiciary) may be justified.

Roughly speaking, in specific issues, we should use efficiency supremacy ("a dollar is a

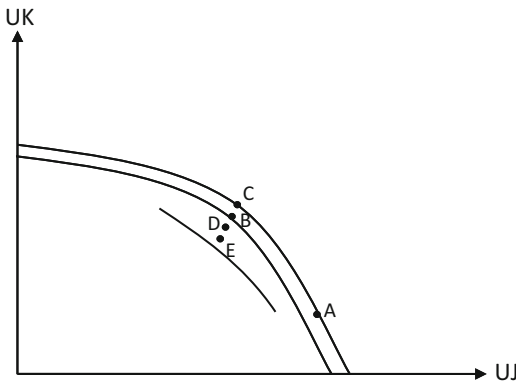
dollar irrespective of recipients”) to maximize the total social wealth. We then also use appropriate redistribution to achieve some desirable level of equality. What about the costs of redistribution? Despite substantial costs of redistribution, the case for efficiency supremacy in specific issues is not compromised. Though redistribution using tax/transfer to achieve equality has substantial disincentive effects, the pursuit of equality in specific issues like judiciary decisions also has the same degree of disincentive effects, if the degree of equality (in real incomes) achieved is the same. This is so because in the work-leisure choice of a rational person (irrationality may cause divergences either way and will likely net out to be negligible), not only will the tax/transfer system (its degree of progressivity or the marginal tax rates at the relevant income levels in particular) be taken into account, but the effective redistributive effects in specific issues will also be taken into account. It is true that information will be imperfect, but again the effects of this imperfection may go either way and will likely net out to be negligible. This is so because people (extreme misers possibly excepted) do not earn incomes for incomes sake but to gain utilities. When you earn higher incomes, if the extra money cannot be used to buy convenient parking spaces (distributed by first-come-first-served instead), the values gained by earning extra incomes will be discounted, equivalent to having a higher marginal income tax rate. Whether we achieve a given degree in equality (in real income or utility) by the tax-transfer system or by additional inefficient but pro-equality measures in specific issues including judiciary decisions, the same degree of incentives will prevail. However, the violation of efficiency supremacy will involve additional efficiency costs and hence is inferior.

For any system using a particular tax/transfer schedule and some inefficient but pro-equality measures in specific issues, we may dismantle these inefficient measures but at the same time make the tax/transfer schedule more equality oriented or more progressive. This could be done in a way that the combined effects are to make every income groups indifferent after and before this change in policy. As we then have the same utility

distribution over groups of people of different earning abilities, the same incentive effects will prevail for all groups. As the change is to dispense with inefficient measures, the government could impose more taxes than it needs to give more subsidies, in order to make all groups indifferent to before the change. The efficiency gain of this efficiency supremacy system then manifests in having a higher government net revenues which could be used to make all groups better off. This result was proved rigorously as a proposition and published in one of a few, if not the very top journal in economics, the *American Economic Review* (Ng 1984). It is somewhat ironical that, right in the heyday (early to mid 1980s) of the criticism of Posner’s wealth maximization, a proposition was proved that is most supportive of Posner’s position, but no one seemed to realize this connection for three decades. (See however Kaplow and Shavell (1994) on the superiority of using income tax/subsidy than the legal system in redistributing income.)

As all groups are made better off, this is at least a quasi-Pareto improvement. It is only quasi, because, in the presence of heterogeneous preferences of different individuals within the same income group, we may not be able to ensure that no one is made worse off. However, those made worse off could be compensated by the gainers within the same income group. This makes the quasi-Pareto change satisfying the Kaldor-Hicks criterion within each and every income group. As the compensation test in the original Kaldor-Hicks criterion is only hypothetical, it may not be widely acceptable if applied across all income groups. If the very rich are made better off by \$10 billion and the very poor are made worse off by \$5 billion, and the compensation is not actually carried out (with the actual compensation, the change would be a Pareto improvement), the result, even if wealth increasing, may be very welfare reducing. However, our quasi-Pareto change here satisfies the compensation test within each and every income group. It is thus much more acceptable as a sufficient condition for a social improvement than the Kaldor-Hicks criterion.

Our argument here may be illustrated using the UPF (utility possibility frontier) curves. In Fig. 1



Wealth Maximization: Efficiency and Equality Considerations, Fig. 1 Disincentive effects

and starting from a situation at point A, it is tempting to think that, if a specific judiciary decision will move us from A to B, it may be welfare improving. Despite the lower wealth or the lower efficiency involved (the UPF through B being lower than the UPF through A), the higher degree in equality at B may make it to be on a higher welfare contour than A. Choosing between A and B as the final point, I would also choose B. However, starting from A, instead of making an inefficient or wealth-decreasing move to B by a decision/policy on a specific issue, we would be better off trying to increase equality from A, aiming at C. If redistribution is costless and C is attainable, it is Pareto superior to B. Everyone will agree that this is a better move. However, in the presence of redistribution costs, it is not feasible to move from A to C. We can only travel from A along the broken utility feasibility frontier (UFF) to D. If the disincentive effects are substantial, D may be Pareto-inferior to B. Does this make the inefficient move from A to B desirable? The answer based on the analysis of this section is: No! The reason is that, if redistribution through a general tax-transfer from A, aiming at C, will end up at D through the disincentive effects, the redistribution through the inefficient move from A to B will also generate similar disincentive effects, moving us from B to E, which is Pareto inferior to D. Thus, the real choice is between A or moving to either D (through general redistribution) or to E (through the inefficient pursuit of equality in specific issues). The choice is clear, either we stay

at A, or if we want equality enough to be willing to pay the redistribution costs, we move to D by general redistribution. Never choose E by violating efficiency supremacy in specific issues.

It is true that an equality-improving measure need not be inefficient. However, under our proposed efficiency-supremacy in specific issues, such efficiency-consistent equality-improving measures will not be ruled out. Only measures that are inefficient are ruled out. In fact, all specific inefficient measures, whether equality-increasing, equality-decreasing, or equality-neutral, should be ruled out. An example of inefficient measure is the use of first-come-first-served in allocating scarce parking spaces in city central areas. Usually, it is pricing at very low levels, making first-come-first-served the effective allocation method, at least partially. The refusal to increase the prices to equate supply and demand results in excess demand with insufficient empty spots. Drivers seeking for empty spots have to waste too much time and petrol, not to speak of the resulting increased congestion and pollution. This inefficient measure achieves some equality at least in parking, as people with low incomes may also afford to park there. This equality effect is presumably the reason for the popularity of such measures. However, this perception is based on seeing only one area without a more overall view. The efficient solution is to increase parking fees to avoid excessive space hunting. This in itself may make the rich (who are willing to pay more for convenient parking spaces) better off and the poor worse off. However, if we also tax the rich more and help the poor more, all groups could be made better off. The same degree in equality in real income or utility could be achieved more efficiently.

Similarly, in judiciary decisions, at least for those affecting economic efficiency, the effects should not be weighted by incomes, such as counting a dollar to the poor as worth two dollars and a dollar to the rich as worth only half a dollar, as in fact proposed and done by some economists in cost-benefit analysis. Such systems of income or distributional weighting appear to be desirable in increasing social welfare, though they are counter to wealth maximization or efficiency. This

belief in the desirability of these systems is again based on seeing the direct effects without seeing the indirect effects. The policy of using distributional weights in cost-benefit analysis also makes earning higher incomes less valuable, increasing the disincentive effects. Instead of trying to pursue equality in specific issues, we should pursue equality generally, leaving the specific issues to be purely efficiency oriented. In principle, this is similar to the point that, to achieve equality, we should try to achieve a high degree of equality in total purchasing power, leaving the choice of how to spend the total incomes to the individuals. The objective of equality is achieved without the distortion of efficient individual choices or the violation of individual freedom of choice. For example, even if we want to achieve complete equality, it is better to equalize posttax/subsidy incomes across the board instead of allocating each and every good equally. Assuming perfect effectiveness (though unlikely in real life), both methods achieve perfect equality. However, the later inefficient policy would give Madonna and Kwang the same amount of lipsticks, but Kwang does not use lipsticks.

It is tempting for a particular judge or analyst to think, "I have no control on the distribution policy which I take as given, then for a particular decision/proposal, if I could promote equality somewhat at only a slight cost in efficiency, this should be a good thing." In fact, something similar probably had taken place for many decades, contributing to the inefficient pursuit of equality in many areas (such as the first-come-first-served system in car parking discussed above). Partly as a result, the total system was then perceived to be too low in terms of efficiency (including incentives) such that, starting from the 1970s, the degrees of progressivity in income taxation have been decreased several times in many countries. Combined with this reverse changes in progressivity, we end up with no more equality, but with many inefficient policies. We would have been better off without those inefficient equality-oriented policies in specific issues and kept the higher degrees in tax progressivity. Short-term good doings lead to long-run badness. Thus, Posner's advocacy of the separation of the judiciary and the legislature

and the related wealth maximization in the judiciary may largely be justified. For specific issues including judiciary decisions, efficiency supremacy irrespective of distribution (a "separation of distribution and efficiency" continuously challenged until now, e.g., Driesen and Malloy 2015, p. 5 of typescript) is acceptable, if supplemented by appropriate distribution in the general equality policy, as this will make all income groups better off.

Reconciling Rights with Welfare Maximization: Concluding Remarks

In the face of strong criticism, Posner (e.g., 1985, 2015) has retreated into pragmatism without giving up wealth maximization. While this provides some workable practicability, the conceptual justification and consistency are lacking. In particular, he seems to be confused and confusing in both accepting and rejecting both utilitarianism and individual rights at different places at the same time. The present entry provides both a justification and an explanation of Posner's position. Posner probably see some merits and problems in both utilitarianism/welfarism and rights but has not been able to reconcile them in a conceptually consistent way, hence, the apparently inconsistent positions. In addition to the previous section, this section tries to provide a consistent position that may lie at the back of Posner's head. At least this is the position I see as consistent and justifies wealth maximization for welfare maximization.

Since welfare is the only ultimate value (Ng 1990), the **ultimate** objective for a society should be social welfare – cannot be wealth, GDP, or any such similar objective things with no intrinsic value. Accepting welfarism and ignoring issues of animal welfare, social welfare is a positive function of individual welfares. Taking account of indirect and long-term effects, welfarism is consistent with the acceptance of certain rights, constitutional provisions, etc. if these may be justified on the ground of long-run welfare maximization. On any specific issue that a court, a government, or a policy adviser is making a

decision/suggestion, we are in a particular point in time with our given history, constitutions, statutory laws, legal precedences, accepted rights, moral principles, given endowments of individuals both in ability and inherited wealth, etc. Ideally, the previous decisions contributing to these “givens” should be based on welfare maximization, but few things in the real world can be ideal. The point here is that these “givens” determines the initial point A in Fig. 1 in the previous section. There, we show that, given any initial point such as A, when we are making a decision/suggestion on a specific issue, we should follow the principle of efficiency supremacy, treating a dollar as a dollar whomever it goes to. But this efficiency supremacy in specific issues (very close to if not exactly wealth maximization in Posner’s sense) should be supplemented by the equitable equality promotion policy in some branch of the government other than the judiciary.

Our position in using welfare maximization as the intrinsic, ultimate, and long-run moral position, but using wealth maximization (for welfare maximization) for specific issues in the short-run situations with many “givens,” may then explain Posner’s apparently conflicting positions under his “hybrid ethical system” (Posner 2015, p. 10) not only in a conceptually consistent way but provides a strong support for the efficiency (all groups better off) of wealth maximization for welfare maximization with equality supplementation. Our position that allows for feelings for justice and short-term “givens” also appears to be less extreme than that of Kaplow and Shavell (2002) who nevertheless provide a very powerful argument for basing policies “exclusively on their effects on individual’s well-being” (p. 465). I accept this at the ultimate level, but make some concessions in dealing with many “givens” in short-run decisions. These include the feelings of fairness, justice, etc. However, ultimately, these are important not intrinsically, but by affecting welfare or happiness directly or indirectly. Frustration from feelings of injustice already counts as negative welfare, and it may generate further negative welfare effects in the future through increasing social disharmony,

etc. Taking account of all these is consistent with welfare maximization. On the other hand, this entry goes further than this and accepts the use of wealth maximization in specific issues including judiciary decisions. However, in contrast to Posner, this is accepted under the qualification of appropriate supplementation in redistribution and as more clearly instrumental for welfare maximization.

Our acceptance of the many “givens” including the feelings of justice in the short run may be regarded as a form of pragmatism. However, since for the long run and intrinsically, we have a consistent guide of welfare maximization, our position is conceptually adequate and consistent. For example, we will not be in an awkward position for all those who believe in rights (or imperatives, including Kantian) for rights sake when faced with conflicting rights or when some rights may lead to disastrous outcomes in terms of welfare. This is so since no deontic, including Kant himself, believes in only one single right/imperative, and that welfare is not important. If one believes in only one single right, e.g., no murder under any circumstance and regards this as more important than any sacrifice in welfare, then he can be consistent. However, he may end up failing to save the whole world from going to hell when an escaped murderer carries out his hell creation plan. When we create/accept more rights/principles to constrain the excess or inadequacy of just one single right, we face the awkward position when two or more of our rights/principles are in conflict with each other. By using rights/principles for welfare maximization, we can provide a rational basis for accepting which rights (e.g., why human rights, animal rights, but no stone rights) and also on how to reconcile different rights when in conflict.

Of course, the effects on future welfare in the observation or violation of different rights by different degrees, or the welfare effects of social disharmony under different contexts, etc., are very difficult to guesstimate, not to say estimate. Nevertheless, these are matters of our lack of relevant information on which further studies may help to lower, not a matter of conceptual or moral philosophical acceptability.

Cross-References

► [Posner, Richard](#)

References

- Campbell D (2012) Welfare economics for capitalists: the economic consequences of Judge Posner. *Ober* 33(6):2233–2274
- Driesen DM, Malloy RP (2015) Critiques of law and economics. Available at SSRN 2572574
- Dworkin RM (1980) Is wealth a value? *J Leg Stud* 9(2):191–226
- Fernandes BG (2015) Legal argumentation and theories of adjudication in the US legal tradition: a critical view of Cass Sunstein's minimalism, Richard Posner's pragmatism and Ronald Dworkin's advocacy of integrity. In: *Argument types and fallacies in legal argumentation*. Springer International Publishing, Cham, pp 203–218
- Hatzis AN, Mercuro N (eds) (2015) *Law and economics: philosophical issues and fundamental questions*, vol 22. Routledge, New York
- Kaplow L, Shavell S (1994) Why the legal system is less efficient than the income tax in redistributing income. *J Leg Stud* 23(2):667–681
- Kaplow L, Shavell S (2002) *Fairness versus welfare*. Harvard University Press; previously published in *Harvard Law Rev* (2001) 114 (4): 966–1388
- Kronman AT (1980) Wealth maximization as a normative principle. *J Leg Stud* 9(2):227–242
- Mathis K (2009) *Efficiency instead of justice? Searching for the philosophical foundations of the economic analysis of law*. Springer, Berlin
- Ng Y-K (1975) Bentham or Bergson? Finite sensibility, utility functions and social welfare functions. *Rev Econ Stud* 42(4):545–569
- Ng Y-K (1984) Quasi-Pareto social improvements. *Am Econ Rev* 74(5):1033–1050
- Ng Y-K (1990) Welfarism and utilitarianism: a rehabilitation. *Utilitas* 2(2):171–193
- Ng YK (2016) Posner's wealth maximization for welfare maximization: separating efficiency and equality considerations, Working paper, Economic Growth Centre, Nanyang Technological University
- Pino-Emhart A (2016) The old and new philosophical foundations of tort law. *Modern Law Rev* 79(3): 504–518
- Posner RA (1973) *The economic analysis of law*. Little Brown, Boston. Many editions published subsequently
- Posner RA (1979) Utilitarianism, economics, and legal theory. *J Leg Stud* 8(1):103–140
- Posner RA (1980a) The value of wealth: a comment on Dworkin and Kronman. *J Leg Stud* 9(2): 243–252
- Posner RA (1980b) The ethical and political basis of the efficiency norm in common law adjudication. *Hofstra Law Rev* 8(3), Essay 2
- Posner RA (1984) Wealth maximization and judicial decision-making. *Int Rev Law Econ* 4(2):131–135
- Posner RA (1985) Wealth maximization revisited. *Notre Dame J Law Ethics Public Pol* 2:85–105
- Posner RA (1990) *The problems of jurisprudence*. Harvard University Press, Cambridge, MA
- Posner RA (2015) Chapter 1, Norms and values in the economic approach to law. In: *Law and economics: philosophical issues and fundamental questions*. Routledge, London, 22:1–15
- Veljanovski CG (1981) Wealth maximization, law and ethics – on the limits of economic efficiency. *Int Rev Law Econ* 1:5–28

Weber, Max

Dustin Garlitz

Department of Philosophy, University of South Florida, Tampa, FL, USA

Abstract

Max Weber (1864–1920) was a German jurist, political economist, and social scientist. This essay examines his intellectual background and academic career in economics and law, his place in the youngest German Historical School, his economic legacy, his work in economic and legal theory and economic and legal history, and his contributions to law and economics in his work *Economy and Society*, among his other notable publications. Focus is given to works of his that engage religion, in which he derives legal norms in a historical fashion. In regard to his contributions to economics, an emphasis is made on political economy, with his economic contributions presented in that context, as well as among a larger historical portrait.

Weber's Academic Career in Economics and Law

Throughout the course of his academic career, Max Weber's notable studies included works in law, legal history, economics, and economic history. Weber's father was a lawyer and a politician.

Weber pursued law and a legal education from the time he first matriculated at the University of Heidelberg in 1882. In his first semester at the University, he enrolled in a course entitled “Roman Law: Pandects and Institutes.” Weber then studied at the University of Strasbourg in 1883 and went on to the University of Berlin in 1884.

It was after earning his degree in law in 1886 that Weber sat for and passed the examination for Referendar, or took and passed the Bar exam, and commenced the practice of law. Weber then started work in the State Law and Agrarian History Faculty at the University of Berlin that year (Turner 2000: xv) and eventually pursued advanced legal studies there. In his legal studies at the University of Berlin, Weber had worked under the directorship of some of the leading representatives of the late nineteenth-century German legal academy. While continuing his legal studies at Berlin, Weber interned in the courts. Weber took a broad approach to his studying of law, which was common in his day. Accordingly, in addition to studying law, he spent time reading theology and philosophy, as well as a significant amount of history and political economy. As such, law and political economy had long been inextricably associated in the German cameralist tradition (Mommsen and Osterhammel 1987: 27).

When Weber became a doctoral student in 1888, he had also begun serving as a law clerk at the Royal District Court of Charlottenburg. His scholarship as a doctoral student bordered between legal and economic history, and he eventually publishing his doctoral dissertation on the history of medieval trading companies in 1889. Weber graduated *magna cum laude* that year with the full title of his dissertation published as “Development of the Principle of Joint Liability and the Separate Fund in the Public Trading Company from the Household and Trade Communities in the Italian Cities.” After Weber’s doctoral degree in law was conferred, he followed by serving as a junior barrister and then became an assistant judge or what was known as an *assessor*. As such, he had become a qualified lawyer in Berlin. After being drawn into legal practice, Weber first doubted his ability and potential to carry out

scholarly research in law. It was only after his chance of working as a syndic in Bremen – which he considered practical work – fell through that he decided to write a postdoctoral thesis, completing his *Habilitation* (Mommsen and Osterhammel 1987: 27). This second thesis was completed in 1891, which was titled *Roman Agrarian History and Its Importance for State and Civil Law*, and earned him the right to teach commercial, German, and Roman law.

Weber was awarded an extraordinary professorship in commercial and German law at the University of Berlin in 1892. After his brief stint at Berlin, Weber left the legal academy to take up a full professorship in political economy at the University of Freiburg in 1894, giving his inaugural academic lecture there titled “The National State and Economic Policy” in 1895. It has been argued that, in the former year, at the age of 30, Weber had lost interest in both the practice and the teaching of law, although he remained deeply committed to resolving basic questions about the nature and social functions of law as they appeared throughout the study of history. Weber went on to accept a chair in economics at the University of Heidelberg in 1896. In 1919, the year before his death, he moved to the University of Munich to take up a chair in economics.

Weber’s publication titled *Roscher and Knies: The Logical Problems of Historical Economics* (Weber 1975) was originally published in three parts in *Schmoller’s Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft*. The publication of Weber’s *Critique of Stammler* (Weber 1977) was essentially an extended review of Rudolf Stammler’s second edition of *The Historical Materialist Conception of Economy and Law: A Sociophilosophical Investigation*. Weber delivered lectures which were given to the public, under the title “Outlines of Universal Social and Economic History,” in the winter semester of 1919–1920, and which formed the publication that is titled *General Economic History* (Weber 2003). Weber’s most controversial work was *The Protestant Ethic and the Spirit of Capitalism* (Weber 2001), which he published in 1905. His magnum opus was his two-volume encyclopedia *Economy and Society* (Weber 2013), published

posthumously in the 1920s. It was in *Economy and Society* that Weber outlined many different types and forms of law and legislation and provided an explanation of the scope of economics and the economy in the framework of the emergent social sciences.

The Youngest German Historical School and Weber's Economic Legacy

Despite Weber's educational background in law, his professional career in academia was mainly spent in economics, in which he held two chairs. More precisely, it was his work in economic history that gave him the reputation as one of the leading figures of the youngest Historical School in Germany, along with Arthur Spiethoff and his frequent interlocutor Werner Sombart, the latter of whom he often held intellectual disagreements with while at times working on similar historical problems. Weber's concerns in economics broadened over the course of his academic career with his main focus eventually shifting to an area of scholarship that was to become known as economic sociology. Weber's concern within economic history – particularly as explicated in his work *The Protestant Ethic and the Spirit of Capitalism* – aligned well with the general climate of fin-de-siècle historical schools in Germany and in England. Weber, unlike others in the German School, spent very little time narrating the role played by economic policies of governments in economic change. Instead, like Sombart, Weber focused much more on the systematic study of modern capitalism, its nature, and the conditions for its emergence.

One of Weber's publications was devoted exclusively to the study of world economic history, the aforementioned *General Economic History*, which as stated previously was based on the transcripts of Weber's lectures in 1919–1920 and taken from his students' notes. The English edition was published in 1927 and was based upon the translation made by economic theorist Frank H. Knight, who spent most of his notable academic career at the University of Chicago. Knight taught a graduate seminar on Weber at Chicago in

the 1930s and wrote an article in 1928 on modern capitalism that dealt with Sombart as well as Weber, in which he sided with Weber. *General Economic History* was essentially an aggregate survey of economic development from ancient times to modernity. The text provided summaries, and in some case revisions, of important arguments that had been found in earlier writings, including succinct characterizations of the patterns of western economic development, as well as insightfully brief explanations of significant economic changes that were sometimes detailed in other writings (Turner 2000: 261).

Weber's Legal History

In his work *Ancient Judaism* (Weber 1967), Weber gave a treatment of the origins of the law, a problem that he and earlier legal scholars saw as occurring over the entire course of Old Testament studies. It has been argued that Weber's work in legal history was primarily a search for examples of laws and legal institutions that corresponded to the various "ideal types" which he attributed to the law (Turner 2000: 223–239). In other works, he compared Hindu law to Western ecclesiastical law, the German *Sachsenspiegel*, and Islamic and Judaic practices, and elsewhere, Islamic law to ancient Roman juristic practice. Weber classified the Canon Law of the Roman Catholic Church as sacred law and attributed its commitment to formal rationality partially to the relationship it had developed in late antiquity with the Roman law, and even more so, due to its transformation in the late eleventh century during the pontificate of Pope Gregory VII, at which time he found that the Church's functionaries were holders of rationally defined bureaucratic offices (Turner 2000: 223–239).

Weber's treatment of English common law adhered to the mainstream view of legal scholarship in his time in that he found modern English law was essentially a continuation of medieval English law. This led Weber to conclude that therefore it was not formally a rationalized type of law. And Weber gave treatments of many other substantial legal systems in the history of

civilization, as the case of his studies of the earliest of Roman law, formal canon law, Germanic law, Chinese law, and other Jewish, Hindu, and Buddhist laws. All of these legal histories figured prominently in different areas of Weber's legal scholarship (Turner 2000: 223–239).

Weber's *Economy and Society*: Economics

In *Economy and Society* (Weber 2013), Weber gives an explication of the economy and its arena of normative and de facto powers (Weber 2013: 311–635). This involves an explanation of the economy and social norms, the economic relationships of organized groups, and basic outlines of those groups. In Volume Two of the work (Weber 2013: 641–1363), Weber focuses on studying economic systems that reigned prior to the emergence of capitalism, such as feudalism (Weber 2013: 1070) and mercantilism (Weber 2013: 1097). At that point, he also pays attention to the basis and premise of economic enterprises, the economic and political power revealed in the study of ancient Germanic history (Weber 2013: 1256), and provides an attempt to outline an economic theory of antiquity (Weber 2013: 1266–1339).

In the first chapter of Part One (Weber 2013: 3), Weber discusses the frameworks of enterprise and formal organization, as well as voluntary and compulsory associations. Next, in the second chapter of Part One (Weber 2013: 63), he gives an analysis of the concept of economic action, a definition of the concept of utility, as well as an explication of the various modes of the economic organization of action (Weber 2013: 71), including typical measures of rational economic action and disparate types of economic organization. Weber continues by studying the media of exchange (Weber 2013: 75), means of payment, money, and the primary consequences of using money and credit and gives an analysis of its market, formal and substantive rationality of economic action (Weber 2013: 85), the rationality of monetary accounting, management and budgeting, the concept and types of profit-making,

and the role of capital (Weber 2013: 90). Also, he discusses the various kinds of calculations, such as substantive conditions of formal rationality in a money economy, market economies, and planned economies (Weber 2013: 109).

Weber's focus then shifts to the various types of economic division of labor (Weber 2013: 114), including types of the technical division of labor, social aspects of the division of labor (Weber 2013: 122), and the definition of occupation and occupational structure in such a context (Weber 2013: 140). Weber continues by studying production (Weber 2013: 150), from the appropriation of the material means of production to the appropriation of managerial functions and the expropriation of workers from the means of production. Followed is an analysis of various forms of communism (Weber 2013: 153), capital goods and capital accounting, the concept of trade and its principal forms (Weber 2013: 156), the condition of maximum formal rationality of capital accounting (Weber 2013: 161), and the principal modes of capitalistic orientation of profit-making (Weber 2013: 164). Weber then concludes by giving an outline of the monetary system of the modern state and the different kinds of money (including currency money) (Weber 2013: 166), restricted money, note money, the formal and substantive validity of money, and the methods and aims of monetary policy (Weber 2013: 180), followed by an excursus that involves the presentation of a critical note on the state theory of money, the nonmonetary significance of political bodies for economic order, the financing of political bodies (Weber 2013: 194), repercussions of public financing on private economic activity (Weber 2013: 199), the influence of economic factors on the formation of organizations, and a formal description of the mainspring of economic activity (Weber 2013: 202).

Weber's *Economy and Society*: Law

In the third chapter of Volume Two (Weber 2013: 941–952), Weber focuses on types of legal domination, including the basis of legitimacy, which he describes as domination and legitimacy. Weber

then outlines the three pure types of authority, first of which is legal authority, second traditional authority, and third charismatic authority (Weber 2013: 1111). In the description of the latter, Weber includes explanations of the routine character of charisma, its history in feudalism, as well as its transformation in a democratic direction. In the eighth chapter of Volume Two, titled “Economy and Law,” and which he labels “The Sociology of Law” (Weber 2013: 641), Weber discusses the fields of substantive law, which includes public law and private law, right-granting law and regulation, government and administration, as well as criminal law and private law, tort and crime, the limitation of power and separation of powers, substantive law and procedure, and the categories of legal thought (Weber 2013: 654). Later in the chapter (Weber 2013: 666), Weber focuses on the forms or creation of rights, including logical categories of legal propositions, which entails liberties and powers such as the freedom of contract, the development of freedom of contract, status contracts and purposive contracts which included the historical origin of the purposive contracts, institutions auxiliary to actionable contract (Weber 2013: 681) such as agency, assignment, negotiable instruments, limitations of freedom of contract, and extension of the effect of a contract beyond its parties and special law as conceived of as associational contacts including the notions of juristic personality, freedom, and coercion (Weber 2013: 729).

The next focus of Weber in the eighth chapter is on the emergence and creation of legal norms (Weber 2013: 753), including the emergence of new legal norms, entailing the presentation of theories of customary law insufficient as explanations, the role of party practices in the emergence and development of legal norms, and the process from irrational adjudication to the emergence of judge-making law (Weber 2013: 758), the development of new law through imposition from above and its approaches to legislation, and the role of the law prophets and of the folk justice of the Germanic assembly, as well as the primary role of law specialists (Weber 2013: 775). Weber then discusses legal honoraries and various

types of legal thought (Weber 2013: 784), including empirical legal training or law as craft; academic legal training, namely, law as a science; and its origins in sacred law, as well as the influence of Roman law (Weber 2013: 792). Weber continues by stressing formal and substantive rationalization as theocratic and secular law (Weber 2013: 809), namely, the general conditions of legal formalism; the substantive rationalization of sacred law; Indian, Chinese, and Persian law; as well as Islamic, Jewish, and Canon law, and then moves from there to explaining codifications (Weber 2013: 839), the driving forces behind the theory of codification, the reception of Roman law, and the development of modern legal logic, including the various types of patrimonial codification (Weber 2013: 856). Finally, Weber focuses on the formal qualities of revolutionary law (Weber 2013: 865), including natural law and the formal qualities of modern law, which entails an explication of revolutionary law and the French civil code in the natural law context, as well as expositions of natural law as the normative standard of positive law (Weber 2013: 866), explanations of the genesis of modern natural law, the transformation of formal into substantive natural law, class relations in natural law ideology (Weber 2013: 871), and the practical significance and disintegration of natural law and legal positivism in law as a vocation (Weber 2013: 875). In his discussion of the formal qualities of modern law (Weber 2013: 880), Weber gives an outline of particularism in modern law, the anti-formalistic tendencies of modern legal development, contemporary Anglo-American law, and lay justice and corporative tendencies in law as a vocation in modernity (Weber 2013: 892).

Weber's Economic and Legal Theory

Although trained in law, Weber did almost all his teaching in economics. He was first appointed to a chair of economics in Freiburg in 1894 and then became a successor to Karl Knies at Heidelberg in 1898. Weber's “Outline on Theoretical Economics,” or *Grundriss*,

consists, in its first part, mainly of thematic headlines and references. He expresses the aims and methods of theoretical economics there, citing the literature of Roscher, Hildebrand, Wagner, and Schmoller, but also of Menger, Sax, and Dietzel. Weber also wrote two pamphlets on Stock and Commodity Exchanges in 1984 and 1896 for the Göttingen Workers' Library, edited by Friedrich Naumann.

Weber was influenced by the legal theorist Georg Jellinek (Schluchter 1985: 83). There has been a significant amount of critical legal scholarship on Weber in relation to Rudolf Jhering, most notably expositions of their theories of action. Weber was a Heidelberg type philosophically and was a proponent of the Southwest German School of Neo-Kantianism, as was Jhering. Heinrich Rickert and Emil Lask were two of the leading figures of the school, and additional research has been conducted on Weber and Rickert among the cultural sciences and in a critical legal context, with Lask's 1905 work in legal philosophy complementing their intellectual thought. Weber's critique of Stammler was conducted in the context of Neo-Kantian legal philosophy. Additionally, legal scholarship that portrays Weber and Austrian legal theorist Hans Kelsen as Neo-Kantian is now emerging, although it is mostly focused on the nexus of law and politics in their works (especially its consequences for political and legal realism) rather than law and economics. In the tradition of the Southwest School of Neo-Kantianism, Weber attempted to theoretically position economics (as he did with law) within the cultural sciences. As his biographer Fritz Ringer has noted, Weber's long-term methodology involved the unification of the cultural sciences and social sciences (Ringer 1997), with law and economics at the core of these intellectual foundations.

Cross-References

- [Administrative Law](#)
- [Austrian School of Economics](#)
- [Cameralism](#)

- [Capitalism](#)
- [Causation](#)
- [Customary Law](#)
- [Law and Economics, History of](#)
- [European Nationality](#)
- [Government](#)
- [Hayek, Friedrich August von](#)
- [Mercantilism](#)
- [Mises, Ludwig von](#)
- [Political Economy](#)
- [Rationality](#)
- [Schmoller, Gustav von](#)

References

- Mommsen WJ, Osterhammel J (eds) (1987) Max Weber and his contemporaries. Allen & Unwin, London
- Ringer F (1997) Max Weber's methodology: the unification of the cultural and social sciences. Harvard University Press, Cambridge, MA
- Schluchter W (1985) The rise of western rationalism: Max Weber's developmental history (trans: Roth G). University of California Press, Berkeley/Los Angeles
- Turner S (ed) (2000) The Cambridge companion to Weber. Cambridge University Press, Cambridge
- Weber M (1967) Ancient Judaism (trans and eds: Gerth HH, Martindale D). The Free Press, New York
- Weber M (1975) Roscher and Knies: the logical problems of historical economics (trans: Oakes G). The Free Press, New York
- Weber M (1977) Critique of Stammler (trans: Oakes G). The Free Press, New York
- Weber M (2001) The Protestant ethic and the spirit of capitalism (trans: Parsons T). Routledge, London/New York
- Weber M (2013). Economy and society. In: Roth G, Wittich C (eds) An outline of interpretative sociology, 2 vols. University of California Press, Berkeley/Los Angeles/London

Further Reading

- Boucock C (2000) In the grip of freedom: law and modernity in Max Weber. University of Toronto Press, Toronto
- Camici C, Gorski PS, Trubek DM (eds) (2005) Max Weber's "Economy and society": a critical companion. Stanford University Press, Stanford
- Gane N (2012) Max Weber and contemporary capitalism. Palgrave Macmillan, Basingstoke/New York
- Hagemann H, Nishizawa T, Ikeda Y (eds) (2010) Austrian economics in transition: from Carl Menger to Friedrich Hayek. Palgrave Macmillan, Basingstoke/New York

- Holton R, Turner BS (2011) Max Weber on economy and society. Routledge, London/New York
- Koslowski P (ed) (1997) Methodology of the social sciences, ethics, and economics in the newer historical school: from Max Weber and Rickert to Sombart and Rothacker. Springer, Berlin
- Kronman AT (1983) Max Weber, Jurists: profiles in legal theory. Stanford University Press, Stanford
- Mardellat P (2009) Max Weber's critical response to theoretical economics. *Eur J Hist Econ Thought* 16(4): 599–624
- Parsons S (2003) Money, time and rationality in Max Weber: Austrian connections. Routledge, New York/London
- Peukert H (2004) Max Weber: Precursor of economic sociology and heterodox economics? *Am J Econ Sociol* 63(5):987–1020
- Radkau J (2009) Max Weber: a biography (trans: Camiller P). Polity Press, Cambridge
- Shionoya Y (2005) The soul of the German historical school: methodological essays on Schmoller, Weber and Schumpeter. Springer, New York
- Sumiya K (2001) Max Weber and the critical succession of the German historical school. In: Yuichi S (ed) The German historical school: the historical and ethical approach to economics. Routledge, London/New York
- Swedberg R (1999) Max Weber as an economist and as a sociologist: towards a fuller understanding of Weber's view of economics. *Am J Econ Sociol* 58(4):561–582
- Swedberg R (2000) Max Weber and the idea of economic sociology. Princeton University Press, Princeton
- Turner SP, Factor RA (1994) Max Weber: the lawyer as social thinker. Routledge, London/New York
- Weber M (1964) The theory of social and economic organization (trans: Henderson AM, Parsons T, ed: Parsons T). The Free Press, New York
- Weber M. (1991) The sociology of religion (trans: Fischeff E). Beacon Press, Boston
- Weber M (1999). In: Swedberg R (ed) Essays in economic sociology. Princeton University Press, Princeton
- Weber M (2002) The history of commercial partnerships in the middle ages (trans: Kaelber L). Rowman & Littlefield Publishers, Lanham
- Weber M (2003) General economic history (trans: Knight FH). Dover Publications, Mineola
- Weber M (2008) Roman agrarian history: in its relation to Roman public and civil law (trans: Frank RI). Regina Books, Camas
- Yagi K (2005) Karl Knies, Austrians, and Max Weber: a Heidelberg connection? *J Econ Stud* 32(4):314–330

Wegfall der Geschäftsgrundlage

► Impracticability

Whistle-Blower Policy

Sandeep Kapur¹ and Anthony Heyes²

¹Birkbeck College, University of London, London, UK

²University of Ottawa, Ottawa, Canada

Abstract

This chapter describes the increasingly important phenomenon of whistle-blowing and identifies the elements of optimal policy responses.

Whistle-blowers

Whistle-blowers are individuals who make public, or report to some external authority, the illegality, gross mismanagement, abuse of power, or other form of wrongdoing that they observe happening in or by an organization of which they are a part. It is usually also accepted that whistle-blowers are motivated by some notion of public interest rather than by personal pecuniary reward.

Whistle-blowers can be based in the private sector, in government departments, or in the “third sector.” Some definitions – for example, that adopted by the Government Accountability Project (GAP) in the United States – limit the category to current *employees* of an organization. High-profile examples of whistle-blowers include Ingvar Bratt (who exposed corrupt practice in the supply of weapons to India by Swedish engineering company Bofors AB), Sherron Watkins (who exposed accounting malpractice at Enron), and Edward Snowden (who made public the eavesdropping and intelligence gathering methods of the National Security Agency in the United States).

Whistle-blowing can serve an important role, both in discouraging wrongdoing in organizations in the first place and in facilitating the ending of such practice if it is already ongoing. Enforcement based on whistle-blowing has the advantage that the whistle-blowing agent – as an insider to the organization – sees things that an external enforcement agency typically cannot. He or she

may also be able to provide the external enforcement agency with the “hard evidence” that the agency needs in order to prosecute the miscreants – or tell the agency how and where to find it. Consider the following example:

On October 5, 2006, United States Coast Guard inspectors from Coast Guard Sector Portland boarded the ship at Vancouver, Washington to conduct a routine inspection. Inspectors reviewed the ship’s Oil Record Book, a document in which ship engineers must record all transfers and discharges of oil. Inspectors were unable to identify any discrepancies but were subsequently contacted by the ship’s Second Engineer who discretely told the inspectors about a flexible hose being used to bypass pollution prevention equipment and send oily sludge directly into the ocean. This prompted inspectors to conduct a second inspection on October 6, during which they located the hose and found additional evidence of its use. As a result of the prosecution the owners (Irika Maritime) were fined \$750,000 and placed on four year probation. The judge noted that the whistleblower should be protected from reprisal, but held no expectation of financial reward when contacting the inspectors. (United States Attorney’s Office 2007, p. 2)

Optimal policy response to whistle-blowing

What is the optimal policy to whistle-blowing? One aspect of policy relates to the personal treatment of whistle-blowers. As insiders, those who choose to blow the whistle are vulnerable to retaliation by their managers or by colleagues who are involved in, or complicit with, the wrongdoing. There is a long history – both anecdotal and academically documented – of whistle-blowers being subject to ostracism and other negative reprisal within their workplaces. There is evidence that the career prospects of employees who become whistle-blowers can be severely damaged by their actions, whether they stay with the firm that was the subject of their complaint or try to move to another organization (the employee can stigmatized by the label that they carry). This often carries over into financial stress, family breakdown, and adverse health impacts. Many organizations have evolved internal codes of practice to protect whistle-blowers from retaliation and other negative outcomes.

To some extent whistle-blowers may benefit from standard protections under employment or common law. But in recent decades most advanced countries have developed specific policies around whistle-blowing to enhance these protections. In the United Kingdom the Public Interest Disclosure Act (1999) and in the United States the Whistleblower Protection Act (1989) – strengthened by the subsequent US Whistleblower Protection Enhancement Act (2007) – provide overarching legal protection. In 2004 the Organisation for Economic Cooperation and Development (OECD) published its revised Principles of Corporate Governance to include whistle-blower protection (Organisation for Economic Cooperation and Development [OECD] 2004). In practice, these policies vary in their effectiveness.

In addition to policies aimed at protecting whistle-blowers from harm, we also observe specific efforts to encourage whistle-blowing. This includes the use of financial rewards or “bounties” for eliciting disclosure of information that can be used for successful prosecution for wrongdoing. The implementation of whistle-blowing “hotlines” – such as that established by the UK Financial Services Authority in 2008 for the reporting of financial fraud – aims to make the process of information disclosure easier. More recently in the United States, the Securities and Exchange Commission has set up a program that offers substantial financial reward for whistle-blowers whose disclosures lead to significant fines for financial misconduct.

The second aspect of whistle-blower policy relates not to how whistle-blowers themselves are treated but how the organization responds to their disclosures: whether such disclosures are always followed by investigation, and the severity of any penalties imposed on those found responsible for wrongdoing. Apart from the sheer cost of investigating every complaint, there are subtle informational constraints that affect the optimal policy design. Formal analysis of the regulatory design has been sparse. A substantial hurdle in setting up a microeconomic model is specifying the nonstandard motivations of whistle-blowers. To the extent they are motivated by some spirit of public interest – the gains from their intervention

are reaped by others – even as they typically bear significant personal costs, their action is difficult to rationalize in a conventional neoclassical economic model populated by self-interested actors (*Homo economicus*).

Theoretical approaches

Heyes and Kapur (2009) provide a *behavioral* economic model of whistle-blowing. Drawing on the well-established empirical literatures of whistle-blower motivations taken from social psychology, sociology, and related disciplines, they identify three sets of reasons why people choose to blow the whistle on wrongdoing in their workplace.

- (1) “Conscience cleansing”: the hypothesis is that employees have moral codes that preclude them from being complicit with – “going along with” or “keeping quiet about” – immoral activities proposed by or engaged in by their principals. Psychologists assert that being complicit with an immoral activity corrupts the self and “. . . whistleblowers disclose because they dread living with a corrupted ‘self’ more than they dread the other outcomes” (Alford 2001, p. 90). This was identified as the most common from the research available from other fields.
- (2) “Welfarism”: individuals obtain “warm glow” from improving welfare and engage in some sort of “cost-benefit” calculus in deciding when to blow the whistle, comparing external benefits of disclosure with its private cost. This is consistent with the notion of “justifiable disclosure” evidenced by Hunt (1997).
- (3) “Punishment motivation”: under this behavioral assumption employees disgruntled with their employer – perhaps for reasons unrelated to the wrongdoing in question – are motivated to hurt or do damage to the firm that employs them. Blowing the whistle on observed malpractice gives them an opportunity to do this.

Heyes and Kapur operationalize these diverse motivations and embed them in a principal-agent model with otherwise standard features. They use

the framework to assess the efficiency of a regulatory enforcement program based on tip-offs from whistle-blowers. They show that optimal policy is sensitive to the motivations attributed to the whistle-blower in any context. For instance, it is easy to see that policy must be more responsive to a “welfarist” whistle-blower than to the potentially frivolous disclosures of a disgruntled employee. Policy may also be sensitive to the fact that whistle-blowing norms vary across cultures (Patel 2003). Their model offers a general framework for analyzing how policy towards whistle-blowing should be tailored to the informational constraints faced by regulators and the complex motivations that lead whistle-blowers to act.

References

- Alford F (2001) Whistleblowers. Cornell University Press, Ithaca
- Heyes A, Kapur S (2009) An economic model of whistleblower policy. *J Law Econ Organ* 25(1):157–182
- Hunt G (1997) Whistleblowing, pages 681–90 in Robert Chadwick, *The encyclopedia of applied ethics*. Academic, New York
- Organisation for Economic Cooperation and Development (OECD) (2004) *Principles of corporate governance* (Revised). OECD, Paris
- Patel C (2003) Some cross cultural evidence on whistleblowing. *J Int Account Res* 26(1):69–96
- US Attorney’s Office, Western District of Washington (2007) Shipping line pays \$750,000 in fines and assessments for illegal ocean Dumping. Available at www.usdoj.gov/usao/waw/press/2007/jan/irika.html

WTO: Procedural Rules

Yong-Shik Lee¹ and Alessandro Romano²

¹The Law and Development Institute, Decatur, GA, USA

²China-EU School of Law, China University of Political Science and Law, Beijing, China

Abstract

The Dispute Settlement Mechanism (DSM) of the World Trade Organization plays a crucial

role in resolving interstate trade disputes. This entry analyzes from a law and economics perspective some of the most important procedural rules of the DSM: (i) the standing, (ii) the allocation of litigation costs, and (iii) the existing remedies.

Introduction

Procedural rules are a set of rules prescribing the formal steps that must be taken in order to enforce legal rights. Because procedural rules define the mechanics of legal cases, they are a key component of any legal system. Unsurprisingly, law and economics has always paid a great deal of attention to procedural rules, yet most of the studies focus on private litigation. This entry examines and summarizes the findings of the relatively scant literature on the efficiency of procedural rules in interstate litigation, particularly the rules of dispute settlement under the World Trade Organization (WTO). There are two key reasons for investigating procedural rules within the WTO. First, the Dispute Settlement Mechanism (DSM) of the WTO is generally regarded as highly successful and is markedly different from the dispute settlement procedures adopted by other international institutions. Second, the DSM presents some features that are in clear contrast with the procedures adopted by national jurisdictions. Therefore, the standard litigation models developed to describe the incentives for litigating in national jurisdictions (e.g., Priest 1980) cannot properly explain the functionalities of the DSM. The most relevant differences are (i) the nature of the parties involved in the litigation, (ii) the allocation of litigation costs, and (iii) the existing remedies.

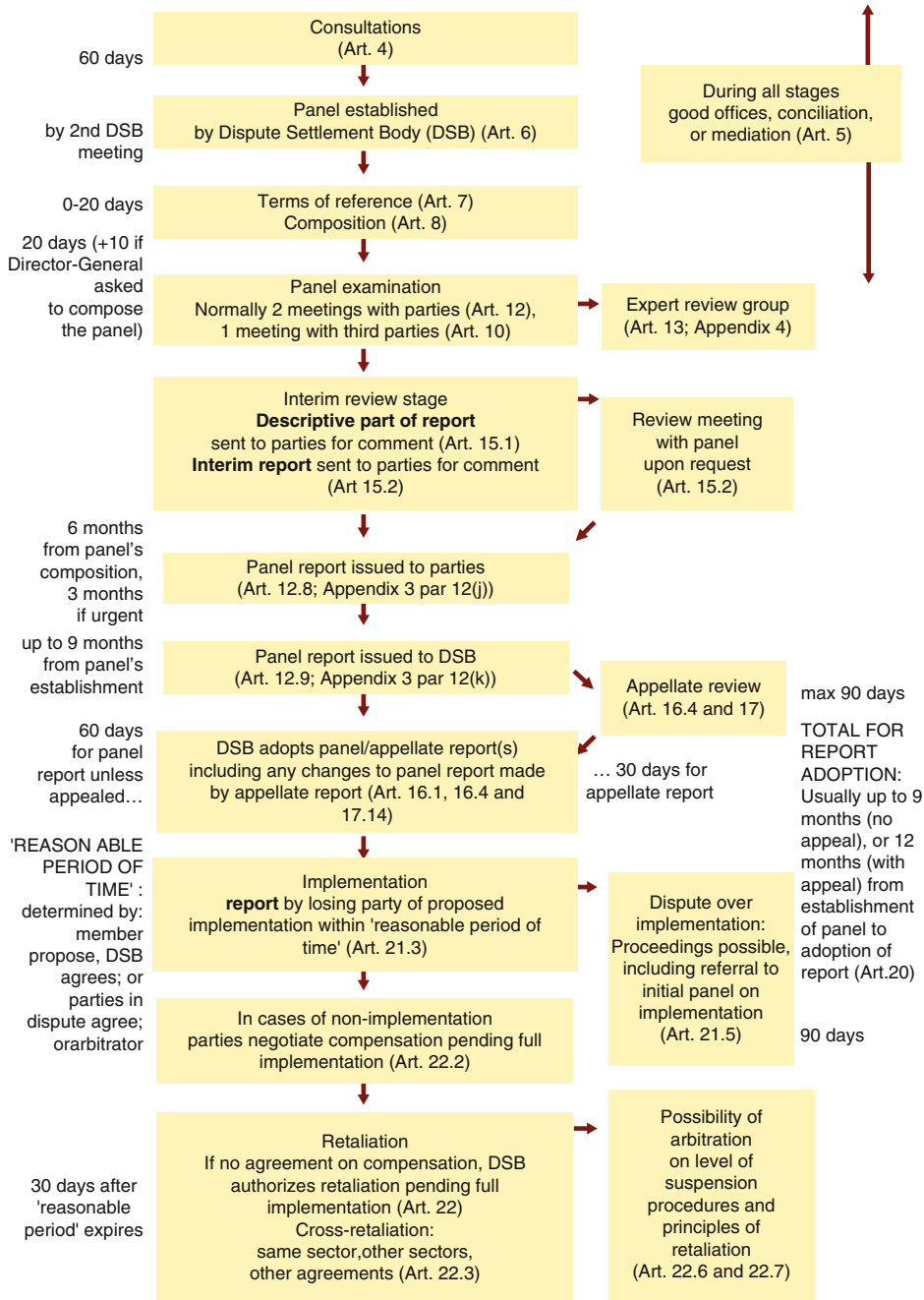
The WTO and Its Dispute Settlement System

The WTO is a result of the multilateral negotiations that took place between 1986 and 1994 (also known as the Uruguay Round). The primary objective of the WTO is to help trade flow as

freely as possible so long as there are no undesirable side effects because this is important for economic development and well-being (www.WTO.org). In other words, the main goal of the WTO is to eliminate, or at least reduce, barriers to trade among countries, including tariff and non-tariff barriers (NTBs). Tariffs are charges applied to goods in relation to the assessed value of the goods at customs, and NTBs are a wide range of regulations and policies that create significant costs for foreign firms in the import or export of products and services. To promote international trade through a rules-based system such as the WTO, it is essential to establish an effective system of dispute resolution. To serve this need, the DSM was created as a part of the WTO agreements (Annex 2 of the Agreement establishing the World Trade Organization). This dispute settlement system, generally considered the central pillar of the WTO, is embodied in the Understanding on Rules and Procedures Governing the Settlement of Disputes (hereinafter, Dispute Settlement Undertaking or DSU). The DSM can be invoked by member countries (members) when a member considers another member to have adopted a trade measure in violation of its obligations under GATT/WTO disciplines, leading to the impairment of any benefits accruing to it under the covered agreements (DSU Art. 3). Since its inception, the DSM has been frequently used to settle trade disputes among members, as demonstrated by the large number of cases, over 500, brought to the DSM between 1995 and 2016. When compared to the number of reports, only 101, adopted in the almost half-century of GATTs existence (Lee 2016), these figures show that members have relied on the WTO DSM much more than on its predecessor to resolve their trade-related disputes.

The WTO dispute settlement process is a highly legalized process, articulated in well-defined multiple phases (see Fig. 1). Its three main phases are described in the following:

1. Every dispute case must begin with a consultation phase for 60 days (DSU Arts. 4.5 and 4.7). The requirement of consultation is consistent with the goal of the DSM that is to settle



WTO: Procedural Rules, Fig. 1 Flow chart of the Dispute Settlement Process of the WTO (Source: www.wto.org, Reproduced with the permission of ©World Trade Organization (WTO) 2016)

disputes rather than making formal rulings (World Trade Organization 2001). The vast majority of cases are settled in the consultation process; as of October 2014, only 156 out of

486 dispute settlement cases had advanced to the panel stage (Lee 2016).

2. The complainant may request the establishment of a panel if the dispute has not been

resolved at the consultation stage. Panels are quasi-judicial bodies that adjudicate disputes arising between members. Panels are created on an ad hoc basis and are composed by either three or five experts detached from any government (DSU Arts. 8.1 and 8.2). As a general rule, the panel issues its report and circulates it publicly 6 months after its composition (DSU Art. 12.8).

3. Both the complainant and the respondent may appeal the panel decision to the Appellate Body (AB). The AB is a permanent body within the WTO DSB composed by seven members. The role of the AB is to review the legal aspects of panel reports and examine the consistency of panel decisions with relevant GATT/WTO rules (DSU Art. 3.2).

The Law and Economics of Procedural Rules

The WTO dispute settlement process is distinctively different from national jurisdictions in three aspects: the standing of parties, the allocation of litigation costs, and the remedies. These differences are further explained below.

Standing: Public Enforcement in the WTO

The DSM does not allow litigation by private parties; only members have standing to bring complaints before the DSB (DSU Art. 10.2). Thus, the WTO has adopted a mechanism of public enforcement (i.e., the parties of an international agreement – in this case the members – have the exclusive standing to litigate in the dispute settlement body) (Sykes 2005). In contrast, most domestic legal systems and many other international fora rely on private enforcement, in which private parties litigate against wrongdoers. For example, virtually all the cases litigated in the International Centre for Settlement of Investment Disputes (ICSID) are initiated by private parties (Choi 2007). Law and economics literature has attempted to explain why the WTO adopted public enforcement. According to Sykes (2005), governments of importing countries have no incentive to allow private parties standing on trade-related

issues and invite challenges to their trade measures by them.

The public enforcement scheme creates an interesting dynamism with respect to private interest groups. Under a private enforcement mechanism, private parties can decide when to litigate a case and base this decision on their own cost-benefit calculation (Priest 1980). In contrast, in the WTO, the firms that are being harmed by a restrictive trade measure have to persuade the government to litigate the case (Luanratana and Romano 2014). However, private parties do not have equal influence over the government (Tullock et al. 2002); strong and concentrated interest groups may influence the government to engage in litigation when it is not in the public interest (Shaffer 2003). On the contrary, small firms operating in fragmented markets may not have sufficient leverage to convince the government to litigate their case, even when litigation would be efficient and in the public interest (Romano et al. 2015). Small firms affected by a restrictive trade measure may not be able to deliver an organized, united voice to the government. This is a collective action problem; free riding and coordination problems prevent small firms from joining forces to influence the government, even if it would be in their best interest to do so (Olson 1965). An additional concern associated with a system of public enforcement is that governments might be reluctant to engage in a dispute due to diplomatic considerations (Davis and Shirato 2007). For example, Japan has not been actively bringing complaints against China, because the latter is Japan's largest export market (Davis and Shirato 2007).

Litigation Costs

The allocation of litigation costs is widely recognized by law and economics scholars as an important determinant of the number and kind of disputes litigated. The underlying rationale is that a rational actor will decide to litigate when the expected benefits from the litigation outweigh its cost (Priest 1980). For example, let us assume that the benefits (B) of winning the case are equal

to 150, the cost of litigation (C) is 100, and the probability of winning the case (P) is 0.5 (for the sake of simplicity, we do not consider the possibility of a settlement). In this case, the plaintiff has no incentive to litigate because the cost of litigation is higher than the expected benefits, or $C > P * B$. Although this view only offers a very stylized description of litigation, it holds some descriptive power in the context of private litigation. By comparison, the WTO presents an additional layer of complexity due to the exclusive standing of the government in the litigation, as well as the absence of any rule requiring firms affected by a trade measure to contribute toward the cost. While there are cases in which firms contribute to the cost of litigation, particularly in developed countries, it is not a general practice. The incentives of private parties and the government can be summarized as follows (Luanratana and Romano 2014):

1. Firms do not internalize the actual cost of litigation. Instead, they consider the share of the costs that they agree to cover (which can be anything ranging from 0 to the full cost of litigation) plus the cost of lobbying the government. This sum is not necessarily equal to the litigation costs, and it can even be higher if there is special resistance on the side of government. In the cited example, let us assume that firms only bear half the litigation costs and that lobbying has a cost of 10. Then the litigation has a positive payoff for the firm because the expected benefits are still greater than the cost ($50 + 10 < 150 * 0.5$).
2. Governments look at the cost of litigation minus the quota borne by the firm and the political costs minus the political benefits of litigating the case.

Therefore, the incentives of the government and the firms cannot be captured by the standard model of private litigation. It is the interplay between the incentives of these two agents, the government and the firms, that determines which cases are brought to the DSB. While there is no evidence that the interplay of the incentives of firms and governments leads to an efficient

selection of cases, data strongly suggest that complainants only start litigation when they have a high likelihood of succeeding. In fact, complainants prevail in around 90% of all cases (Guzman 2002; Hoekman et al. 2009; Keisuke 2003), compared to about 50% in the ICSID (Turk 2011), suggesting that strong cases are selected and brought before the DSB. The literature has not agreed on a single explanation for the observed pattern (e.g., Colares 2009; Guzman 2002).

Remedies

Remedies have been the subject of extensive discussion in the context of compliance with DSB decisions (Bronckers and Van den Broek 2005). Generally, an AB or panel decision does not entail monetary compensation (see, e.g., Mavroidis 2000) but requires the infringing member to remove the noncomplying measure (DSU Art. 19) within a reasonable amount of time (DSU Art. 12.3). Therefore, unlike cases in private litigation, the DSM does not aim to compensate the complainant for the infringement but to prevent continuing rule violations by ensuring withdrawal or modification of measures that are inconsistent with WTO agreements (DSU, Art. 3.7). If the infringing member does not withdraw or modify the measure within a reasonable amount of time, the complainant has the right to retaliate (DSU Art. 22.2) by suspending or withdrawing trade concessions in the form of raising customs duties and other nontariff barriers.

Law and economics scholars have raised concerns about the efficiency of the existing remedies, because the absence of monetary compensation might negatively affect deterrence. Optimal deterrence is achieved only when the injurer (the infringing member) internalizes the harm by compensating the victim (the complainant) for the entire harm caused (Shavell 2009). If no monetary compensation is to be awarded, then the decision on the introduction of a restriction to trade (trade measure) will largely be based on a political calculus, such as the possibility of drawing domestic support for introducing a trade measure that benefits its constituency or political ramifications in

international relations. There is no reason to assume that these political considerations systematically lead to an efficient outcome.

Moreover, retaliation seems to be a poor surrogate for compensation due to a number of reasons. First, retaliation is not always a possibility, particularly for small developing countries that cannot realistically impact the trade of the infringing country (Bronckers and Van den Broek 2005). Second, retaliation may affect the trade of the retaliating country more adversely than that of the infringing member (Bronckers and Baetens 2013). Third, retaliation generally targets sectors of the economy of the infringing country not related to the trade measure (Bronckers and Van den Broek 2005). For instance, as a retaliation against the EU for its decision not to comply with the DSB ruling on its banana regime, the United States restricted the import of Italian batteries (*European Communities Regime for the Importation, Sale and Distribution of Bananas*). Some scholars have argued that retaliation can be effective if the interest group affected by retaliation (Italian battery producers) effectively lobbies the government to lift the restrictive measure in order to end the retaliation (Nzelibe 2005). It is ultimately an empirical question whether exporting interest groups will mobilize frequently enough and can effectively induce their government to comply with the DSB decision. Lastly, retaliation is in and of itself a restriction to trade and hence runs counter to the goals of the WTO (Bronckers and Van den Broek 2005).

Future Research

There are two main paths of law and economics research, analytical (descriptive or positive) and normative. The former addresses the existing rules and institutions and their effects, whereas the latter explains their possible modification in order to achieve a given goal (Coleman 1980; Lee 2014, 2016). It is not straightforward to anticipate the directions in which these two paths of research will proceed in the domain of international trade law in general, as well as in the study of the WTO in particular.

There is an active debate as to whether the existing procedural rules are penalizing developing countries. In fact, a large number of developing countries, particularly least-developed ones, have not used the DSM (Bown and Hoekman 2005). Possible reasons for the underrepresentation of developing countries might be their political and economic weakness vis – vis more powerful trading partners, their lack of the expertise in litigating cases, and the high costs associated with litigating a case in the WTO system (Davis and Bermeo 2009). Understanding to what extent developing countries are disadvantaged by the status quo can fuel normative analyses aimed at identifying ways to facilitate their participation in the DSM. From this perspective, scholars have started to emphasize the role that many actors, including law schools and nongovernmental organizations, can play (Bown and Hoekman 2005). A fundamental guiding principle of normative research in this domain ought to be realism. Any change in the WTO requires the consensus of 164 members; therefore, radical changes to the status quo in the DSM are unlikely to take place in a near future (Grewal 2008).

Cross-References

► [International Litigation and Arbitration](#)

References

- Bown CP, Hoekman BM (2005) WTO dispute settlement and the missing developing country cases: engaging the private sector. *J Int Econ Law* 8:861–890
- Bronckers M, Baetens F (2013) Reconsidering financial remedies in WTO dispute settlement. *J Int Econ Law* 16:281–311
- Bronckers M, Van den Broek N (2005) Financial compensation in the WTO improving the remedies of WTO dispute settlement. *J Int Econ Law* 8:101–126
- Choi WM (2007) The present and future of the investor-state dispute settlement paradigm. *J Int Econ Law* 10:725–739
- Colares JF (2009) A theory of WTO adjudication: from empirical analysis to biased rule development. *Vand J Trans Law* 42:383–439
- Coleman JL (1980) Efficiency, exchange, and auction: philosophic aspects of the economic approach to law. *Calif Law Rev* 68:221–249

- Davis CL, Bermeo SB (2009) Who files? Developing country participation in GATT/WTO adjudication. *J Politics* 71:1033–1049
- Davis CL, Shirato Y (2007) Firms, governments, and WTO adjudication: Japan's selection of WTO disputes. *World Politics* 59:274–313
- Grewal DS (2008) *Network power: the social dynamics of globalization*. Yale University Press, New Haven
- Guzman AT (2002) *The political economy of litigation and settlement at the WTO*. UC Berkeley Public Law Research Paper 98
- Hoekman BM, Horn H, Mavroidis PC (2009) Winners and losers in the panel stage of the WTO dispute settlement. In: Thomas C, Trachtman JP (eds) *Developing countries in the WTO legal system*. Oxford University Press, New York, p 151
- Keisuke I (2003) Why does the World Trade Organization appear neoliberal? The puzzle of the high incidence of guilty verdicts in WTO adjudication. *J Public Pol* 23:1–21
- Lee YS (2014) *Safeguard measures in world trade: the legal analysis*. Edward Elgar Publishing, Cheltenham
- Lee YS (2016) *Reclaiming development in the World Trading System*. Cambridge University Press, Cambridge
- Luanratana W, Romano A (2014) Stare decisis in the WTO: myth, dream, or a sirens song? *J World Trade* 48:773–794
- Mavroidis PC (2000) Remedies in the WTO legal system: between a rock and a hard place. *Eur J Int Law* 11:763–813
- Nzelibe J (2005) The credibility imperative: the political dynamics of retaliation in the World Trade Organization's dispute resolution mechanism. *Theor Inq Law* 6:215–254
- Olson M (1965) *The logic of collective action*. Harvard University Press, Cambridge, MA
- Priest GL (1980) Selective characteristics of litigation. *J Leg Stud* 9:399–421
- Romano A, Sotis C, Yifei YB (2015) ICSID vs. WTO: an economic analysis of procedural rules. *N C J Int Law Com Reg* 41:31–58
- Shaffer GC (2003) *Defending interests: public-private partnerships in WTO litigation*. Brookings Institution Press, Washington, DC
- Shavell S (2009) *Economic analysis of accident law*. Harvard University Press, Cambridge, MA
- Sykes AO (2005) Public versus private enforcement of international economic law: standing and remedy. *J Leg Stud* 34:631–666
- Tullock G, Brady GL, Seldon A (2002) *Government failure: a primer in public choice*. Cato Institute, Washington, DC
- Turk MC (2011) Why does the complainant always win at the WTO? A reputation-based theory of litigation at the World Trade Organization. *31NW J Int Law Bus* 31:385
- World Trade Organization *Trading into the future* (2001) Available at: https://www.wto.org/english/res_e/download_e/tif.pdf